2017

Towards a science of human stories: using sentiment analysis and emotional arcs to understand the building blocks of complex social systems

Andrew James Reagan
University of Vermont

Towards a science of human stories: using sentiment analysis and emotional arcs to understand the building blocks of complex social systems

A Dissertation Presented

by

Andrew J. Reagan

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Sciences

May, 2017

Defense Date: February 24, 2017
Dissertation Examination Committee:

Christopher M. Danforth, Ph.D., Advisor
Peter Sheridan Dodds, Ph.D., Advisor
Jacques A. Bailly, Ph.D., Chairperson
Samuel V. Scarpino, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of Graduate College

Abstract

We can leverage data and complex systems science to better understand society and human nature on a population scale through language — utilizing tools that include sentiment analysis, machine learning, and data visualization. Data-driven science and the sociotechnical systems that we use every day are enabling a transformation from hypothesis-driven, reductionist methodology to complex systems sciences. Namely, the emergence and global adoption of social media has rendered possible the real-time estimation of population-scale sentiment, with profound implications for our understanding of human behavior. Advances in computing power, natural language processing, and digitization of text now make it possible to study a culture's evolution through its texts using a "big data" lens.

Given the growing assortment of sentiment measuring instruments, it is imperative to understand which aspects of sentiment dictionaries contribute to both their classification accuracy and their ability to provide richer understanding of texts. Here, we perform detailed, quantitative tests and qualitative assessments of 6 dictionary-based methods applied to 4 different corpora, and briefly examine a further 20 methods. We show that while inappropriate for sentences, dictionary-based methods are generally robust in their classification accuracy for longer texts. Most importantly they can aid understanding of texts with reliable and meaningful word shift graphs if (1) the dictionary covers a sufficiently large enough portion of a given text's lexicon when weighted by word usage frequency; and (2) words are scored on a continuous scale.

Our ability to communicate relies in part upon a shared emotional experience, with stories often following distinct emotional trajectories, forming patterns that are meaningful to us. By classifying the emotional arcs for a filtered subset of 4,803 stories from Project Gutenberg's fiction collection, we find a set of six core trajectories which form the building blocks of complex narratives. We strengthen our findings by separately applying optimization, linear decomposition, supervised learning, and unsupervised learning. For each of these six core emotional arcs, we examine the closest characteristic stories in publication today and find that particular emotional arcs enjoy greater success, as measured by downloads. Within stories lie the core values of social behavior, rich with both strategies and proper protocol, which we can begin to study more broadly and systematically as a true reflection of culture. Of profound scientific interest will be the degree to which we can eventually understand the full landscape of human stories, and data driven approaches will play a crucial role.

Finally, we utilize web-scale data from Twitter to study the limits of what social data can tell us about public health, mental illness, discourse around the protest movement of #BlackLivesMatter, discourse around climate change, and hidden networks. We conclude with a review of published works in complex systems that separately analyze charitable donations, the happiness of words in 10 languages, 100 years of daily temperature data across the United States, and Australian Rules Football games.

Citations

Material from this dissertation has been submitted for publication in *EPJ Data Science* on January 10, 2017, with the preprint available as follows:

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., Dodds, P. S.. (2016). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. *Preprint available online at https://arxiv.org/abs/1512.00531*.

AND

Material from this dissertation has been published in *EPJ Data Science* on November 4, 2016, in the following form. Within a month of publication, our paper was the most shared of all papers published in *EPJ Data Science*, as measured by Altmetric.

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., Dodds, P. S.. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science 5*(1), 31.

Table of Contents

Chapter 1

Introduction and Literature Review

Epoch watch: Welcome to the Sociotechnocene.

-@peterdodds 2012-01-10

1.1 Introduction

Individual words encapsulate information and emotion as the building blocks that we use to capture experiences in stories. Beyond words, multi-word expressions (phrases), conceptual metaphor, and complicated grammar (syntax) coalesce to provide incredible expressive power. Attempts to quantify semantic content build atop syntactic understand of language with the aim of transforming a model of meaning that has proven useful to our own cognitive machinery into something more readily applicable for another purpose (e.g., summarization by a computer). One such goal of semantic understanding is to measure the sentiment expressed in written communication, which is broadly known as sentiment analysis. The next evolution of natural language systems will tackle the harder-yet problems of pragmatics, where narrative understanding and generation can enable common-sense reasoning on par with human intuition.

In our work, we transfer the emotion of single, isolated words into a one-dimensional happiness measure to build the Hedonometer. Leveraging the Hedonometer technology and modern computational power, we analyze digitized text with the ultimate goal of understanding stories. This dissertation proceeds as follows: in this chapter we explore the foundations of sentiment analysis and narrative structure. In Chapter 2 we benchmark and compare methods for sentiment analysis.

In Chapter 3 we apply these methods and extract dominant emotional arcs from digitized text. In Chapter 4, we discuss contributions made to published work in the broader science of complex systems. Finally, in Chapter 5 we offer some concluding remarks.

Next, we examine prior work in natural language processing, sentiment analysis, and computational narrative understanding.

1.2 Sentiment analysis

The field of Natural Language Processing (NLP) has been around since the advent of computers, but is growing rapidly alongside computational advances. While major advances have been made, there remain many open problems. We focus here on a specific NLP problem, namely understanding the emotional content of language. We refer to the emotional content in a written text broadly as the sentiment. In addition to the summaries given in recent review articles (Giachanou and Crestani, 2016), the landscape of tools and technologies is expanding quickly and sentiment analysis systems are deployed to tackle important challenges. As we will see, sentiment analysis is a sub-field of NLP that can benefit from advancement in other realms of NLP as well (e.g., phrase partitioning).

Applications of sentiment analysis span academia, industry, and government. Just some of the current uses include predicting elections (Tumasjan et al., 2010), product sales (Liu et al., 2007), stock market movement (Bar-Haim et al., 2011), and tracking public opinion (Cody et al., 2015). NLP and measures of sentiment are used to analyze consumption of information from the media, and societal level decisions are driven by this flow of public opinion online. Beyond individual and collective decisions, corporate success demands an understanding of the public sentiments directed towards their products.

Advances in Artificial Intelligence (AI) have elucidated the distinction between problems that are hard for computers and those that are hard for humans—a difference that is not obvious at the outset. Determining sentiment is one such task: understanding the sentiment of our friends and colleagues through informal text is relatively easy for us, but it is hard to codify in a computer algorithm. As we will see, machine learning (often broadly referred to as AI) is finding uses in all areas of Natural Language Processing (NLP), including advancing the state-of-the-art in sentiment classification and sentiment dictionary creation. While sentiment analysis benefits from machine learning to create classifiers and sentiment dictionaries, the output of sentiment detection also aids higher level approaches to language understanding.

1.2.1 Psychology of emotion

With few exceptions, current sentiment analysis methods aim to detect sentiment one-dimensionally, giving a score on a range from negative to positive sentiment. While this pragmatic approach proves useful, Jack et al. (2014) conjectured that there are four basic emotions, Ekman (1992) names six, and Plutchik (1991) identifies two additional basic emotions in humans. These theories are only the most well known classifications, with at least 90 such classifications being given over the past century, as noted by Plutchik (2001). Through the use of brain imaging and fMRI techniques, researchers in neuroscience have also attempted to distinguish whether basic emotions are best captured as discrete categories (anger, fear) or underlying dimensions (valence, arousal). Altogether they have found consistent neural locations for basic emotions but no one-to-one mapping, and further research is still needed (Harrison et al., 2010; Hamann, 2012).

The widely acknowledged six basic emotions identified by Paul Eckman are:

- *happy*,
- *surprised*,
- *afraid*,
- *disgusted*,
- *angry*,
- and *sad*.

In Figure 1.1, a visualization of these six basic emotions is shown. As noted in the caption, these six emotions serve as a basis for more complex emotions. The eight basic emotions of Plutchik (1991) are shown as the variations along four dimensions in Figure 1.2. While we do not expect that each of the six basic emotions have orthogonal representations in their embodiment in language (as orthogonality may be inferred from the Figures, is found in facial expression, and underlies the theory), a basis of more than a single dimension is likely necessary to represent the full range of emotion. The basic emotions theory rejects that all emotions can be represented as either positive of negative states, and this should extend to language. Indeed, attempts to cast the basic emotions as either positive (e.g., happy) or negative (e.g., sad) are subjective, e.g. by Robinson (2008) classifying *pride* as a negative emotion. According to Ekman (1992), basic emotions are distinguished by nine characteristics:

1. Distinctive universal signals.
2. Presence in other primates.
3. Distinctive physiology.
4. Distinctive universals in antecedent events.
5. Coherence among emotional response.
6. Quick onset.
7. Brief duration.
8. Automatic appraisal.
9. Unbidden occurrence.

To this end, in Figure 1.3 the theory of Russell (1980) attempts to find the core dimensions of emotion using data from emotions manually labelled for 28 adjectives. The explained variance by the first two principal components would provide an indication of how well we can capture emotion with two abstract dimensions, however this is not provided by Russell (1980). Each of these theories expands upon the single dimension considered further in sentiment analysis: positive and negative. More complex emotions can be constructed from combinations of the basic emotions (e..g., `delight = joy + surprise`), which is not possible from combinations of simply positive and negative states (e.g., it would be nonsensical to find coefficients a, b for the abstract categories positive and negative to satisfy `delight = a*positive + b*negative`).

An alternative to basic, discrete emotions being the building blocks for all emotions is to place all emotions in the dimensions of *valence*, *arousal*, and *dominance*, often referred to as "norms" and measured alongside *concreteness* and *age of acquisition* (Lindquist et al., 2016). In the literature, the term *valence* is used interchangeably to mean the negative/positive emotional dimension.

The positivity bias in language is frequency-independent, as long as the frequency selections are rank ordered (see Dodds et al. (2015a) and Chapter 4). Schrauf and Sanchez (2004) asked participants to write as many emotion words as they could think of in two minutes, and found that participants were able to recall a larger list of negative emotional words. At least one theory for this difference, as elaborated in Koch et al. (2016), posits that this difference is because positive words

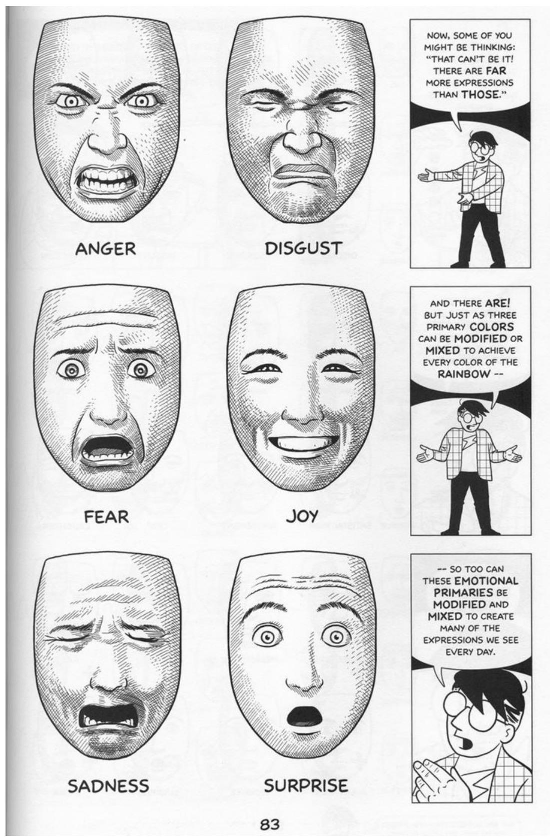


Figure 1.1: The six emotions of Ekman (1992), illustrated here by McCloud (2006). In principle, the entire range of human emotions can be constructed from this minimal "basis", e.g., the emotion delight is the addition of joy and surprise. This theory of basic emotions distinguishes these emotions as being fundamentally distinct, adapted for fundamental life tasks, and universally present through evolution (or, perhaps, universal social learning). In particular the distinction between basic emotions is not explained by variation in dimensions of arousal, pleasantness, or activity.

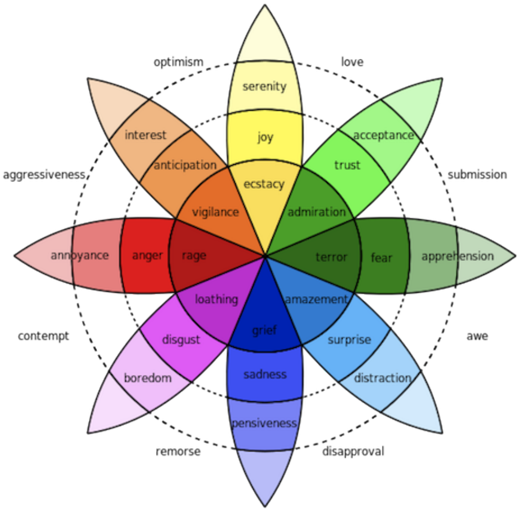


Figure 1.2: Schematic of the eight emotions from Plutchik (1991). The commonly known eight names (e.g., joy, etc.) are one row out from the center. In addition to the six emotions of Ekman (1992) we find anticipation and trust on the first level.

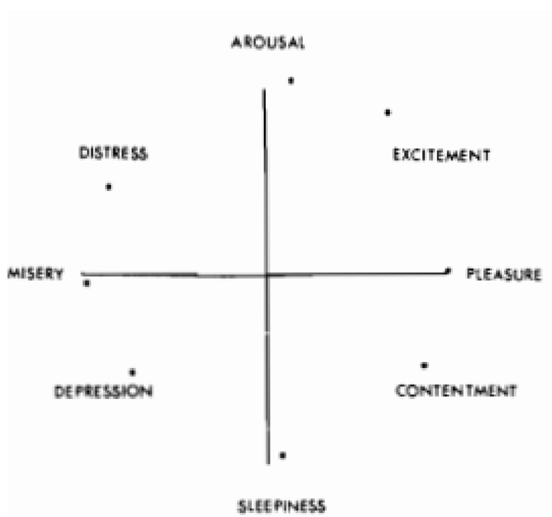


Figure 1.3: Eight emotions on the arousal–pleasure axis of Russell (1980), who finds these axis to be the best representation of emotion. To this end, using 28 emotional words manually annotated for the characteristics which they share, Russell finds the two major principal components in a Principal Component Analysis, establishing this "circular ordering." This circular ordering agrees well with the mental model of emotional states used by psychologists at the time.

are more similar than negative words. In one of six tests, they show that the scores for positive words are more tightly clustered than the scores for negative words from the Warriner & Kuperman sentiment dictionary.

In addition to the emotion of expression, we note that other work attempts to measure personality traits of individuals based on their expressions (rather than the sentiment of the expressions themselves), specifically Kosinski et al. (2013) and Youyou et al. (2015). As an example, given a person's micro-blog post, the algorithms developed by Kosinski et al. (2013) are trained to measure whether the person is an introvert or extrovert. These attempts fundamentally differ from sentiment analysis by measuring traits of an individual rather than traits of the expression, though in practice the two goals make use of similar machine learning techniques.

For the remainder of this chapter, we will assume that each emotion is being measured on a scale from -4 $\rightarrow$ 4, with 0 representing no presence of emotion and a score of -4/4 representing the maximum negative/positive emotional priming. While some dictionaries benefit from considering emotion on a different scale for human evaluation (e.g. "labMT" with $1 \rightarrow 9$ or "AFINN" with $-5 \rightarrow 5$), we make this choice to speak more generally about each sentiment dictionary we test.

1.2.2 Goals of sentiment analysis

It may help to first frame the problem of detecting sentiment in text, and we will utilize the generalization given by Bing Liu in his 2012 book *Sentiment Analysis and Opinion Mining* (Liu, 2012). Here, our goal is to detect and *understand* the average sentiment of a document using the words contained within: *document-level sentiment classification*. Our definition extends that of Liu (2012) to include the goal of better *understanding* text through sentiment detection, and this goal is complementary (and in some cases outright necessary) to achieve classification. While document length varies, Liu (2012) subdivides finer-grained classification into two categories: (1) classifying sentence-level sentiment and (2) classifying entity-level sentiment. Sentence-level sentiment is detecting sentiment in sentences, and entity-level sentiment aims to predict sentiments that are directed at named entities (e.g., products, people, or corporations). We express caution in pursuing these latter goals using existing methodology, namely in classifying short, informal text. We will examine in Chapter 2 how dictionary based approaches are effective at the document level, but fail at the sentence level (and by extension fail at the entity level as well). Several examples of different sentences are also given in Liu (2012), highlighting the difficulty of classifying individual sentences, and we share these examples here.

The accuracy of classifying documents correctly as positive or negative is commonly measured using precision, recall, and F-score statistics, as in Ribeiro et al. (2016). These measures assess the classification accuracy, but do not attempt to assess the qualitative goal of achieving better understand of text with sentiment analysis (an area on which our work will build). Both of these goals can be assessed with ground truth data, and next we review publicly available data sets for sentiment evaluation.

1.2.3 Publicly available annotated data

Review papers such as those by Giachanou and Crestani (2016) attempt to capture the many advances in the field, including applications of machine learning with training data, although they only identify 3 of the 17 sentiment dictionaries that we list in Chapter 2. They identify the lack of benchmarks as important issue (Giachanou and Crestani, 2016):

Short name	Description	# Samples	Referenced By
STS,Tweets_STF,STS-Test	Stanford Twitter Sentiment	499	G, R, S
Sanders,Tweets_SAN,Sanders	Sanders Corpus	3424	G, R, S
HCR,HCR	Health Care Reform	4616	G, S
OMD,Tweets_DBT,OMD	Obama-McCain Debate	3298	G, R, S
SS-Tweet,Tweets_RN_I,SS-Twitter	SentiStrength Twitter Dataset	4243	G, R, S
SemEval,Tweets_Semeval,SemEval	SemEval Datasets	6087	G, R, S
STS-Gold,STS-Gold	STS-Gold	2036	G, S
DETC,DETC	Dialogue Earth Twitter Corpus	N/A	G, S
Tweets_RND_IV	aisopos_ntua	500	R
Comments_TED	TED Comments	839	R
Comments_BBC	SentiStrength BBC Comments	1000	R
Comments_Digg	SentiStrength Digg Comments	1077	R
Reviews_I	SentiStrength Myspace Reviews	1041	R
RW	SentiStrength Runners World Forum	1046	R
Comments_YTB	SentiStrength YouTube Comments	3407	R
Amazon	VADER Amazon Reviews	3708	R
Reviews_II	VADER Movie Reviews	10605	R
Comments_NYT	VADER NYT Comments	5190	R
Tweets_RND_II	VADER Tweets	4200	R
Tweets_RND_III	DAI-Labor English MT	3771	R
ORT	Opinion Retrieval Twitter	5051	L

Table 1.1: Summary of publicly available Twitter datasets tagged with sentiment labels. In respect of Twitter's Terms of Service, lists of the Tweet IDs are provided, as well as a script to download the Tweets through Twitter's public API (note some data may not longer be available). We shorted the references as follows as G: Giachanou and Crestani (2016), S: Saif et al. (2013), R: Ribeiro et al. (2016), and L: Luo et al. (2012).

> One of the main challenges in evaluating approaches that address Twitter-based sentiment analysis is the absence of benchmark datasets. In the literature, a large number of researchers have used the Twitter API to crawl tweets and create their own datasets, whereas others evaluate their methods on collections that were created by previously reported studies. One major challenge in creating new datasets is how the tweets should be annotated. There are two approaches that have been followed for annotating the tweets according to their polarity: manual annotation and distant supervision.

To this end, we note the availability of datasets below and attempt to collect each dataset enumerated by Giachanou and Crestani (2016); Saif et al. (2013); Ribeiro et al. (2016) in Table 1.1 and make them accessible in one place online. In addition to these public datasets, some academic groups choose not to release their tagged data, and there are claims of very large datasets held by private companies in the sentiment analysis space. Given the time and cost associated with obtaining high quality training data, and the ubiquity of machine learning for sentiment analysis in industry, the training data itself can be viewed as a trade secret.

In addition to the tagged datasets above, we attempt to provide a comprehensive list of sentiment dictionaries in Table 2.1.

1.2.4 Natural Language Processing techniques

As itself a tool for NLP, sentiment analysis leverages approaches that are applied more broadly (e.g., classification), and can benefit, if only slightly, from other such techniques. In this section, we provide a very brief overview of techniques for processing raw text, detecting boundaries of multi-word expressions, disambiguating word senses, tagging parts-of-speech, and dependency parsing.

Tokenization

Here, we consider words as the basis for our computation, and the process of extracting words from raw text is often referred to as "tokenization". The simplest tokenization procedure is splitting raw text strings on spaces, with words being any contiguous non-space characters. For well structured (formal) writing, this approach presents few false positive matches, but this approach is often too simple for processing informal text (e.g., Twitter), where grammar is not reliable. To improve upon the aforementioned approach, we build a list of known "word characters" (e.g., the letters a-z, the apostrophe, hyphen, etc.) and extract all contiguous sequences of these characters as words. An example regular expression implementing this approach is provided in Section A.1.2. The final consideration here are the various uses of individual words; the representation of a word differs based on, but not limited to, the different classes, inflection, contractions, possessive use, and/or pluralization of the word. Depending upon the ultimate use case, a choice can be made for how to process words. A common choice is to reduce words to their root, a process called "stemming", which removes the inflection from words, a popular implementation is provided by Porter (2001). A widely used source for annotated data based on word stems is the morphology of WordNet (Fellbaum, 1998). In the approach that we adopt for sentiment analysis, we attempt to retain the most complete representation of words, without removing the information about usage that may be contained beyond a word's root or stem. This achieves a very basic and computationally efficient disambiguation between word senses.

Multi-word Expressions

In addition to tokenization, the meaningful units of language often span multiple words. These multi-word expressions, or "phrases", can also be extracted from tokenized words. Here we summarize two state-of-the-art approaches from Handler et al. (2016) and Williams (2016).

Williams, J. R. (2016). Boundary-based MWE segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.

Williams performs boundary-based MWE segmentation with text partitioning, building on prior work that introduces random and serial partitioning algorithms, and showing that phrase frequency follows Zipf's law more closely than words alone. Trained models for partitioning rely on (1) phrase likelihood from "informed random partitioning", (2) entries the Wiktionary, and (3) annotated corpora. The model is general purpose for pattern recognition, and can be run using text data or PoS tags, combining the output phrases for higher recall. Altogether, this achieves state-of-the-art performance with flexible application to any text-based corpora.

Handler, A., M. J. Denny, H. Wallach, and B. O'Connor (2016). Bag of what? simple noun phrase extraction for text analysis. *NLP+ CSS 2016*, 114.

Handler and colleagues build upon prior work that defines a grammar of PoS labels for noun phrases. In essence, the approach uses patterns to match noun phrases. The implementation realizes computational feasibility with a Finite State Transducer (FST) compiled to find all matches of their pattern represented by a Finite State Grammar (FSG). As an example of this general type of approach, the pattern of word labels Adjective Noun Noun (encoded ANN) would be successfully matched by the grammar `(A|N)*N(N)*`, where the $*$ represents 0 or more matches of the previous expression (as in standard regular expression syntax, otherwise known as the Kleene star). The availability of reliable part-of-speech tags is assumed by this approach, although this is known to be a harder problem for informal text (e.g., social media).

We conclude that both of these available methods, and even the "naive" method described by Mikolov and Dean (2013) offer an improvement upon unigram models for bag-of-words approaches to sentiment analysis, which includes the methods used in this dissertation. Sentiment dictionaries only contain ratings for single words, and extending existing dictionary ratings to MWEs is a widely acknowledged area for future research.

Word Sense Disambiguation (WSD)

To get a sense of the Word Sense Disambiguation (WSD) problem, here we examine a scholarly competition: The English All-Words Task of the SENSEVAL-2 series. The SENSEVAL competitions began in 1998, and the second and third instantiations took place in 2001 and 2004. After 2004, the scope of tasks was broadened and the name switched to SemEval, being held again in 2007, 2010, and 2012–2017 every year. First, we summarize the construction of the benchmark by Snyder and Palmer (2004), and then we examine the winning entry from Decadt et al. (2004).

Snyder, B. and M. Palmer (2004). The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 41–43. Association for Computational Linguistics.

To develop the training and testing data for Senseval-3, Snyder and Palmer extracted approximately 5,000 words from two Wall Street Journal articles and one excerpt from the Brown Corpus. Word sense was annotated by two people using Wordnet senses, and then settled by a third party, for a total of 2,212 words and multi-word-expressions. They found the inter-annotator agreement at 72.5%, representing a practical upper bound for the performance of computational methods.

Decadt, B., V. Hoste, W. Daelemans, and A. Van den Bosch (2004). Gambl, genetic algorithm optimization of memory-based wsd. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 108–112. Association for Computational Linguistics.

GAMBL is a "word expert" approach to WSD in which a word sense classifier is trained for each individual word. The parameters of this classifier are optimized using a genetic algorithm, and the method achieves the best precision/recall of .652.

Part-of-Speech tagging

Part-of-Speech (PoS) tagging aims to disambiguate between the various forms that a word can take: verb, pronoun, preposition, adverb, conjunction, participle, and article are eighth of the most well recognized categories. This information tells us how a word relates to the neighboring words around it, and finer grained taxonomies of parts of speech in English contain more than 80 types. To train and test algorithms for this task, large annotated corpora such as the Penn Treebank are available form Marcus et al. (1993) and OntoNotes .

Abney, S. (1997). Part-of-speech tagging and partial parsing. In *Corpus-based methods in language and speech processing*, pp. 118–136. Springer.

Abney (1997) elaborates upon the work of Church (1988) and DeRose (1988) to develop a reasonable, approximate approach to PoS tagging. State-of-the-art approaches can be classified into rule-based and stochastic, the latter making extensive use of Hidden Markov Models (HMMs) to represent state as a latent variable.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180. Association for Computational Linguistics.

Toutanova et al. (2003) develop a PoS tagger with improved accuracy which is competitive in terms of both speed and accuracy with any attempt since. This is achieved by using a cyclic dependency network to represent the state of the tagger, and achieves 97.24% accuracy on the Penn Treebank corpus. The tagger is used by Manning et al. (2014) in the most recent release the Stanford CoreNLP natural language processing toolkit.

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Existing PoS taggers excel at the task in well structured language but are not applicable to short, informal text. In Owoputi et al. (2013), large-scale unsupervised word clustering and lexical features are used to achieve 93% accuracy on Twitter. In addition, guidelines for manually annotating this type of text are provided.

The application of PoS tagging in stand-alone tests on tagged corpora has achieved high rates of accuracy on both formal and informal text. It now stands to reason that this addition of information for individual words and MWEs have applications in an end-to-end system for sentiment analysis.

Dependency Parsing

Dependency parsing aims to extract the syntactic relationship between the words used in a sentence. Also referred to as syntax parsing, dependency parsing is one more NLP tool that aims to solve a disambiguation problem: of all possible dependency parses, choosing the most appropriate. In many cases, this disambiguation is between two parses that are both grammatically valid, but nonsensical otherwise; consider the different interpretations of "They ate the pizza with anchovies" (seen in Figure 1.4). In the prior examples, anchovies could either be utensils or a topping or their friends, but this is obvious to us with commonsense knowledge. Other examples that I found compelling for parsing are garden path sentences—those which confuse the common human parsing by leading our parse down the wrong path—such as "the old man the boat" or "the horse ran past the barn fell". Both examples are valid senses, but are easy to read incorrectly on the first pass. The dependency parsing algorithms that we examine next solve each of the examples we have just given correctly by utilizing neural network approaches that find the most probable parse.

We note that PoS tagging, a shallower form of parsing, is about twenty times faster than parsing, for applications where computational costs of parsing are a bottleneck (Handler et al., 2016). State-of-the-art approaches from both Chen and Manning (2014) and Andor et al. (2016) achieve parse accuracies over 90%.

Chen, D. and C. D. Manning (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*, pp. 740–750.

In Chen and Manning (2014), a dependency parser is built that uses dense features of the surrounding text to improve upon both the accuracy and speed of current parsers. For performance, they note their "parser is able to parse more than 1000 sentences per second at 92.2% unlabeled attachment score on the English Penn Treebank".

Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Andor et al. (2016) from Google Inc. (now Alphabet) improve further on the accuracy of neural network parsers and release a pre-trained model for general consumption. Their pre-trained model is *Parsey McParseface* and they note that "for dependency parsing on the Wall Street Journal we achieve the best-ever published unlabeled attachment score of 94.61%".

Much like PoS tagging, dependency parsing algorithms extract meaningful information at the sentence level with high accuracy. An open challenge for sentiment analysis is the incorporation of this local information while retaining interpretability across large corpora.

Heuristics

In our pursuit to understand and evaluate sentiment analysis methods at a human level, it is intuitive yet deceiving to consider individual sentences. At the level of individual sentences, the bag of words approach is no longer useful. One attempt to improve these models for short text is to incorporate rules that are manually encoded to fit a given model for language, relying on the grammatical structure of language. Such a rule might be to consider negation words such as "not" to reverse the polarity of the following sentiment word, such that "not w_i" would be combined and assigned the score of "$-w_i$".

Various attempts to incorporate rule-based heuristics and dictionary approaches for sentiment analysis include the work of Thelwall et al. (2012) and Hutto and Gilbert (2014). The systems developed by Kiritchenko et al. (2014), Wilson et al. (2005), and Polanyi and Zaenen (2006) incorporate a rule for negation. An analysis of the usefulness of different features for Twitter sentiment analysis is performed by Agarwal et al. (2011), including PoS and binary lexicon features. Perhaps unsurprisingly, the polarity of words is the single most useful feature. The analysis showed that the most useful combination is the one of PoS with the polarity of words. Hutto and Gilbert (2014) report an increase on in the F1 score for binary Tweet classification of 2.1% using negation, extended vowels (“happy” to “haaapy”), punctuation, and capitalization as cues.

1.2.5 Building corpus-specific sentiment dictionaries

Categorization

Previous work on building sentiment dictionaries using data, as opposed to human evaluation, has taken various forms. We categorize these approaches by three main categories; (1) the type of data that is used to gain information about how words are similar, (2) how the data is processed, and (3) which methods are used to infer semantic properties.

Types of data include:

- Thesaurus
- Word associations
- Unstructured text corpora

Data processing

- Network from structured data
- Network for POS patterns
- Word embedding vectors
- Vectors similarity (cosine distance, etc) $\rightarrow$ networks (k-NN, etc)

Some of the methods employed:

- Graph clustering
- Graph label propagation
- Orthogonal subspace projection on embedding

We distinguish these approaches from machine learning approaches that estimate emotion of words from tagged training data in that these approaches extend existing scores about words.

Chronologically, the first approach here is by Hatzivassiloglou and McKeown (1997), and the most recent we have found is the work of Rothe et al. (2016). We will proceed by summarizing the main result of each paper, casting the methodology into one of the aforementioned categories.

Previous approaches

First, we take a close look at the earliest effort to build a corpus-specific sentiment dictionary to get a deeper sense of the steps involved in this task.

Hatzivassiloglou, V. and K. R. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 174–181. Association for Computational Linguistics.

Hatzivassiloglou and McKeown (1997) use a four-pronged approach: (1) adjectives are extracted from large text corpora that are linked by conjunctions ("and" or "but"), (2) a log-linear regression determines whether they are synonyms/antonyms to make a graph of positive/negative connections, (3) a clustering algorithm is run for two clusters, and (4) the cluster with the greatest average frequency is labeled as the positive words. The 1987 WSJ corpus is used, with PoS tags for adjectives and conjunctions. They report 82% accuracy on the binary classification of word pairs as synonym or antonym, and 90% accuracy on semantic orientation (predicting manual labels on 1336 adjectives). Their approach does not rely on existing word scores, but nevertheless forms the basis for future work that does incorporate existing sentiment dictionary data.

Now that we have seen one approach in more detail, we will look ahead to methodology that more closely informs our own work. The years following saw an expansion in the methods, processing, and data used to automatically extend affective word scores, including work (Turney, 2002; Turney and Littman, 2003; Taboada and Grieve, 2004; Kim and Hovy, 2004; Hu and Liu, 2004; Esuli and Sebastiani, 2006; Das and Chen, 2007; Kaji and Kitsuregawa, 2007; Blair-Goldensohn et al., 2008; Bestgen et al., 2008; Rao and Ravichandran, 2009). We start again in more depth with recent work of Velikovich, directly applicable to extending data sets that we are familiar with (e.g., labMT).

Velikovich, L., S. Blair-Goldensohn, K. Hannan, and R. McDonald (2010). The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010*

***Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 777–785. Association for Computational Linguistics.**

In the paper from Velikovich et al. (2010), many of the specifics of the approach are left out. We review this paper because the methodology outlined is very similar in spirit to all of the approaches that follow. For a domain corpus, they use n-grams up to length 10 scraped from 4 billion web pages, however the details of this corpus are left vague. They then use the cosine distance between context vectors from these n-grams to build a k nearest neighbor (k-NN) network with $k = 15$ (the method used to generate context vectors is again left to the reader). Seed words within the network are labeled with positive and negative sentiment, and scores for all n-grams are determined by shortest paths to the seed set, a using a generic graph propagation algorithm. For results, Velikovich et al. (2010) report that their effort compares favorably to the manually constructed lexicon from Wilson et al. (2005) and a lexicon from WordNet used in Blair-Goldensohn et al. (2008).

Bestgen, Y. and N. Vincze (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods 44*(4), 998–1006.

Bestgen and Vincze (2012) begin by taking 300-dimensional word embeddings from the Singular Value Decomposition (SVD) of the word co-occurrence matrix of the TASA corpus, comprised of 44K documents. They use these embeddings to build a k-NN network, and then use the DIC-LSA technique of Bestgen et al. (2008) with the ANEW dictionary (using the dictionary scores to measure correlations with words in the network). This approach extends the ANEW dictionary by adding scores to additional words, directly using the scores in the ANEW dictionary itself. For different values of k, the score for each word in the network is taken to be the average of it's neighbors (the k closest words in the embedding space), and for words with scores from ANEW, the node value itself is held-out. By using only the most extreme words (those in ANEW with scores closer to 1 and closer to 9), they achieve an correlation coefficient (Cohen's Kappa) of .53–.94 on sets of all–190 of the words from ANEW (the latter .94 correlation achieved with using the 190 most extreme words in ANEW). In addition, they provide ratings using their method for 17,000 English words.

Tang, D., F. Wei, B. Qin, M. Zhou, and T. Liu (2014). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *COLING*, pp. 172–182.

Tang et al. (2014) train a neural network (NN) to learn phrase sentiment from phrase embeddings using a graph collected from Urban Dictionary and Tweets with emoticons. The Tweets with emoticons are used to embed all phrases in a two dimensional space with the loss function as a hybrid between word context (e.g., word2vec) and emoticon label context (happy or sad). A network of words is extracted from Urban Dictionary and used to apply label propagation for positive (good, `:)`), negative (poor, `:(`), and neutral words (when, he) across the network (which includes phrases). The word embeddings and scores from label propagation are used as features for a ternary sentiment classifier that is trained to predict scores from label propagation. Their system outperforms those tested for the SemEval 2013 competition by attaining a performance of macro F1 score .78, and their final dataset, TS-Lex, is composed of 65,685 words with sentiment scores and provided online.

Amir, S., R. Astudillo, W. Ling, P. C. Carvalho, and M. J. Silva (2016). Expanding subjective lexicons for social media mining with embedding subspaces. *arXiv preprint arXiv:1701.00145*.

Their approach to lexicon expansion "consists of training models to predict the labels of pre-existing lexicons, leveraging unsupervised word embeddings as features" (Amir et al., 2016). Correlations between their method and existing continuous datasets had a maximum of 0.68, an improvement over support vector regression. The resulting lexicon out-performed other methods in Tweet classification, although not all methods were compared.

Hamilton, W. L., K. Clark, J. Leskovec, and D. Jurafsky (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

Hamilton et al. (2016) utilize the approach set out in Velikovich et al. (2010) to generate corpus specific word embeddings using SVD and propagating sentiment labels on inferred k-NN network. The most novel part of the approach measures the uncertainty in predicted labels with bootstrapping procedure that holds out fractions of seed set (with a seed set of 10 words, holding out 2). They claim to measure performance with correlations to existing dataset of Warriner et al. (2013), but not found in results.

Mandera, P., E. Keuleers, and M. Brysbaert (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology 68*(8), 1623–1642.

Mandera et al. (2015) measure sensitivity of the performance of corpus specific sentiment dictionaries to the number of words in the training data. They split the Warriner et al. (2013) corpus into training and testing sets at different thresholds (e.g., 70/30 and 80/20). Networks are built using k-NN and Random Forests on four different distances metrics, and the best performance is attained from the SVD of PMI embedding and a k-NN with $k = 30$. They show that accuracy for this best method varies from .61–.72 between a 10/90 to 50/50 split into testing and training. The reported accuracy leads the authors to cast doubts on the efficacy of automated approaches, but their survey is not exhaustive and the next methods we will explore improve upon the accuracy.

Van Rensbergen, B., S. De Deyne, and G. Storms (2016). Estimating affective word covariates using word association data. *Behavior Research Methods 48*(4), 1644–1652.

Van Rensbergen et al. (2016) estimate word scores using word association data for 14K dutch words, finding the best correlation between this method and human evaluation for k-NN algorithm (also tried "Orientation towards Paradigm Words"). For $k = 10$ they obtained correlations for valence, arousal, and dominance of .91, .84, and .85. This performance is considerably better than was achieved by Mandera et al. (2015) for English using corpus derived word similarity. These results highlight the sensitive differences between word analogy tasks for human readers and the information extracted by vector space embedding methods.

Rothe, S., S. Ebert, and H. Schütze (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Rothe et al. (2016) transform the embedding space of works via optimization of certain dimensions onto known semantic properties. This amounts to reducing the 300 or so dimensions typically used for vector space embedding into less than three dimensions. They apply Stochastic Gradient Descent (SGD) to learn a transformation Q that orthogonalizes the embedding matrix A under the constraint of establishing a sentiment dimension. This approach is more successful than embedding words directly into such a low dimension space, agreeing with previous work that has show vector embedding performs best with more than 100 dimensions, while extracting the relevant semantic information for sentiment analysis. For lexicon creation, their approach labeled "Densifier" achieves the statistically significant best performance on SemEval 2015 Task 10E with Kendall's τ of .654.

Altogether, these approaches provide a roadmap and demonstrate the possibility of constructing a high-quality, general purpose, phrase based sentiment dictionary.

1.2.6 Visualization

Lacking from the bulk of research that applies sentiment analysis, but crucial for validation and understanding, is visualization of sentiment analysis. Despite limited attempts by researchers in sentiment analysis to use visualization to understand their analysis, online tools have been built to allow anyone to build simple visualizations in a straightforward way (Viegas et al., 2007). Motivation for our choice of dictionary-based methods along with a straightforward averaging algorithm for generating scores is that the analysis can be visualized to be understood. The averaging algorithm is linear and this allows for the comparison of the individual word contributions to text sentiment classification, both enabling greater understanding and validating the analysis.

An overview of previous approaches to text visualization can be found in Heer (2014) and Cao and Cui (2016). We note the four goals of text visualization as identified by Heer: understanding, comparison, grouping, and correlation. Here, we focus on the task of understanding. A selection of recent work that builds on this task is available from Hearst (2009) (Chapter 11), Chuang et al. (2012), Van Ham et al. (2009), and Chuang et al. (2012).

Visualizations of readable portions of text are able to communicate the results of analysis at that level, such as the syntactic parse visualization in 1.4. On a sentence level, we can see with or

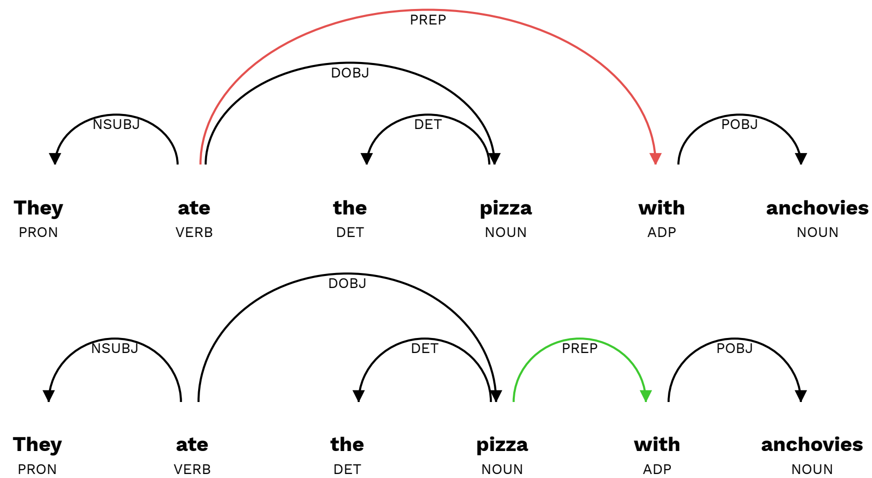


Figure 1.4: Visualization of a syntactic dependency parse with the displaCy tool from Honnibal (2015), a companion to the spaCy package for NLP in Python. The tool doubles as an annotation tool with key-based input for efficient manual dependency tagging.

without added visual clues (e.g., colored backgrounds or font size) which individual words have either positive or negative scores, and how their balance contributes to the average-based classification. When rules become involved, this process is more complicated and it may be necessary to utilize a sentence diagram to understand the classification at even the individual sentence level. Neither of these approaches scale to visualize more than individual sentences, a fundamental shortcoming in working with big data.

Next, we examine tag clouds as a tool to understand text and the results of text analysis.

Tag clouds

Tag clouds are a popular method for displaying the results of text analysis, with the size of text being used to represent one variable from the analysis and the layout of words with random locations, angles, and color, generally positioned to minimize white space. Various attempts have been made to assess the efficacy of tag clouds compared to more traditional statistical information visualizations such as bar charts with a consensus that they are less effective, though aesthetically pleasing: see Halvey and Keane (2007), Rivadeneira et al. (2007), and Hearst and Rosner (2008). One popular package for producing word clouds layouts is “Wordle” from Feinberg (2009).

Since tag clouds by wordle have random layouts, improvements that incorporate relevant information into the layout itself have been considered. In Schrammel et al. (2009) they compare the performance and likability of four approaches: alphabetic, random, similarity on Flickr, and distance in WordNet. From 64 participants, they find that "semantically clustered tag clouds can provide improvements over random layouts in specific search tasks and that they tend to increase the attention towards tags in small fonts compared to other layouts".

In Lohmann et al. (2009) tag cloud layouts are compared on three tags and results show that there is no single best layout. The three tasks they test and the best layout for each are:

- Finding a specific tag: Sequential layout with alphabetical sorting.
- Finding the most popular tags: Circular layout with decreasing popularity.
- Finding tags that belong to a certain topic: Thematically clustered layout.

It is also confirmed using eye tracking that tag clouds are scanned (not read), attention is focused on the center of the tag cloud, and they all perform sub-optimally for looking up specific words.

A study of the social (non-academic) use of Wordle is done by Viegas et al. (2009), finding that the existence of tools for building custom Wordles was crucial to their popularity and that 35/49% of men/women under the age of 20 did not know that frequency of usage is used for the font size.

Adding a time component to tag clouds with the use of "sparklines", Lee et al. (2010) find that SparkClouds are able to communicate trends as well. New layouts attempt to incorporate additional information to tag clouds through layout and color, such as the TAGGLE system of Emerson et al. (2015).

Moving beyond tag clouds, we briefly present word shift graphs in the next section.

Word shift graphs

An indispensable, scientific tool for visualizing text analysis is the word shift graph. The graph was first designed and put to use by Dodds and Danforth (2009) to understand the result of sentiment analysis. An online, interactive version of the graphs are used widely at hedonometer.org, and more details on the use of these graphs is available at compstorylab.org. The important difference between the word shift graph and tag cloud is that the word shift graph uses both spatial dimensions meaningfully, encoding the ranking of words in the vertical direction and the relevant statistical value

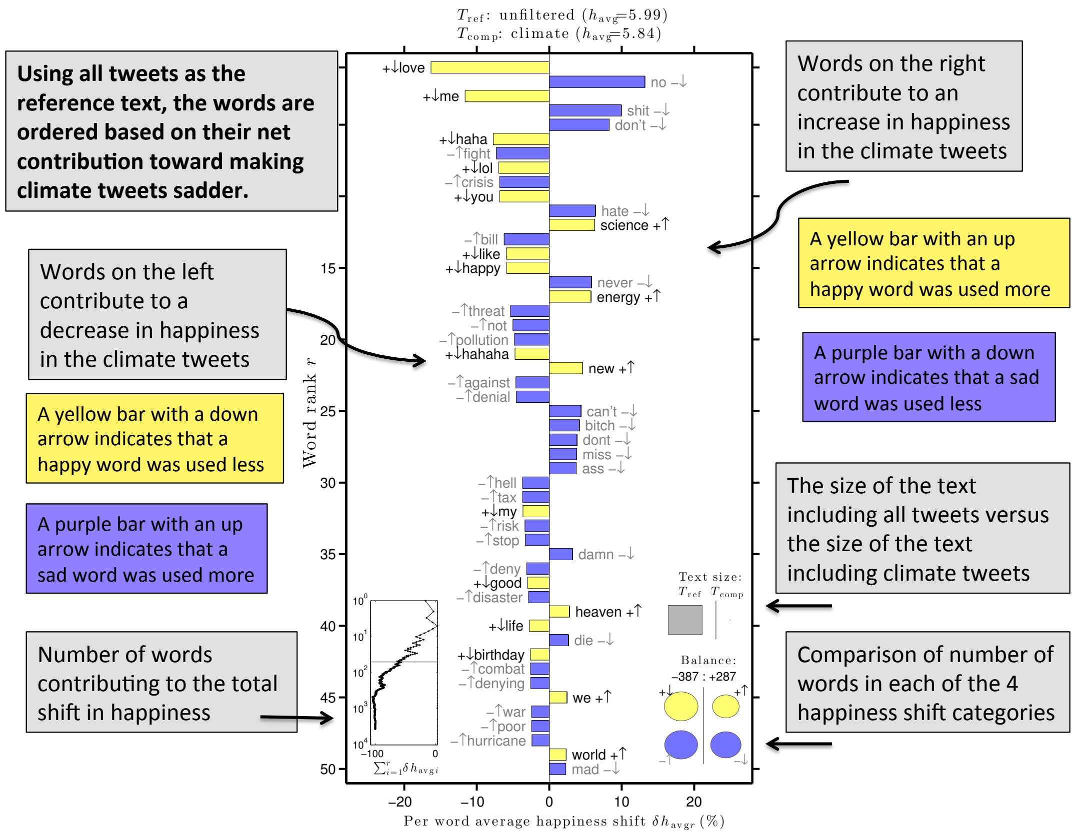


Figure 1.5: We quote the following caption and re-use the figure from Cody et al. (2015): A word shift graph comparing the happiness of tweets containing the word "climate" to all unfiltered tweets. The reference text is roughly 100 billion tweets from September 2008 to July 2014. The comparison text is tweets containing the word "climate" from September 2008 to July 2014. A yellow bar indicates a word with an above average happiness score. A purple bar indicates a word with below average happiness score. A down arrow indicates that this word is used less within tweets containing the word "climate". An up arrow indicates that this word is used more within tweets containing the word "climate". Words on the left side of the graph are contributing to making the comparison text (climate tweets) less happy. Words on the right side of the graph are contributing to making the comparison text more happy. The small plot in the lower left corner shows how the individual words contribute to the total shift in happiness. The gray squares in the lower right corner compare the sizes of the two texts, roughly 10^7 vs 10^{12} words. The circles in the lower right corner indicate how many happy words were used more or less and how many sad words were used more or less in the comparison text.

in the horizontal direction, enabling comparison between the values. We present a closer examination of an example word shift graph in Figure 1.5.

We elaborate more on the construction, present use cases where the word shift graph helps us understand successes and failures of sentiment analysis, and generally make extensive use of the word shift graph as a tool in Chapter 3. A future effort could aim to assess the efficacy of the word shift graph for text-based research, by performing a task-level user study.

1.2.7 Benchmarking literature

In this section we review recent efforts to benchmark sentiment analysis methods for their performance.

Liu, B. (2012, May). *Sentiment analysis and opinion mining.* Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.
This book from Bing Liu provides a broad overview of sentiment analysis, and the many different problems that it hopes to address as well as summaries of many common approaches. Liu provides a framework to understand the aspects of sentiment analysis, with the levels of analysis (aspect, sentence, document level), and goals including classification and opinion summarization. In Chapter 8, a discussions of the methods for generating sentiment dictionaries is presented, and includes manual, dictionary-based, and corpus based approaches. Survey methods are not considered (the well-known ANEW dictionary is absent), and there is some confusion between methods that use a dictionary to propagate scores and those that use features of a corpus to propagate scores (Velikovich et al. (2010) incorrectly classified as the former). While the references are extensive, no analysis is conducted to understand how the different approaches for generating sentiment dictionaries perform. Despite these shortcomings, the book is a broad and very useful guide to the landscape of sentiment analysis.

Hutto, C. J. and E. Gilbert (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media.*
This paper is focused on the development of a new dictionary-based method for sentiment analysis that incorporates a rule-based system and a dictionary tailored to social media. While other papers that introduce dictionaries for sentiment analysis have made comparisons between methods (e.g., LIWC correlations between the 2001, 2007, and 2015 dictionaries on their website), we include this as a benchmark because of the uncommon rigor in the comparisons made. In

particular, Hutto and Gilbert compare their new method VADER to 11 other sentiment analysis methodologies. They compare to seven dictionary-based methods and four ML methods, and find favorable correlations between the classification of Tweets for the dictionary based methods. In addition they perform tests to measure the performance gains to be had using four rules, and word sense disambiguation, finding mean F1 performance gains of 2 points on individual Tweets. These rules are a subset of those employed by Thelwall et al. (2012). The comparisons between sentiment dictionaries focus on the classification performance, and do not provide any insight into what properties of the dictionaries contributes to their performance. In addition, no effort is made to use sentiment analysis as more than a binary classifier, a shortcoming that we will address.

Giachanou, A. and F. Crestani (2016, June). Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv. 49*(2), 28:1–28:41.

This extensive survey from Giachanou *et al.*provides an overview and categorization of methods used to quantify sentiment on Twitter. No quantitative comparisons are made between the methods themselves. The broad categories of the methods they find are based on those from Liu (2012):

- Machine Learning.
- Lexicon-Based.
- Hybrid (Machine Learning & Lexicon-Based).
- Graph-Based.

The focus is on ML approaches (as they note: "The majority of [Twitter Sentiment Analysis] methods use a method from the field of machine learning").

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto (2016, jul). SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci. 5*(1), 23.

This recent benchmark from Ribeiro *et al.*was published while our work was under review, having been submitted after our preprint was released on the arXiv. The comparisons made by Ribeiro *et al.*utilize a variety of methods, and provide measures of performance for all methods based on F1 scores. The methods selected include commercial, ML, and dictionary-based, and they are applied for four corpora. Beyond metrics of classification performance, no insight is provided into the reasons why certain methods out-perform others, nor is any focus on understanding texts through sentiment (or using visualization), the key tenets of our effort in Chapter 2.

1.3 Emotional arcs

Stories provide a useful framing to condense our experience, and through this they are both ubiquitous and powerful. In 2011, a DARPA initiative "Narrative Networks" (DARPA, 2011) said the following in relation to security:

> Narratives exert a powerful influence on human thoughts and behavior. They consolidate memory, shape emotions, cue heuristics and biases in judgment, influence in-group/out-group distinctions, and may affect the fundamental contents of personal identity. It comes as no surprise that because of these influences stories are important in security contexts: for example, they change the course of insurgencies, frame negotiations, play a role in political radicalization, influence the methods and goals of violent social movements, and likely play a role in clinical conditions important to the military such as post-traumatic stress disorder.

The ubiquitous nature of stories is summed up well in Dodds (2013):

> We humans are storytelling and story-finding machines: *homo narrativus*, if you will. In making sense of the world, we look for the shapes of meaningful narratives in everything. Even in science, we enjoy mathematical equations and algorithms because they are a kind of universal story. Fluids—the oceans and atmosphere, the blood in your body, honey—all flow according to a single, beautiful set of equations called the Navier-Stokes equations.
>
> In our everyday, human stories, far away from science, we have a limited (if generous) capacity to entertain randomness—we are certainly not *homo probabilisticus*. Too many coincidences in a movie or book will render it unbelievable and unpalatable. We would think to ourselves, "that would never happen in real life!" This skews our stories. We tend to find or create story threads where there are none. While it can sometimes be useful to err on the side of causality, the fact remains that our tendency toward teleological explanations often oversteps the evidence.

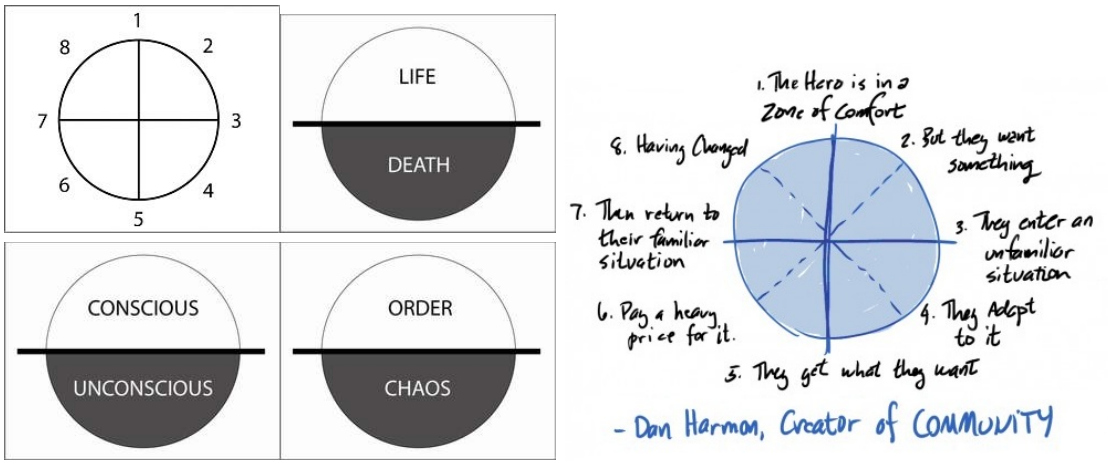


Figure 1.6: Harmon cycles with and without labels, as used to develop the show Community. The cyclical nature of the story has roots in the "monomyth" of Campbell (1949).

In Chapter 3 we consider previous work that finds between one and 36 different plot types: Campbell (1949); Harris (1959); Abbott (2008); Booker (2006); Polti (1921). Of these, the work of Campbell and Moyers (1991) has gained popular attention as a result of the expositions of Dan Harmon in writing the show *Community* (Raftery, 2011). In a series of online posts, Harmon elaborates on the "monomyth" and its incorporation into the writing of the *Star Wars* movies (Volger, 1992). The plot here is cyclical, and therefore represented on a circle, and the argument goes that all well constructed plots can be arranged to fit into this mold. The basic circle consists of 8 locations, starting and ending in the same place, and show a labeled visualization of these locations in Figure 1.6.

Lacking from the existing work considering theories of plot is a strong grounding in empirical evidence or stability of the "universal" theories across culture. It is precisely this shortcoming which we hope to address, by using a broad collection of Fiction stories within western culture.

1.3.1 Story graphs, plot diagrams, and inferring causality

With the distinction between plot, structure, and emotional trajectory in mind, there have also been attempts to discover plot using data driven methods. Brewer and Lichtenstein (1980) makes the distinction between plot and structure is made even clearer. Through experimentation with different structures, Brewer and Lichtenstein find that the resulting affect in readers is different, with some structures being considered stories and others not (the authors single out "suspense and resolution" and "surprise and resolution" as indicative of stories).

Plot units were first introduced by Lehnert (1981), and form the basis for most all efforts that follow.

Using topic modeling, both Schmidt (2015b) and Jockers (2013) find known patterns of plot across many thousand stories. In Piper (2015a), computational analysis is applied to realize the potential of distant reading (a term owing to Moretti (2013)) to find and test scholarly insights. In Winston (2011), a system called "Genesis" is developed to compare plot summaries and infer causal connections between events, with the broad aim of the system formalized as the *Strong Story Hypothesis*:

> The mechanisms that enable humans to tell, understand, and recombine stories separate human intelligence from that of other primates.

In his Master's Thesis, Awad (2013) extends the Genesis system to model differences in American and Chinese stories by adding commonsense rules that differ between cultures. With commonsense rules, *Genesis* is able to measure story coherence.

Work by Regneri et al. (2010) learns event scripts from written descriptions of events that may not always exist in written form (implicit scripts, like shopping), using a graph-based ("temporal script graph") algorithm and data collected on Amazon's Mechanical Turk. The algorithm is tested to detect similar events with differing descriptions.

The Analogical Story Merging (ASM) system is developed using "Bayesian model merging" for story categorization and is applied to 15 Russian folktales (Finlayson, 2011). The test folktales are annotated for 18 aspects of meaning by 12 annotators using a tool developed for this task. The folktale categories defined by Vladmir Propp are predicted by ASM and the system achieves a Rand Index of 0.511 (a measure of the similarity between clusters).

In Elson (2012a) a Story Intention Graph (SIG) is developed to model stories and implemented to measure similarity and analogy. Elson's propositional similarity metric is used to predict human judgments of story similarity and outperforms human annotation (is better than inter-annotator agreement) on Aesop's fables.

The AESOP system of Goyal et al. (2013) converts narrative texts into their plot unit model (where plot units are "conceptual knowledge structure to represent the affect states of and emotional tensions between characters in narrative stories"). AESOP performs four steps: "affect state recognition, character identification, affect state projection, and link creation." Performance is inspected on a set of Aesop's fables, similar to Elson (2012a).

In *Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel*, Piper (2015a) applies Multi-Dimensional Scaling (MDS) on representations of novels in a VSM (Vector Space Model — vectors of word frequencies), and performs hierarchical clustering to understand the differences between novels and autobiographies.

1.3.2 Story generation

In *Plot Induction and Evolutionary Search for Story Generation*, McIntyre and Lapata (2010) build upon their previous work to train a story planner from extracted events, their participants, and preceding relationships from a large corpus. Their system is used to to generate simple, 4 or 5 sentences stories that are mildly coherent.

The Neukom Institute at Dartmouth hosts a competition for algorithms to produce short stories, in a true-fashion Turing test (Neukom Institute, 2016). In the 2016 competition, algorithms and writers were given a one-word prompt and tasked to write a 500-word short story. The stories were then judged by a panel consisting of David Cope, Lynn Neary, and David Krakauer to be either human or machine written. Each judge received 8 human written stories and 3 machine generated stories, one from each of the 3 entrants into the competition. To quote their results:

> No machine won, but one submission generated by Toksu and Ibrahim on the seed "thesaurus" "fooled" one of the judges!

With no first place award, the second place award and $1000 prize was awarded to Judy Malloy whose algorithm rearranged sentences from "Another Party in Woodside".

1.3.3 Character Identification and Networks

Much work on computational understanding of stories has focused on the extraction and analysis of character networks. The ideas behind character networks were first examined in the original work of Moretti (Moretti, 2000, 2007; Schulz, 2011; Moretti, 2013), and have been used widely in Digital Humanities research. Below we highlight work that has caught our attention.

Elson et al. (2010) use character name chunking, quoted speech attribution and conversation detection to generate character networks from a collection of British novels. They find a lack of support for characterizations provided by literacy scholars and suggest an alternative explanation. Namely, the do not find support for the hypothesis that 19th century fiction novels have (1) social networks that differ by the setting of the novel (rural vs. urban) and that (2) novels with more characters have less dialogue (an inverse relationship is suggested by the so-called "chronotype" theories). Instead they find that the point of view of narration (first vs. third person) is strongly correlated with the This work applies the distant reading philosophy by first carefully selecting a corpus of books and consulting previous literary research before doing analysis, an approach we aspire to emulate. Elson later extended this work with models of discourse (Elson, 2012b).

Bamman *et al.*use Bayesian models, word embedding, and state-of-the-art NLP techniques to learn personas of characters in literature (Bamman et al., 2014) and in film (Bamman et al., 2014). Their analysis is performed across a large corpora of 15,099 books selected from Hathitrust, 42,306 wikipedia movie plot summaries for film, and is shown to replicate the classification of character roles by a literary scholar. A similar effort is undertaken by Valls-Vargas et al. (2014), utilizing PoS annotations from syntactic parsing to detect characters in a small set of stories, and using "action matrices" in another attempt (Valls-Vargas et al., 2014) to encode Propp's narrative theory. They are able to automatically detect the roles of characters within 10 folktales (developing a system they refer to as "Voz").

These methods have also been used to examine popular culture. In a blog post, Gabasova (2015) finds the most central character in Star Wars. Xanthos et al. (2016) elaborate on the method of constructing and visualizing character networks, an example of their work for Shakespeare is available as a poster: http://www.martingrandjean.ch/network-visualization-shakespeare/. Min and Park (2016) perform an in-depth study of Victor Hugo's *Les Misérables*, proposing using the growth of edges in and characters in the network over time to compare different works (with each edge/character curve normalized to sum to 1 at the end of each book). More recently, Wu (2016) has made an interactive exploration of the play Hamilton using discourse and the character network, and Meeks and Averick built an interactive exploration of the dialogue in the show Archer (Meeks and Averick, Meeks and Averick)

To compare character networks across movies, Ruths (2016) uses network alignment to map characters between the *Stars Wars* movies *The Force Awakens* and *A New Hope* revealing both expected and surprising similarities. For example R2-D2 maps to BB-8 and Chewbacca maps to Chewbacca, as we might expect. However, the main characters have more surprising alignments from the interaction networks, with Luke mapping to Poe, Obi-wan mapping to Kylo Ren, and Darth Vader mapping Rey. A particular problem in using character networks that span an entire movie, TV show, or book is that multiple story lines can intersect in ways that are not accounted for by the method. Bost et al. (2016) examine conversation in TV shows using a smoothing of narration to overcome the multiple narrative problem, finding protagonists more readily than using simpler interaction networks.

1.3.4 Frames for NLP

The seminal work by Schank and Abelson (1977) (and earlier efforts by Rumelhart (1975)) laid the groundwork for scripts as a framework for cognitive algorithmic computation. Research programs separately advancing AI capabilities and NLP tasks have made use of this framework. Although existing knowledge bases such as SUMO (Niles and Pease, 2001), Cyc (Lenat, 1995) or FrameNet (Fillmore et al., 2003) contain such script-like knowledge to a certain extent, their coverage is severely limited. Increases in computational power have realized the building of systems for script-based event detection, and there have been many efforts made in the past decade to advance such systems. Schemata such as NarrativeML to annotate narratives are reviewed by Mani (2012). Next, we very briefly highlight some of these approaches, focusing particularly on the research program of Chambers due to the accessibility of the papers and the breadth of research by himself and his students.

In a series of papers Chambers et al. (2007); Chambers and Jurafsky (2008, 2009, 2010); Chambers (2013) set to classify temporal relations between events, apply unsupervised learning to detect narrative event chains and entities involved, build a database of narrative schemata, and find schemata in large corpora with probability-based models. A narrative event chain is defined as two events linked by a common actor. Event chains are identified in text through co-reference between a single entity, ordered by a trained classifier, and all possible event chains are restricted through a clustering approach in Chambers et al. (2007); Chambers and Jurafsky (2008). Both Cheung et al. (2013) (using the proposed approach of O'Connor (2013)) and Chambers (2013) utilize generative models for inducing event schemata, with the former utilizing a HMM over latent event variables and the latter using a entity-driven model. Recent work from Pichotta and Mooney (2015) improves on the baseline results of Chambers in detecting scripts using Recurrent Neural Networks (RNNs, particularly a flavor known as Long Short Term Memory (LSTM)) and architectures adapted to this task.

Corpora used by Chambers and by others include the FrameNet from Baker et al. (1998), Timebank Corpus from Pustejovsky et al. (2003), Opinion Corpus from Mani et al. (2006), Narrative Schema Database from Chambers and Jurafsky (2010), the Media Frames Corpus by Card et al. (2015), and most recently the Story Cloze Dataset from Mostafazadeh et al. (2016). As an example, Do et al. (2011) use a primarily unsupervised approach to specifically learn causality between events in the Penn Discourse Treebank, and Roemmele et al. (2017) use an RNN on the Story Cloze dataset. The understanding and generation of stories with these data sets and new models may hold promise for major advances in the field of NLP. Cambria and White (2014) has suggested that the next wave of NLP advances that aim to decode stories (a move from "bag of words" approaches to "bag of narratives") may very well be a breakthrough in understanding human nature.

Along those lines, stories have been explored as a model to training artificial intelligence systems for commonsense reasoning. Advanced in this area all recognize and leverage the utility of stories for sense-making (Bex and Bench-Capon, 2010; Bex, 2013; Li et al., 2012; Riedl, 2016).

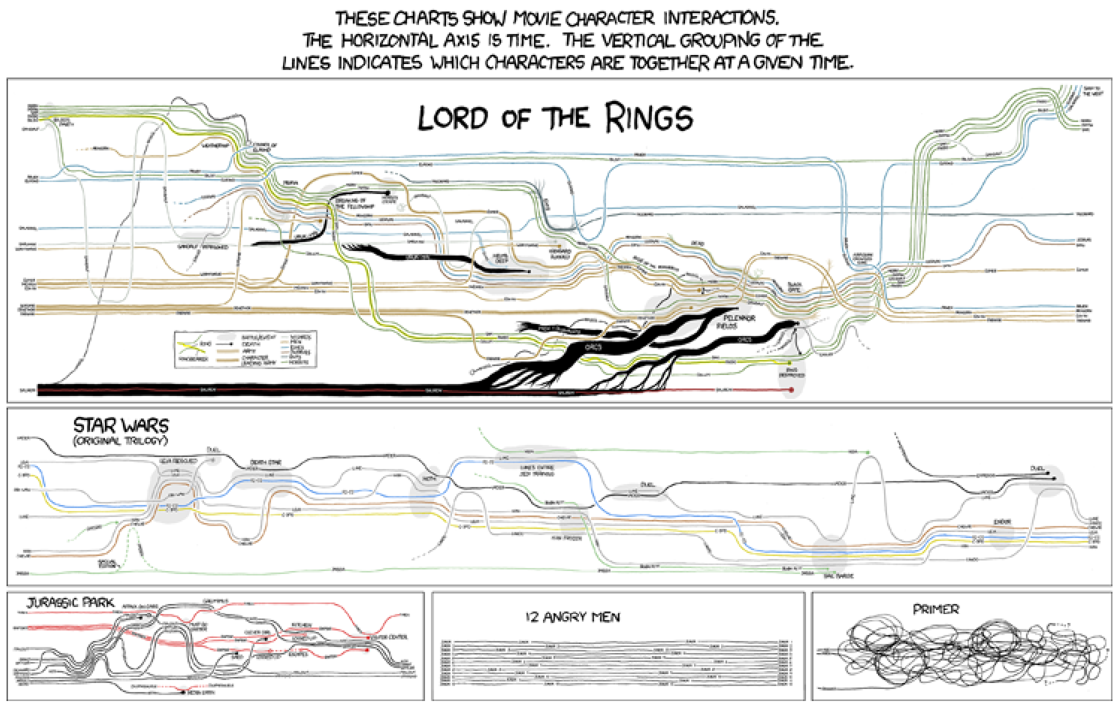


Figure 1.7: XKCD number 657 by Munroe (2009) shows the time evolution of character co-occurrence in Lord of the Rings, Star Wars, Jurassic Park, 12 Angry Men, and Primer. Munroe adds: "in the LotR map, up and down correspond LOOSELY to northwest and southeast respectively." The width of lines correspond to the number of characters in each group, which applies here to the Orcs in Lord of the Rings.

1.3.5 Visualization

Stories as a model for understanding are not readily visualized, as finding a proper encoding for the mental models we use is difficult. Nevertheless, efforts at capturing the essence of story in a visual form are omnipresent in art and automated attempts to generate such mappings are attempted (recall Figure 1.6). The illustrated movie maps of DeGraff and Harmon (2015) make representations of movies in the limited space of two pages by using three dimensions to show the movement of time and place. The web comic XKCD draws inspiration from the well-known visualization of Napoleon's march by Minard and maps the interaction of characters with time as a x-axis and character proximity as distance in the y-axis of a chart, see Munroe (2009) and Figure 1.7. Ogawa and Ma (2010) attempt to automatically build XKCD-style plots for software development, and an image of their reproduction of the XKCD *Lord of the Rings* visualization is shown by Cao and Cui (2016).

1.3.6 Emotional arcs

The emotional arcs drawn by Vonnegut (1981) are simpler, using time again on the x-axis and representing the fortune of the main character in the vertical direction. Vonnegut explicitly draws a connection between the New Testament and Cinderella, a story that has incredible popular appeal. Other story arcs named by Vonnegut are the "Man in the Hole" and the "Boy meets girl" arcs.

With the same goal of finding commonalities between stories as Vonnegut (1981), in a series of blog posts Jockers (2014) lays out a strategy for generating emotional arcs and eventually finds six story types using hierarchical clustering. Our work in Chapter 3 is an continuation of a very different core methodology that we first propose in Dodds et al. (2015a). Though the core methodology is markedly different, we note that Jocker's first blog post appeared 10 days before the pre-print of our paper As we note in Chapter 3 as well, the distinction between plot and emotional arc as well as correct use of using sentiment analysis tools distinguish our contributions from those of Jockers (2014).

Attempts to analyze plot more directly than emotional arc have been increasing in the past few years. Cherny (2016) applies machine learning over a bag-of-words analysis to predict action and sex scenes using Naive Bayes (NB) and Stochastic Gradient Descent (SGD). Training data is crowd-sourced from two ratings of 500 word chunks on the survey platform Mechanical Turk (MT), and Cherny develops novel visualizations of the relationships between topics in chapters. Reiter et al. (2014) use an unsupervised method to generate and compare event-based representations of rituals and folktales, but we were unable to obtain their manuscript. Piper (2015a) analyzes the differences between the first and second half of novels about "conversion." We revisit the approach by Schmidt (2015b) here: he uses Latent Semantic Analysis (LSA) and dimensionality reduction to find patterns of plot in a reduced 2-dimensional topic space. While this is an interesting approach, we would not expect the coefficients of the first two modes in the SVD to hold particular relationships between themselves. Most recently, the approach of measuring sentiment using sentences and smoothing has been published by Gao et al. (2016).

The most similar approach to ours (perhaps based on our method from Dodds et al. (2015a), though they cite Vonnegut) was an effort by sentiment analysis startup Indico's Dan Kuster, available at https://indico.io/blog/plotlines/ (Kuster, 2015). Kuster uses sliding windows and dynamic time warping as a distance metric between emotional arcs, and on single books the method is indeed very similar to ours, yet they don't extend to mine for patterns across a large corpus.

1.3.7 Suzyhet and validation

The work of Jockers (2014) has been publicly debated in the online sphere. The back-and-forth between Matt Jockers and Annie Swafford (and others) has happened in blog posts (Swafford, 2015), comments on blogs, and on Twitter. The extent of this debate is documented in two parts by Clancy (2015) (available online: https://storify.com/clancynewyork/contretemps-a-syuzhet and https://storify.com/clancynewyork/a-fabula-of-syuzhet-ii). We attempt to briefly summarize some of the discussion of prominent scholars in digital humanities and how this relates to our own work on emotional arcs, particularly the comments of Bamman, Piper, Schmidt, Enderle, and Underwood.

Bamman (2015) elaborates on the discussion around on how to measure validity of emotional arcs Bamman (2015) goes on to build a survey to perform the validation proposed by Piper (2015b) and Weingart (Weingart). Bamman's survey for Shakespeare's Romeo and Juliet takes responses from 5 participants on Mechanical Turk for each scene on a -5 to 5 scale along with a free text reasoning for the score. We plot the mean of these ratings along with our measure of the emotional arc (the happiness of the words in the play for a sliding window of 10000 words and 200 time points) of the play in Figure 1.8. This approach could, of course, be extended to provide additional formal validation of the methods and parameters used in our study of emotional arcs. However, non-expert annotations are not always a proper gold-standard (Snow et al., 2008), and there may even be (we might even expect) valid interpretations of a story that produce different emotional responses. In this case, we would expect that our automated method would find one of these arcs, and the goal of a more advanced system could be to find more than one arc for a given book.

In addition to the problems identified by Swafford, Schmidt (2015a) builds on Enderle (2015) and highlights the problem that the low pass filter needs to be circular. These discussions have provided many interesting future directions for this work and the validation of computational approaches to narratives.

Our own work on emotional arcs (Chapter 3) has attracted a great deal of popular attention and has been noticed by those in the digital humanities community, particularly by Schmidt (2016) and

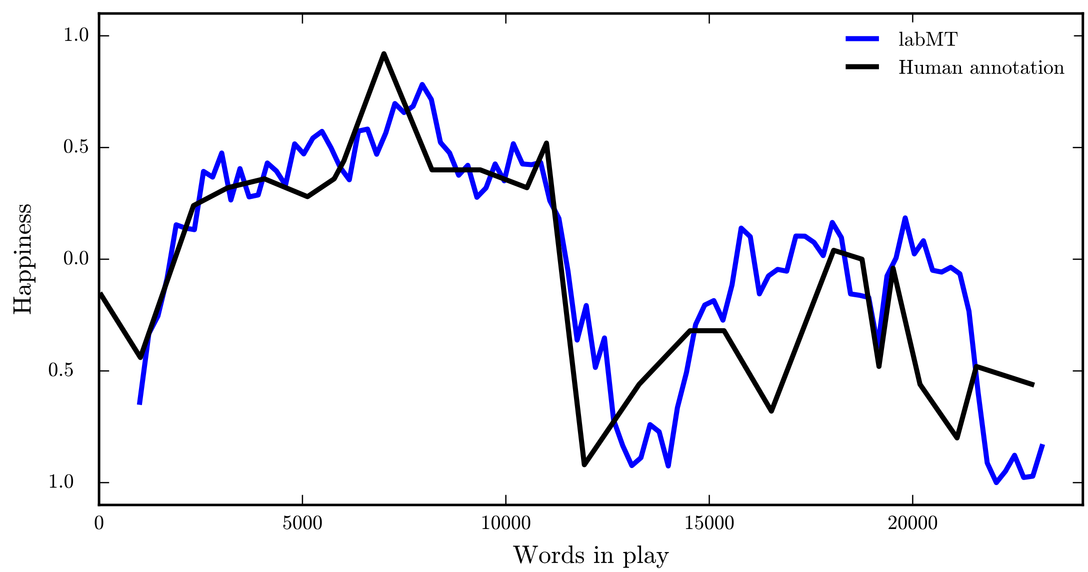


Figure 1.8: Emotional arcs of Shakespeare's Romeo and Juliet, generated with the labMT sentiment dictionary and the average of 5 human annotations on each scene. The labMT approach generated 100 time points, with 2000 rated words at each point shown, $\delta_h = 1$, and ignoring scene boundaries (the same approach used in general). The human annotation data is from a survey conducted in Bamman (2015) with 5 responses for each of the 26 scenes in the play, points are shown on the x-axis in the center of each scene's words. The survey collected responses from -5 to 5, which we have re-scaled linearly to -1 to 1 (by dividing by 5), and the labMT data is re-scaled by first mean centering the time series, then multiplying by the inverse of the absolute maximum (such that the time series will touch -1 or 1 in the direction of the absolute maximum).

Enderle (2016). We address the concerns raised in both of these critical takes in our work. Drawing directly from the suggestions from Schmidt, we utilize the Library of Congress classification to produce a better selection of texts from Project Gutenberg in our published manuscript, a notable improvement from the pre-print corpus he analyzes. In our treatment, we carefully consider the choice of a suitable null hypothesis to test whether there is structure in the emotional arcs of real stories. Our first pass used the emotional arcs of the same books with randomly shuffled words ("word salad" books), for a corpora that has no narrative structure but the same emotional words. The final version of our null model generates stories from a bigram Markov chain trained on the actual text. These "nonsense" narratives have no real structure, but resemble written English. For more complete details and sample text from each method, see Section B.3. Other reasonable attempts could consider shuffling sentences of paragraphs, however Brownian noise and arbitrary random walks are not sensible comparisons. In particular, the singular value spectrum of Brownian noise is arbitrary.

In the next Chapter, we test sentiment analysis methods for performance in classification and providing understanding of emotional text, methods that form the basis of our study into emotional arcs.

Chapter 2

Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs.

The emergence and global adoption of social media has rendered possible the real-time estimation of population-scale sentiment, which has profound implications for our understanding of human behavior. Given the growing assortment of sentiment-measuring instruments, it is imperative to understand which aspects of sentiment dictionaries contribute to both their classification accuracy and their ability to provide richer understanding of texts. Here, we perform detailed, quantitative tests and qualitative assessments of 6 dictionary-based methods applied to 4 different corpora, and briefly examine a further 20 methods. We show that while inappropriate for sentences, dictionary-based methods are generally robust in their classification accuracy for longer texts. Most importantly they can aid understanding of texts with reliable and meaningful word shift graphs if (1) the dictionary covers a sufficiently large portion of a given text's lexicon when weighted by word usage frequency; and (2) words are scored on a continuous scale.

2.1 Introduction

As we move further into what might be called the Sociotechnocene—with increasingly more interactions, decisions, and impact being made by globally distributed people and algorithms—the myriad human social dynamics that have shaped our history have become far more visible and measurable than ever before. Driven by the broad implications of being able to characterize social systems in microscopic detail, sentiment detection for populations at all scales has become a prominent research arena. Attempts to leverage online expression for sentiment mining include prediction of stock markets (Bollen et al., 2011; Si et al., 2013; Chung and Liu, 2011; Ruiz et al., 2012), assessing responses to advertising, real-time monitoring of global happiness (Dodds et al., 2015a), and measuring a health-related quality of life (Alajajian et al., 2016). The diverse set of instruments produced by this work now provide indicators that help scientists understand collective behavior, inform public policy makers, and, in industry, gauge the sentiment of public response to marketing campaigns. Given their widespread usage and potential to influence social systems, understanding how these instruments perform and how they compare with each other has become imperative. Benchmarking their ability to provide insight into sentiment, and their performance, both focuses future development and provides practical advice to non-experts in selecting a sentiment dictionary.

We identify sentiment detection methods as belonging to one of three categories, each carrying their own advantages and disadvantages:

1. Dictionary-based methods (Dodds et al., 2015a; Bradley and Lang, 1999; Pennebaker et al., 2001; Wilson et al., 2005; Liu, 2010; Warriner et al., 2013),
2. Supervised learning methods (Liu, 2010), and
3. Unsupervised (or deep) learning methods (Socher et al., 2013).

Here, we focus on dictionary-based methods, which all center around the determination of a text T's average happiness (sometimes referred to as *valence*) with sentiment dictionary D through the equation:

$$h_{\mathrm{D}}^{T} = \frac{\sum_{w \in D} h_{\mathrm{D}}(w) \cdot f^{T}(w)}{\sum_{w \in D} f^{T}(w)} = \sum_{w \in D} h_{\mathrm{D}}(w) \cdot p^{T}(w), \tag{2.1}$$

where we denote each of the words in a given sentiment dictionary D as words w, word sentiment scores as $h_{\mathrm{D}}(w)$, word frequency as $f^T(w)$, and normalized frequency of w in T as $p^T(w) = f^T(w)/\sum_{w\in D} f^T(w)$. In this way, we measure the happiness of a text in a manner analogous to taking the temperature of a room. While other simple sentiment metrics may be considered, we will see that analyzing individual word contributions is important and that this equation allows for a straightforward, meaningful interpretation.

Dictionary-based methods offer two distinct advantages which we find necessary: (1) they are in principle corpus agnostic (applicable to corpora without ground truth data available) and (2) in contrast to black box (highly non-linear) methods, they offer the ability to "look under the hood" at words contributing to a particular score through *word shift graphs* (defined fully later; see also (Dodds and Danforth, 2009; Dodds et al., 2011)). Indeed, if we are at all concerned with understanding why a particular scoring method varies—e.g,, our undertaking is scientific—then word shift graphs are essential tools. In the absence of word shift graphs, or similar devices, any explanation of sentiment trends is missing crucial information and rises only to the level of opinion or guesswork (Golder and Macy, 2011; Garcia et al., 2015; Dodds et al., 2015b; Wojcik et al., 2015).

As all methods must, dictionary-based "bag-of-words" approaches suffer from various drawbacks, and three are worth stating up front. First, they are only applicable to corpora of sufficient size, well beyond that of a single sentence (Ribeiro et al., 2016) (widespread usage in this misplaced fashion does not suffice as a counterargument). We directly verify this assertion on individual Tweets, finding that some sentiment dictionaries perform admirably, however the average (median) F1-score on the STS-Gold data set is 0.50 (0.54) from all datasets (Table A.1), others having shown similar results for dictionary methods with short text (Ribeiro et al., 2016). Second, state-of-the-art learning methods with a sufficiently large training set for a specific corpus will outperform dictionary-based methods on same corpus (Liu, 2012). However, in practice the domains and topics to which sentiment analysis are applied are highly varied, such that training to a high degree of specificity for a single corpus may not be practical and, from a scientific standpoint, will severely constrain attempts to detect and understand universal patterns. Third, words may be evaluated out of context or with the wrong sense. A simple example is the word "miss" occurring frequently when evaluating articles in the Society section of the New York Times. This kind of contextual error is something we can readily identify and correct for through word shift graphs, but would remain hidden to users of nonlinear learning methods.

We lay out our paper as follows. We list and describe the dictionary-based methods we consider in Sec. Dictionaries, Corpora, and Word Shift Graphs, and outline the corpora we use for tests in Subsec. Corpora Tested. We present our results in Sec. Results, comparing all methods in how they perform for specific analyses of the New York Times (NYT) (Subsec. New York Times Word Shift Analysis), movie reviews (Subsec. Movie Reviews Classification and Word Shift Analysis), Google Books (Subsec. Google Books Time Series and Word Shift Analysis), and Twitter (Subsec. Twitter Time Series Analysis). In Subsec. Brief Comparison to Machine Learning Methods, we make some basic comparisons between dictionary-based methods and machine learning approaches. We provide concluding remarks in Sec. Conclusion and bolster our findings with figures, tables, and additional analysis in the Supporting Information.

2.2 Sentiment Dictionaries, Corpora, and Word Shift Graphs

2.2.1 Sentiment Dictionaries

The words "sentiment dictionary," "lexicon," and "corpus" are often used interchangeably, and for clarity we define our usage as follows.

Sentiment Dictionary: Set of words (possibly including word stems) with ratings.

Corpus: Collection of texts which we seek to analyze.

Lexicon: The words contained within a corpus (often said to be "tokenized").

We test the following six sentiment dictionaries in depth:

labMT — language assessment by Mechanical Turk (Dodds et al., 2015a).

ANEW — Affective Norms of English Words (Bradley and Lang, 1999).

WK — Warriner and Kuperman rated words from SUBTLEX by Mechanical Turk (Warriner et al., 2013).

Dictionary	# Entries	Range	Construction	License	Ref.
labMT	10222	$1.3 \rightarrow 8.5$	Survey: MT, 50 ratings	CC	(Dodds et al., 2015a)
ANEW	1034	$1.2 \rightarrow 8.8$	Survey: FSU Psych 101	Free for research	(Bradley and Lang, 1999)
LIWC07	4483	[-1,0,1]	Manual	Paid, commercial	(Pennebaker et al., 2001)
MPQA	7192	[-1,0,1]	Manual + ML	GNU GPL	(Wilson et al., 2005)
OL	6782	[-1,1]	Dictionary propagation	Free	(Liu, 2010)
WK	13915	$1.3 \rightarrow 8.5$	Survey: MT, 14–18 ratings	CC	(Warriner et al., 2013)
LIWC01	2322	[-1,0,1]	Manual	Paid, commercial	(Pennebaker et al., 2001)
LIWC15	6549	[-1,0,1]	Manual	Paid, commercial	(Pennebaker et al., 2001)
PANAS-X	20	[-1,1]	Manual	Copyrighted paper	(Watson and Clark, 1999)
Pattern	1528	$-1.0 \rightarrow 1.0$	Unspecified	BSD	(De Smedt and Daelemans, 2012)
SentiWordNet	147700	$-1.0 \rightarrow 1.0$	Synset synonyms	CC BY-SA 3.0	(Baccianella et al., 2010)
AFINN	2477	[-5,-4, ...,4,5]	Manual	ODbL v1.0	(Nielsen, 2011)
GI	3629	[-1,1]	Harvard-IV-4	Unspecified	(Stone et al., 1966)
WDAL	8743	$0.0 \rightarrow 3.0$	Survey: Columbia students	Unspecified	(Whissell et al., 1986)
EmoLex	14182	[-1,0,1]	Survey: MT	Free for research	(Mohammad and Turney, 2013)
MaxDiff	1515	$-1.0 \rightarrow 1.0$	Survey: MT, MaxDiff	Free for research	(Kiritchenko et al., 2014)
HashtagSent	54129	$-6.9 \rightarrow 7.5$	PMI with hashtags	Free for research	(Zhu et al., 2014)
Sent140Lex	62468	$-5.0 \rightarrow 5.0$	PMI with emoticons	Free for research	(Mohammad et al., 2013)
SOCAL	7494	$-30.2 \rightarrow 30.7$	Manual	GNU GPL	(Taboada et al., 2011)
SenticNet	30000	$-1.0 \rightarrow 1.0$	Label propogation	Citation requested	(Cambria et al., 2014)
Emoticons	132	[-1,0,1]	Manual	Open source code	(Gonçalves et al., 2013)
SentiStrength	2615	[-5,-4, ...,4,5]	LIWC+GI	Free for research	(Thelwall et al., 2010)
VADER	7502	$-3.9 \rightarrow 3.4$	MT survey, 10 ratings	Freely available	(Hutto and Gilbert, 2014)
Umigon	927	[-1,1]	Manual	Public Domain	(Levallois, 2013)
USent	592	[-1,1]	Manual	CC	(Pappas et al., 2013)
EmoSenticNet	13188	[-10,-2,-1,0,1,10]	Bootstrapped extension	Non-commercial	(Poria et al., 2013)

Table 2.1: Summary of dictionary attributes used in sentiment measurement instruments. We provide all acronyms and abbreviations and further information regarding sentiment dictionaries in Subsec. Dictionaries. We test the first 6 dictionaries extensively. The range indicates whether scores are continuous or binary (we use the term binary for sentiment dictionaries for which words are scored as ± 1 *and optionally 0).*

MPQA — The Multi-Perspective Question Answering (MPQA) Subjectivity Dictionary (Wilson et al., 2005).

LIWC{01,07,15} — Linguistic Inquiry and Word Count, three versions (Pennebaker et al., 2001).

OL — Opinion Lexicon, developed by Bing Liu (Liu, 2010).

We also make note of 18 other sentiment dictionaries:

PANAS-X — The Positive and Negative Affect Schedule — Expanded (Watson and Clark, 1999).

Pattern — A web mining module for the Python programming language, version 2.6 (De Smedt and Daelemans, 2012).

SentiWordNet — WordNet synsets each assigned three sentiment scores: positivity, negativity, and objectivity (Baccianella et al., 2010).

AFINN — Words manually rated -5 to 5 with impact scores by Finn Nielsen (Nielsen, 2011).

GI — General Inquirer: database of words and manually created semantic and cognitive categories, including positive and negative connotations (Stone et al., 1966).

WDAL — Whissel's Dictionary of Affective Language: words rated in terms of their Pleasantness, Activation, and Imagery (concreteness) (Whissell et al., 1986).

EmoLex — NRC Word-Emotion Association Lexicon: emotions and sentiment evoked by common words and phrases using Mechanical Turk (Mohammad and Turney, 2013).

MaxDiff — NRC MaxDiff Twitter Sentiment Lexicon: crowdsourced real-valued scores using the MaxDiff method (Kiritchenko et al., 2014).

HashtagSent — NRC Hashtag Sentiment Lexicon: created from Tweets using Pairwise Mutual Information with sentiment hashtags as positive and negative labels (here we use only the unigrams) (Zhu et al., 2014).

Sent140Lex — NRC Sentiment140 Lexicon: created from the "sentiment140" corpus of Tweets, using Pairwise Mutual Information with emoticons as positive and negative labels (here we use only the unigrams) (Mohammad et al., 2013).

SOCAL — Manually constructed general-purpose sentiment dictionary (Taboada et al., 2011).

SenticNet — Sentiment dataset labeled with semantics and 5 dimensions of emotions by Cambria *et al.*, version 3 (Cambria et al., 2014).

Emoticons — Commonly used emoticons with their positive, negative, or neutral emotion (Gonçalves et al., 2013).

SentiStrength — an API and Java program for general purpose sentiment detection (here we use only the sentiment dictionary) (Thelwall et al., 2010).

VADER — method developed specifically for Twitter and social media analysis (Hutto and Gilbert, 2014).

Umigon — Manually built specifically to analyze Tweets from the sentiment140 corpus (Levallois, 2013).

USent — set of emoticons and bad words that extend MPQA (Pappas et al., 2013).

EmoSenticNet — extends SenticNet words with WNA labels (Poria et al., 2013).

All of these sentiment dictionaries were produced by academic groups, and with the exception of LIWC, they are provided free of charge. In Table 2.1, we supply the main aspects—such as word count, score type (continuum or binary), and license information—for the sentiment dictionaries listed above. In the GitHub repository associated with our paper, `https://github.com/andyreagan/sentiment-analysis-comparison`, we include all of the sentiment dictionaries except LIWC.

The labMT, ANEW, and WK sentiment dictionaries have scores ranging on a continuum from 1 (low happiness) to 9 (high happiness) with 5 as neutral, whereas the others we test in detail have scores of ± 1, and either explicitly or implicitly 0 (neutral). We will refer to the latter sentiment dictionaries as being binary, even if neutral is included. Other non-binary ranges include a continuous scale from -1 to 1 (SentiWordNet), integers from -5 to 5 (AFINN), continuous from 1 to 3 (GI), and continuous from -5 to 5 (NRC). For coverage tests, we include all available words, to gain a full sense of the breadth of each sentiment dictionary. In scoring, we do not include neutral words from any sentiment dictionary.

We test the labMT, ANEW, and WK dictionaries for a range of stop words (starting with the removal of words scoring within $\Delta_h = 1$ of the neutral score of 5) (Dodds et al., 2011). The ability to remove stop words—a common practice for text pre-processing—is one advantage of dictionaries that have a range of scores, allowing us to tune the instrument for maximum performance, while retaining all of the benefits of a dictionary method. We will show that, in agreement with the original paper introducing labMT and looking at Twitter data, a $\Delta_h = 1$ is a pragmatic choice in general (Dodds et al., 2011).

Since we do not apply a part of speech tagger, when using the MPQA dictionary we are obliged to exclude words with scores of both +1 and -1. The words and stems with both scores are: blood, boast* (we denote stems with an asterisk), conscience, deep, destiny, keen, large, and precious. We choose to match a text's words using the fixed word set from each sentiment dictionary before stems, hence words with overlapping matches (a fixed word that also matches a stem) are first matched by the fixed word.

2.2.2 Corpora Tested

For each sentiment dictionary, we test both the coverage and the ability to detect previously observed and/or known patterns within each of the following corpora, noting the pattern we hope to discern:

1. The New York Times (NYT) (Sandhaus, 2008): Goal of understanding differences between sections and ranking by sentiment (Subsec. New York Times Word Shift Analysis).
2. Movie reviews (Pang and Lee, 2004): Goal of discerning how emotional language differs in positive and negative reviews and how these differences influence classification accuracy (Subsec. Movie Reviews Classification and Word Shift Analysis).
3. Google Books (Lin et al., 2012): Goal of understanding time series (Subsec. Google Books Time Series and Word Shift Analysis).
4. Twitter: Goal of understanding time series (Subsec. Twitter Time Series Analysis).

For the corpora other than the movie reviews and small numbers of tagged Tweets, there is no publicly available ground truth sentiment, so we instead make comparisons between methods and examine how words contribute to scores. We note that measuring how patterns of sentiment compares with societal measures of well being would also be possible (Mitchell et al., 2013). We offer greater detail on corpus processing below, and we also provide the relevant scripts on GitHub at `https://github.com/andyreagan/sentiment-analysis-comparison`.

2.2.3 Word Shift Graphs

Sentiment analysis is often applied to classify text as positive or negative. Indeed if this were the only use case, the value added by sentiment analysis would be severely limited. Instead we use sentiment analysis as a lens that allow us to see how the emotive words in a text shape the overall content. This is accomplished by first analyzing each word to find its individual contribution to the difference in sentiment scores between two texts. Most importantly, the final step is to examine the words themselves, ranked by their individual contribution. Of the four corpora that we analyze, three rely on this type of qualitative analysis: using the sentiment dictionary as a tool to better understand the sentiment of the corpora rather than as a binary classifier.

To make this possible, we must first find the contribution of each word individually. Starting with the ANEW sentiment dictionary and two texts which we label reference and comparison, we take the difference of their sentiment scores $h_{\text{ANEW}}^{(\text{comp})}$ and $h_{\text{ANEW}}^{(\text{ref})}$, rearrange a few things, and arrive at

$$h_{\text{ANEW}}^{\text{comp}} - h_{\text{ANEW}}^{\text{ref}} = \sum_{w \in ANEW} \underbrace{\left[h_{\text{ANEW}}(w) - h_{\text{ANEW}}^{\text{ref}}\right]}_{+/-} \underbrace{\left[p^{\text{comp}}(w) - p^{\text{ref}}(w)\right]}_{\uparrow/\downarrow}$$

Each word w in the summation contributes to the sentiment difference between the texts according to (1) its sentiment relative to the reference text ($+/-$ = more/less emotive), and (2) its change in frequency of usage ($\uparrow$ / $\downarrow$ = more/less frequent). As a first step, it is possible to visualize this sorted word list in a table, along with the basic indicators of how its contribution is constituted. We use word shift graphs to present this information in the most accessible manner to advanced users. For further detail, we refer the reader to our instructional post and video at http://www.uvm.edu/storylab/2014/10/06/.

2.3 Results

In Fig 2.1, we show a direct comparison between word scores for each pair of the 6 dictionaries tested. Overall, we find strong agreement between all dictionaries with the exceptions we note below. As a guide, we will provide more detail on the individual comparison between the labMT dictionary and the other five dictionaries by examining the words whose scores disagree across dictionaries shown in Fig 2.2. We refer the reader to the S2 Appendix for the remaining individual comparisons.

To start with, consider the comparison of the labMT and ANEW dictionaries on a word-for-word basis. Because these dictionaries share the same range of values, a scatterplot is the natural way to visualize the comparison. Across the top row of Fig 2.1, which compares labMT to the other 5 dictionaries, we see in Panel B for the labMT-ANEW comparison that the RMA best fit (Rayner, 1985) is

$$h_{\text{labMT}}(w) = 0.92 * h_{\text{ANEW}}(w) + 0.40$$

for words w in both labMT and ANEW. The 10 words with farthest from the line of best fit shown in Panel B of Fig 2.2 are (with labMT, ANEW scores in parenthesis): lust (4.64, 7.12), bees (5.60, 3.20), silly (5.30, 7.41), engaged (6.16, 8.00), book (7.24, 5.72), hospital (3.50, 5.04), evil (1.90, 3.23), gloom (3.56, 1.88), anxious (3.42, 4.81), and flower (7.88, 6.64). We observe that these words have high standard deviations in labMT. While the overall agreement is very good, we should expect some variation in the emotional associations of words, due to chance, time of survey, and demographic variability. Indeed, the Mechanical Turk users who scored the words for the labMT set in 2011 are evidently different from the University of Florida students who took the ANEW survey in 2000.

Comparing labMT with WK in Panel C of Fig 2.1, we again find a fit with slope near 1, and with a smaller positive shift: $h_{\text{labMT}}(w) = 0.96 * h_{\text{WK}}(w) + 0.26$. The 10 words farthest from the best fit line, shown in Panel B of Fig 2.2, are (labMT, WK): sue (4.30, 2.18), boogie (5.86, 3.80), exclusive (6.48, 4.50), wake (4.72, 6.57), federal (4.94, 3.06), stroke (2.58, 4.19), gay (4.44, 6.11), patient (5.04, 6.71), user (5.48, 3.67), and blow (4.48, 6.10). Like labMT, the WK dictionary used a Mechanical Turk online survey to gather word ratings. We speculate that the variation is due to differences in the number of scores required for each word in the surveys, with 14–18 in WK and 50

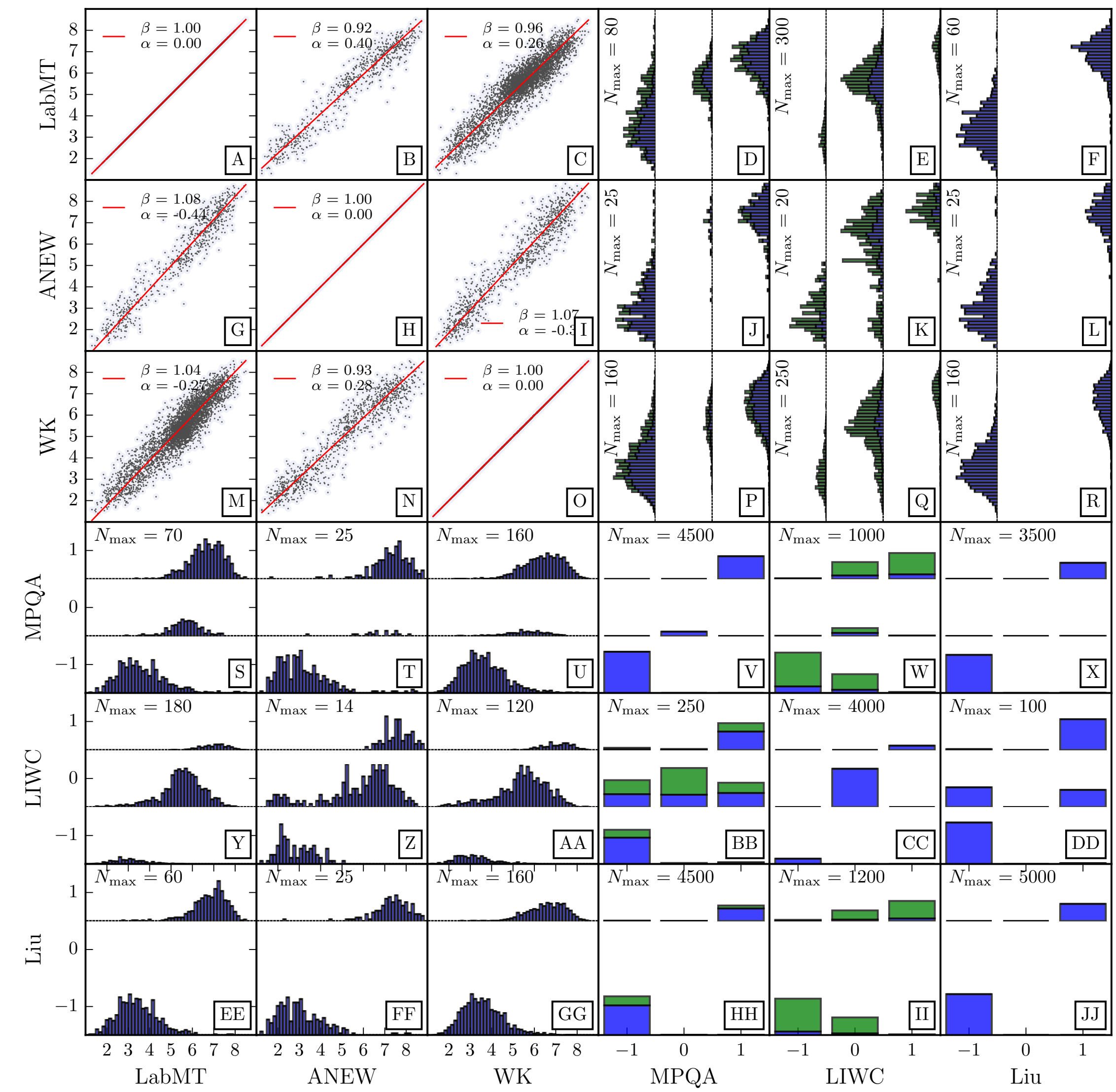


Figure 2.1: Direct comparison of the words in each of the dictionaries tested. For the comparison of two dictionaries, we plot words that are matched by the independent variable "x" in the dependent variable "y". Because of this, and cross stem matching, the plots are not symmetric across the diagonal of the entire figure. Where the scores are continuous in both dictionaries, we compute the RMA linear fit. When a sentiment dictionary contains both fixed and stem words, we plot the matches by fixed words in blue and by stem words in green. The axes in the bar plots are not of the same height, due to large mismatches in the number of words in the dictionaries, and we note the maximum height of the bar in the upper left of such plots. Detailed analysis of Panel C can be found in (Dodds et al., 2015b). We provide a table for each off-diagonal panel in the S2 Appendix with the words whose scores exhibit the greatest mismatch, and a subset of these tables in Fig 2.2.

in labMT. For an in depth comparison of these sentiment dictionaries, see reference (Dodds et al., 2015b).

To compare the word scores in a binary sentiment dictionaries (those with ± 1 or $\pm 1, 0$) to the word scores in a sentiment dictionary with a 1–9 range, we examine the distribution of the continuous scores for each binary score. Looking at the labMT-MPQA comparison in Panel D of Fig 2.1, we see that more of the matches are between words without stems (blue) than those with stems (orange), and that each score in -1, 0, +1 from MPQA corresponds to a wider range of scores in labMT. We examine the shared individual words from labMT with high sentiment scores and MPQA with score -1, both the happiest and the least happy in labMT with MPQA score 0, and the least happy when MPQA is 1 (Fig 2.2 Panels C-E). The 10 happiest words in labMT matched by MPQA words with score -1 are: moonlight (7.50), cutest (7.62), finest (7.66), funniest (7.76), comedy (7.98), laughs (8.18), laughing (8.20), laugh (8.22), laughed (8.26), laughter (8.50). This is an immediately troubling list of evidently positive words rated as -1 in MPQA. We observe the top 5 are matched by the stem "laugh*" in MPQA. The least happy 5 words and happiest 5 words in labMT matched by words in MPQA with score 0 are: sorrows (2.69), screaming (2.96), couldn't (3.32), pressures (3.49), couldnt (3.58), and baby (7.28), precious (7.34), strength (7.40), surprise (7.42), and song (7.58). We see that these MPQA word scores are departures from the other dictionaries, warranting concern about their scores. The least happy words in labMT with score +1 in MPQA that are matched by MPQA are: vulnerable (3.34), court (3.78), sanctions (3.86), defendant (3.90), conviction (4.10), backwards (4.22), courts (4.24), defendants (4.26), court's (4.44), and correction (4.44). These words have sentiments that appear to vary with context.

While it would be simple to adjust these ratings in the MPQA dictionary going forward, we are naturally led to be concerned about existing work using MPQA that does not examine words contributing to overall sentiment. We note again that the use of word shift graphs of some kind would have exposed these problematic scores immediately.

For the labMT-LIWC comparison in Panel E of Fig 2.1 we examine the same matched word lists as before. The 10 happiest words in labMT matched by words in LIWC with score -1 are: trick (5.22), shakin (5.29), number (5.30), geek (5.34), tricks (5.38), defence (5.39), dwell (5.47), doubtless (5.92), numbers (6.04), shakespeare (6.88). From Panel F of Fig 2.2, the least happy 5

A: LabMT comparison with ANEW

Word	h_{LabMT}	h_{ANEW}	h_{diff}
lust	4.64	7.12	1.72
bees	5.60	3.20	1.66
silly	5.30	7.41	1.43
engaged	6.16	8.00	1.20
book	7.24	5.72	1.15
hospital	3.50	5.04	1.15
evil	1.90	3.23	1.09
gloom	3.56	1.88	1.05
anxious	3.42	4.81	1.05
flower	7.88	6.64	1.00

B: LabMT comparison with WK

Word	h_{LabMT}	h_{WK}	h_{diff}
sue	4.30	2.18	1.39
boogie	5.86	3.80	1.39
exclusive	6.48	4.50	1.36
wake	4.72	6.57	1.35
federal	4.94	3.06	1.25
stroke	2.58	4.19	1.24
gay	4.44	6.11	1.23
patient	5.04	6.71	1.22
user	5.48	3.67	1.21
blow	4.48	6.10	1.20

C: LabMT comparison with MPQA's negative words

Word	h_{LabMT}	h_{MPQA}
fine	6.74	-1
game	6.92	-1
cartoon	7.20	-1
eternal	7.20	-1
moon	7.28	-1
fun	7.96	-1
comedy	7.98	-1
laugh	8.22	-1
laugh	8.22	-1
laughter	8.50	-1

D: LabMT comparison with MPQA's neutral words

Word	h_{LabMT}	h_{MPQA}
screaming	2.96	0
pressures	3.49	0
pressure	3.66	0
plead	3.67	0
mean	3.68	0
baby	7.28	0
precious	7.34	0
strength	7.40	0
surprise	7.42	0
surprise	7.42	0

E: LabMT comparison with MPQA's positive words

Word	h_{LabMT}	h_{MPQA}
vulnerable	3.34	+1
court	3.78	+1
conviction	4.10	+1
craving	4.46	+1
excuse	4.58	+1
bull	4.62	+1
striking	4.70	+1
offset	4.72	+1
admit	4.74	+1
repair	4.76	+1

F: LabMT comparison with LIWC's neutral words

Word	h_{LabMT}	h_{LIWC}
lack	3.16	0
couldn't	3.32	0
cannot	3.32	0
never	3.34	0
against	3.40	0
rest	7.18	0
greatest	7.26	0
couple	7.30	0
million	7.38	0
billion	7.56	0

Figure 2.2: We present the specific words from Panels G, M, S and Y of Fig 2.1 with the greatest mismatch. Only the center histogram from Panel Y of Fig 2.1 is included. We emphasize that the labMT dictionary scores generally agree well with the other dictionaries, and we are looking at the marginal words with the strongest disagreement. Within these words, we detect differences in the creation of these dictionaries that carry through to these edge cases. Panel A: The words with most different scores between the labMT and ANEW dictionaries are suggestive of the different meanings that such words entail for the different demographic surveyed to score the words. Panel B: Both dictionaries use surveys from the same demographic (Mechanical Turk), where the labMT dictionary required more individual ratings for each word (at least 50, compared to 14) and appears to have dampened the effect of multiple meaning words. Panels C–E: The words in labMT matched by MPQA with scores of -1, 0, and +1 in MPQA show that there are at least a few words with negative rating in MPQA that are not negative (including the happiest word in the labMT dictionary: "laughter"), not all of the MPQA words with score 0 are neutral, and that MPQA's positive words are mostly positive according to the labMT score. Panel F: The function words in the expert-curated LIWC dictionary are not emotionally neutral.

neutral words and happiest 5 neutral words in LIWC, matched in LabMT from LIWC words (i.e., using the word stems in LIWC to match across LabMT, directionality matters), are: negative (2.42), lack (3.16), couldn't (3.32), cannot (3.32), never (3.34), millions (7.26), couple (7.30), million (7.38), billion (7.56), millionaire (7.62). The least happy words in labMT with score +1 in LIWC that are matched by LIWC are: merrill (4.90), richardson (5.02), dynamite (5.04), careful (5.10), richard (5.26), silly (5.30), gloria (5.36), securities (5.38), boldface (5.40), treasury's (5.42). The +1 and -1 words in LIWC match some neutral words in labMT, which is not alarming. However, the problems with the "neutral" words in the LIWC set are immediate: these are not emotionally neutral words. The range of scores in labMT for these 0-score words in LIWC formed the basis for Garcia *et al.*'s response to (Dodds et al., 2015a), and we point out here that the authors must not have looked at the words, an all-too-common problem in studies using sentiment analysis (Garcia et al., 2015; Dodds et al., 2015b).

For the labMT-OL comparison in Panel E of Fig 2.1 we again examine the same matched word lists as before (except the neutral word list because OL has no explicit neutral words). The 10 happiest words in labMT matched by OL's negative list are: myth (5.90), puppet (5.90), skinny (5.92), jam (6.02), challenging (6.10), fiction (6.16), lemon (6.16), tenderness (7.06), joke (7.62), funny (7.92). The least happy words in labMT with score +1 in OL that are matched by OL are: defeated (2.74), defeat (3.20), envy (3.33), obsession (3.74), tough (3.96), dominated (4.04), unreal (4.57), striking (4.70), sharp (4.84), sensitive (4.86). Despite nearly twice as many negative words in OL as positive words (at odds with the frequency-dependent positivity bias of language (Dodds et al., 2015a)), after examining the words which are the most differently scored and seeing how quickly the labMT scores move into the neutral range, we can conclude that these dictionaries generally agree with the exception of only a few bad matches.

Direct comparisons between the word scores in sentiment dictionaries, while evidently tedious, have brought to light many problematic word scores. In addition, this analysis serves as a template for further comparisons of the words across new sentiment dictionaries. The six sentiment dictionaries under careful examination in the present study are further analyzed in the Supporting Information. Next, we examine how each sentiment dictionary can aid in understanding the sentiments contained in articles from the New York Times.

2.3.1 New York Times Word Shift Analysis

The New York Times corpus (Sandhaus, 2008) is split into 24 sections of the newspaper that are roughly contiguous throughout the data from 1987–2008. With each sentiment dictionary, we rate each section and then compute word shift graphs (described below) against the baseline, and produce a happiness ranked list of the sections.

To gain understanding of the sentiment expressed by any given text relative to another text, it is necessary to inspect the words which contribute most significantly by their emotional strength and the change in frequency of usage. We do this through the use of word shift graphs, which plot the contribution of each word w from the sentiment dictionary (denoted $\delta h_{\text{ANEW}}(w)$) to the shift in average happiness between two texts, sorted by the absolute value of the contribution. We use word shift graphs to both analyze a single text and to compare two texts, here focusing on comparing text within corpora. For a derivation of the algorithm used to make word shift graphs while separating the frequency and sentiment information, we refer the reader to Equations 2 and 3 in (Dodds et al., 2011). We consider both the sentiment difference and frequency difference components of $\delta h_{\text{ANEW}}(w)$ by writing each term of Eq. B.1 as in (Dodds et al., 2011):

$$\delta h_{\text{ANEW}}(w) = 100 \frac{h_{\text{ANEW}}(w) - h_{\text{ANEW}}^{\text{ref}}}{h_{\text{ANEW}}^{\text{comp}} - h_{\text{ANEW}}^{\text{ref}}} \left[p(w)^{\text{comp}} - p(w)^{\text{ref}} \right]. \tag{2.2}$$

An in-depth explanation of how to interpret the word shift graph can also be found at `http://hedonometer.org/instructions.html#wordshifts`.

To both demonstrate the necessity of using word shift graphs in carrying out sentiment analysis, and to gain understanding about the ranking of New York Times sections by each sentiment dictionary, we look at word shift graphs for the "Society" section of the newspaper from each sentiment dictionary in Fig 2.3, with the reference text being the whole of the New York Times. The "Society" section happiness ranks 1, 1, 1, 18, 1, and 11 within the happiness of each of the 24 sections in the dictionaries labMT, ANEW, WK, MPQA, LIWC, and OL, respectively. These graphs show only the very top of the distributions which range in length from 1030 (ANEW) to 13915 words (WK).

First, using the labMT dictionary, we see that the words "graduated", "father", and "university" top the list, which is dominated by positive words that occur more frequently ($+\uparrow$). These more

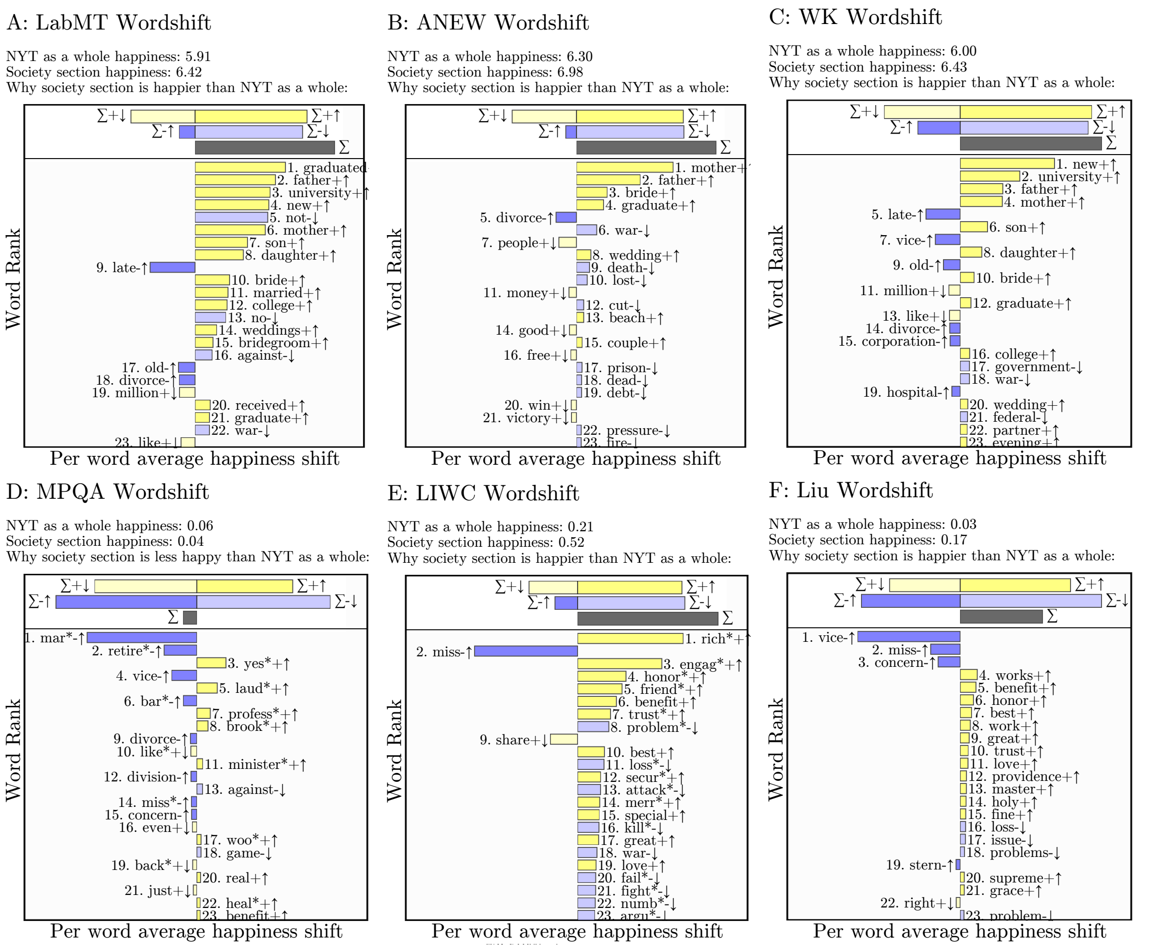


Figure 2.3: New York Times (NYT) "Society" section shifted against the entire NYT corpus for each of the six dictionaries listed in tiles A–F. We provide a detailed analysis in Sec. 2.3.1. Generally, we are able to glean the greatest understanding of the sentiment texture associated with this NYT section using the labMT dictionary. Additionally we note the labMT dictionary has the most coverage quantified by word match count (Figure in S3 Appendix), we are able to identify and correct problematic words scores in the OL dictionary, and we see that the MPQA dictionary disagrees entirely with the others because of an overly broad stem match.

frequent positive words paint a clear picture of family life (relationships, weddings, and divorces), as well as university accomplishment (graduations and college). In general, we are able to observe with only these words that the "Society" section is where we find the details of these events.

From the ANEW dictionary, we see that a few positive words have increased frequency, lead by "mother", "father", and "bride". Looking at this shift in isolation, we see only these words with three more ("graduate", "wedding", and "couple") that would lead us to suspect these topics are present in the "Society" section.

The WK dictionary, with the most individual word scores of any sentiment dictionary tested, agrees with labMT and ANEW that the "Society" section is the happiest section, with somewhat similar set of words at the top: "new", "university", and "father". Low coverage of the New York Times corpus (see Fig A.3) resulted in less specific words describing the "Society" section, with more words that go down in frequency in the shift. With the words "bride" and "wedding" up, as well as "university", "graduate", and "college", it is evident that the "Society" section covers both graduations and weddings, in consensus with the other sentiment dictionaries.

The MPQA dictionary ranks the "Society" section 18th of the 24 NYT sections, a departure from the other rankings, with the words "mar*", "retire*", and "yes*" the top three contributing words (where "*" denotes a wildcard "stem" match). Negative words increasing in frequency ($-\uparrow$) are the most common type near the top, and of these, the words with the biggest contributions are being scored incorrectly in this context (specifically words "mar*", "retire*", "bar*", "division", and "miss*"). Looking more in depth at the problems created by the first out of context word match, we find 1211 unique words match "mar*". The five most frequent, with counts in parenthesis, are married (36750), marriage (5977), marketing (5382), mary (4403), and mark (2624). The score for these words in MPQA is -1, in stark contrast to the scores in other sentiment dictionaries (e.g., the labMT scores are 6.76, 6.7, 5.2, 5.88, and 5.48). These problems plague the MPQA dictionary for scoring the New York Times corpus, and without using word shift graphs would have gone completely unseen. In an attempt to fix contextual issues by fixing corpus-specific words, we remove "mar*,retire*,vice,bar*,miss*" and find that the MPQA dictionary ranks the Society section of the NYT at 15th of the 24 sections

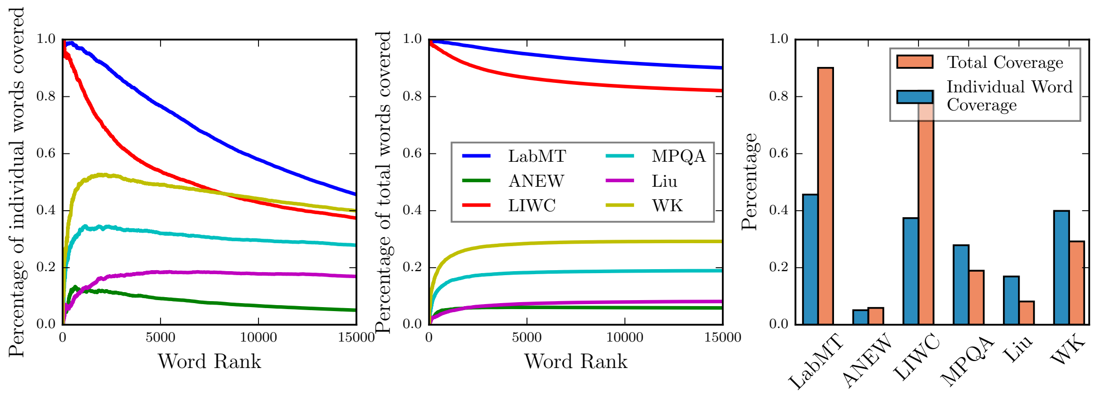


Figure 2.4: Coverage of the words in the movie reviews by each of the dictionaries. We observe that the labMT dictionary has the highest coverage of words in the movie reviews corpus both across word rank and cumulatively. The LIWC dictionary has initially high coverage since it contains some high-frequency function words, but quickly drops off across rank. The WK dictionary coverage increases across word rank and cumulatively, indicating that it contains a large number of less common words in the movie review corpus. The OL, ANEW, and MPQA have a cumulative coverage of less than 20% of the lexicon.

The second binary sentiment dictionary, LIWC, agrees well with the first three dictionaries and ranks the "Society" section at the top with the words "rich*", "miss", and "engage*" at the top of the list. We immediately notice that the word "miss" is being used frequently in the "Society" section in a different sense than was coded for in the LIWC dictionary: it is used in the corpus to mean "the title prefixed to the name of an unmarried woman", but is scored as negative in LIWC (with the likely intended meaning "to fail to reach an target or to acknowledge loss"). We would remove this word from LIWC for further analysis of this corpus (we would also remove the word "trust" here). The words matched by "miss*" aside, LIWC finds some positive words going up ($+\uparrow$), with "engage*" hinting at weddings. Without words that capture the specific behavior happening in the "Society" section, we are unable to see anything about college, graduations, or marriages, and there is much less to be gained about the text from the words in LIWC than some of the other dictionaries we have seen. Nevertheless, LIWC finds consensus with the "Society" section ranked the top section, due in large part to a lack of negative words "war" (rank 18) and "fight*" (rank 22).

The OL sentiment dictionary departs from the consensus and ranks the "Society" section at 11th out of the 24 sections. The top three words, "vice", "miss", and "concern", contribute largely with respect to the rest of distribution, of which two are clearly being used in the wrong sense. For a more reasonable analysis we would remove both "vice" and "miss" from the OL dictionary to score this text. For a more reasonable analysis we remove both “vice” and “miss” from the OL dictionary to score this text, and in doing so the happiness goes from 0.168 to 0.297, making the “Society” section the second happiest of the 24 sections. Focusing on the words, we see that the OL dictionary finds many positive words increasing in frequency ($+\uparrow$) that are mostly generic. In the word shift graph we do not find the wedding or university events as in sentiment dictionaries with more coverage, but rather a variety of positive language surrounding these events, for example 4. “works”, “benefit” (5), “honor” (6), “best” (7), “great” (9), “trust” (10), “love” (11), etc. While this does not provide insight into the topics, the OL sentiment dictionary with fixes from the word shift graph analysis does provide details on the emotive words that make the “Society” section one of the happiest sections.

In conclusion, we find that 4 of the 6 dictionaries score the “Society” section at number 1, and in these cases we use the word shift graph to uncover the nuances of the language used. We find, unsurprisingly, that the most matches are found by the labMT dictionary, which is in part built from the NYT corpus (see S3 Appendix for coverage plots). Without as much corpus-specific coverage, we note that while the nuances of the text remain hidden, the LIWC and OL dictionaries still highlight the positive language in this section. Of the two that did not score the “Society” section at the top, we are able to assess and repair the MPQA and the OL dictionaries by removing the words “mar*,retire*,vice*,bar*,miss*” and “vice,miss”, respectively. By identifying words used in the wrong sense/context using the word shift graph, we directly improve the sentiment score for the New York Times corpus from both MPQA and OL dictionaries closer to consensus. While the OL dictionary, with two corrections, agrees with the other dictionaries, the MPQA dictionary with five corrections still ranks the Society section of the NYT as the 15th happiest of the 24 sections.

In the first Figure in S4 Appendix we show scatterplots for each comparison, and compute the Reduced Major Axes (RMA) regression fit (Rayner, 1985). In the second Figure in S4 Appendix we show the sorted bar chart from each sentiment dictionary.

2.3.2 Movie Reviews Classification and Word Shift Graph Analysis

For the movie reviews, we first provide insight into the language differences and secondly perform binary classification of positive and negative reviews. The entire dataset consists of 1000 positive and 1000 negative reviews, as rated with 4 or 5 stars and 1 or 2 stars, respectively. We show how well each sentiment dictionary covers the review database in Fig 2.4. The average review length is 650 words, and we plot the distribution of review lengths in S5 Appendix. We average the sentiment of words in each review individually, using each sentiment dictionary. We also combine random samples of N positive or N negative reviews for N varying from 2 to 900 on a logarithmic scale, without replacement, and rate the combined text. With an increase in the size of the text, we expect that the dictionaries will be better able to distinguish positive from negative. The simple statistic we use to describe this ability is the percentage of distributions that overlap the average.

To analyze which words are being used by each sentiment dictionary, we compute word shift graphs of the entire positive corpus versus the entire negative corpus in Fig 2.5. Across the board, we see that a decrease in negative words is the most important word type for each sentiment dictionary, with the word "bad" being the top word for every sentiment dictionary in which it is scored (ANEW does not have it). Other observations that we can make from the word shift graphs include a few words that are potentially being used out of context: "movie", "comedy", "plot", "horror", "war", "just".

In the lower right panel of Fig 2.6, the percentage overlap of positive and negative review distributions presents us with a simple summary of sentiment dictionary performance on this tagged corpus. The ANEW dictionary stands out as being considerably less capable of distinguishing positive from negative. In order, we then see WK is slightly better overall, labMT and LIWC perform similarly better than WK overall, and then MPQA and OL are each a degree better again, across the review lengths (see below for hard numbers at 1 review length). Two Figures in S5 Appendix show the distributions for 1 review and for 15 combined reviews.

Classifying single reviews as positive or negative, the F1-scores are: labMT .63, ANEW .36, LIWC .53, MPQA .66, OL .71, and WK .34 (see Table A.4). We roughly confirm the rule-of-thumb that 10,000 words are enough to score with a sentiment dictionary confidently, with all dictionaries except MPQA and ANEW achieving 90% accuracy with this many words. We sample the number of reviews evenly in log space, generating sets of reviews with average word counts of 4550, 6500, 9750, 16250, and 26000 words. Specifically, the number of reviews necessary to achieve 90% accuracy is 15 reviews (9750 words) for labMT, 100 reviews (65000 words) for ANEW, 10 reviews (6500 words) for

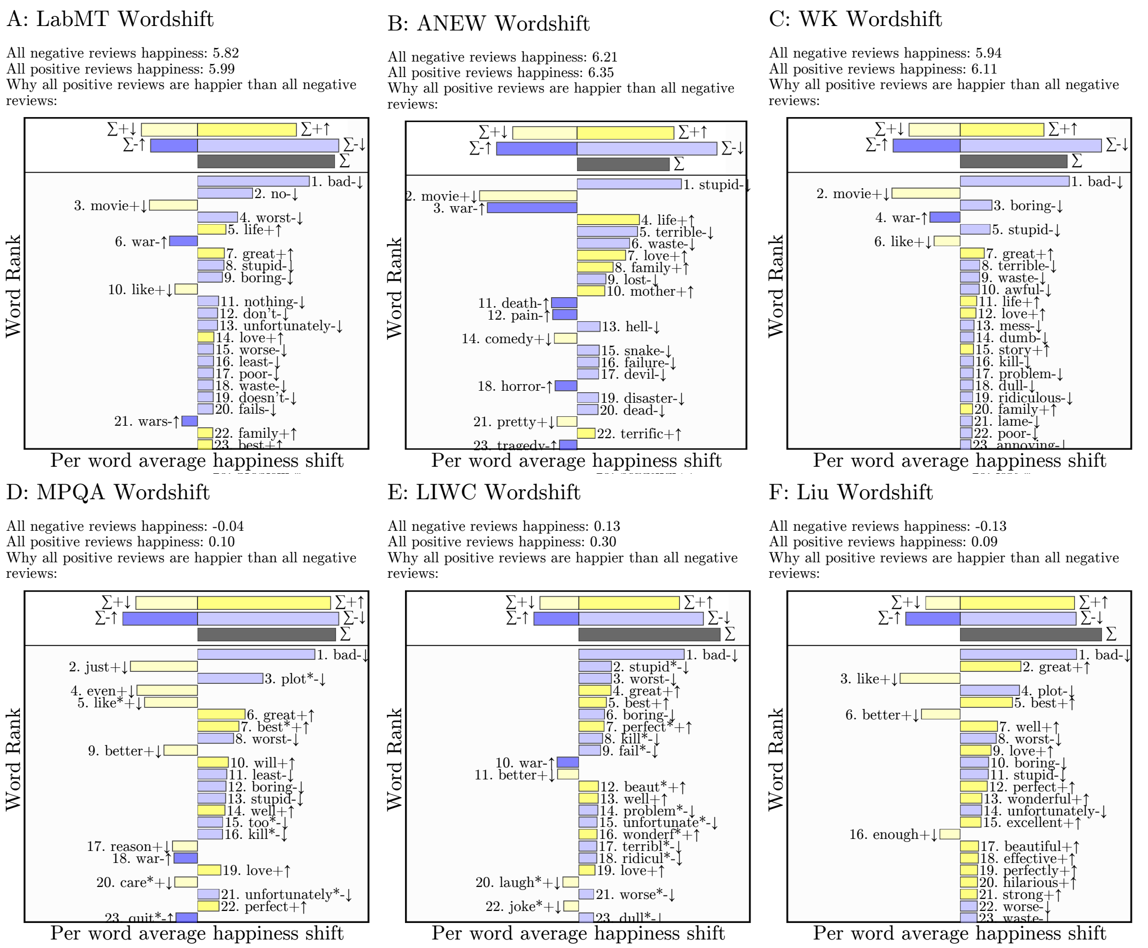


Figure 2.5: Word shift graphs for the movie review corpus. By analyzing the words that contribute most significantly to the sentiment score produced by each sentiment dictionary we are able to identify words that are problematic for each sentiment dictionary at the word-level, and generate an understanding of the emotional texture of the movie review corpus. Again we find that coverage of the lexicon is essential to produce meaningful word shift graphs, with the labMT dictionary providing the most coverage of this corpus and producing the most detailed word shift graphs. All words on the left hand side of these word shift graphs are words that individually made the positive reviews score more negatively than the negative reviews, and the removal of these words would improve the accuracy of the ratings given by each sentiment dictionary. In particular, across each sentiment dictionary the word shift graphs show that domain-specific words such as "war" and "movie" are used more frequently in the positive reviews and are not useful in determining the polarity of a single review.

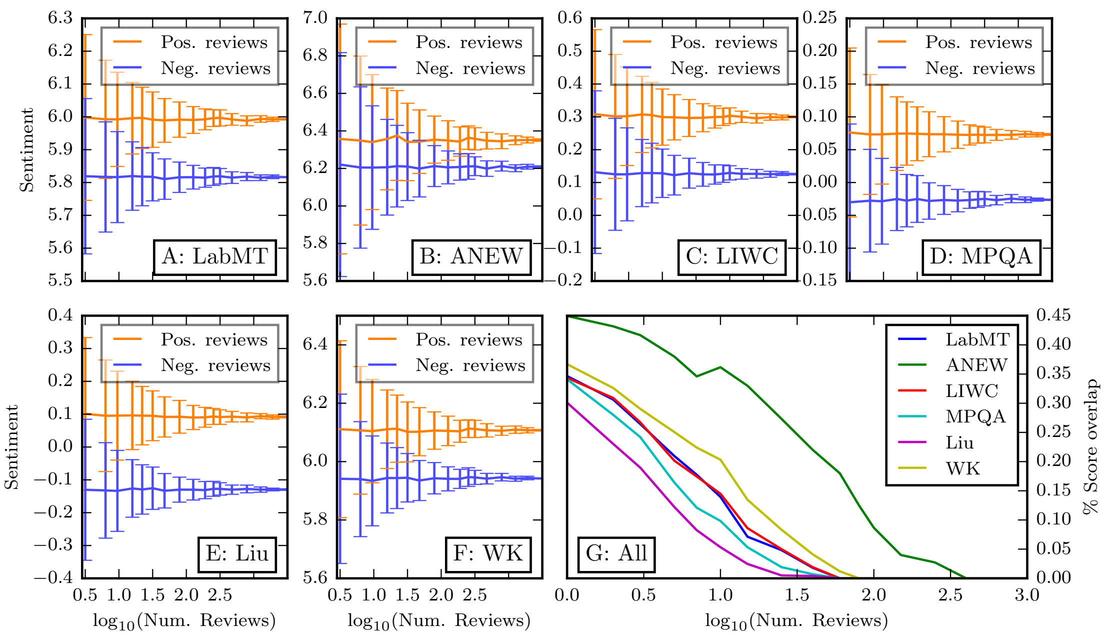


Figure 2.6: The score assigned to increasing numbers of reviews drawn from the tagged positive and negative sets. For each sentiment dictionary we show mean sentiment and 1 standard deviation over 100 samples for each distribution of reviews in Panels A–F. For comparison we compute the fraction of the distributions that overlap in Panel G. At the single review level for each sentiment dictionary this simple performance statistic (fraction of distribution overlap) ranks the OL dictionary in first place, the MPQA, LIWC, and labMT dictionaries in a second place tie, WK in fifth, and ANEW far behind. All dictionaries require on the order of 1000 words to achieve 95% classification accuracy.

LIWC, 10 reviews (6500 words) for MPQA, 7 reviews (4550 words) for OL, and 25 reviews (16250 words) for WK.

While we are analyzing the movie review classification, which has ground truth labels, we will take a moment to further support our claims for the inaccuracy of these methods at the sentence level. The OL dictionary, with the highest performance classifying individual movie reviews of the 6 dictionaries tested in detail, performs worse than guessing at classifying individual sentences in movie reviews. Specifically, 76.9/74.2% of sentences in the positive/negative reviews sets have words in the OL dictionary, and of these OL achieves an F1-score of 0.44. The results for each sentiment dictionary are included in Table A.5, with an average (median) F1 score of 0.42 (0.45) across all dictionaries. While these results do cast doubt on the ability to classify positive and negative reviews from single sentences using dictionary based methods, we note that we need not expect the sentiment of individual sentences to be strongly correlated with the overall review polarity.

2.3.3 Google Books Time Series and Word Shift Analysis

We use the Google books 2012 dataset with all English books (Lin et al., 2012), from which we remove part of speech tagging and split into years. From this, we make time series by year, and word shift graphs of decades versus the baseline. In addition, to assess the similarity of each time series, we produce correlations between each of the time series.

Despite claims from research based on the Google Books corpus (Michel et al., 2011), we keep in mind that there are several deep problems with this beguiling data set (Pechenick et al., 2015). Leaving aside these issues, the Google Books corpus nevertheless provides a substantive test of our six dictionaries.

In Fig 2.7, we plot the sentiment time series for Google Books. Three immediate trends stand out: a dip near the Great Depression, a dip near World War II, and a general upswing in the 1990's and 2000's. From these general trends, a few dictionaries waver: OL does not dip as much for WW2, OL and LIWC stay lower in the 90's and 2000's, and labMT with $\Delta_h = 0.5, 1.0$ go downward near the end of the 2000's. We take a closer look into the 1940's to see what each sentiment dictionary is picking up in Google Books around World War 2 in Figure in S6 Appendix.

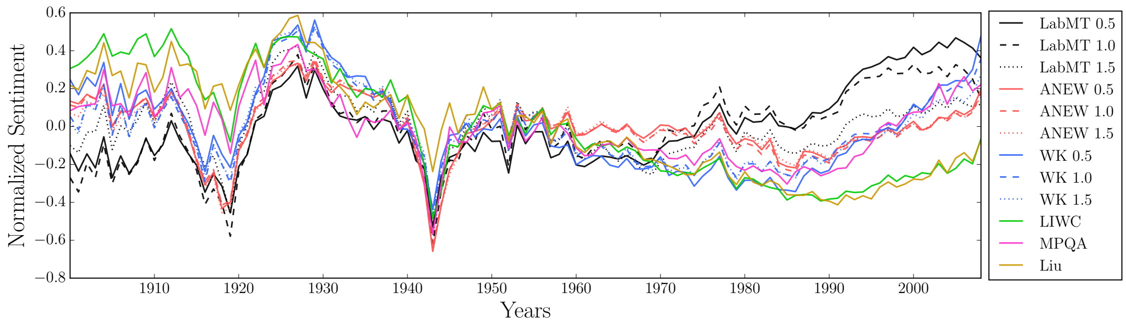


Figure 2.7: Google Books sentiment time series from each sentiment dictionary, with stop values of 0.5, 1.0, and 1.5 from the dictionaries with word scores in the 1–9 range. To normalize the sentiment score, we subtract the mean and divide by the absolute range. We observe that each time series has increased variance, with a few pronounced negative time periods, and trending positive towards the end of the corpus. The score of labMT varies substantially with different stop values, although remaining highly correlated, and finds absolute lows near the World Wars. The LIWC and OL dictionaries trend down towards 1990, dipping as low as the war periods.

In each panel of the word shift Figure in S6 Appendix, we see that the top word making the 1940's less positive than the the rest of Google Books is "war", which is the top contributor for every sentiment dictionary except OL. Rounding out the top three contributing words are "no" and "great", and we infer that the word "great" is being seen from mention of "The Great Depression" or "The Great War", and is possibly being used out of context. All dictionaries but ANEW have "great" in the top 3 words, and each sentiment dictionary could be made more accurate if we remove this word.

In Panel A of the 1940's word shift Figure in S6 Appendix, beyond the top words, increasing words are mostly negative and war-related: "against", "enemy", "operation", which we could expect from this time period.

In Panel B, the ANEW dictionary scores the 1940's of Google Books lower than the baseline as well, finding "war", "cancer", and "cell" to be the most important three words. With only 1030 words, there is not enough coverage to see anything beyond the top word "war," and the shift is dominated by words that go down in frequency.

In Panel C, the WK dictionary finds the the 1940's with slightly less happy than the baseline, with the top three words being "war", "great", and "old". We see many of the same war-related words as in labMT, and in addition some positive words like "good" and "be" are up in frequency.

The word “first” could be an artifact of first aid, a claim that could be substantiated with further analysis of the Google Books corpus at the 2-gram level beyond the scope of this manuscript.

In Panel D, the MPQA dictionary rates the 1940’s slightly less happy than the baseline, with the top three words being “war”, “great”, and “differ*”. Beyond the top word “war”, the score is dominated by words decreasing in frequency, with only a few words up in frequency. Without specific words increasing in frequency as contextual guides, it is difficult to obtain a good glance at the nature of the text. For this reason, having a higher coverage of the words in the corpus enables understanding.

In Panel E, the LIWC dictionary rates the 1940’s nearly the same as the baseline, with the top three words being “war”, “great”, and “argu*”. When the scores are nearly the same, although the 1940’s are slightly higher happiness here, the word shift is a view into how the words of the reference and comparison text vary. In addition to a few war related words being up and bringing the score down (“fight”, “enemy”, “attack”), we see some positive words up that could also be war related: “certain”, “interest”, and “definite”. Although LIWC does not manage to find World War II as a low point of the 20th century, the words that contribute to LIWCs score for the 1940’s compared to all years are useful in understanding the corpus.

In Panel F, the OL dictionary rates the 1940’s as happier than the baseline, with the top three words being “great”, “support”, and “like”. With 7 positive words up, and 1 negative word up, we see how the OL dictionary misses the war without the word “war” itself and with only “enemy” contributing from the words surrounding the conflict. The nature of the positive words that are up is unclear, and could justify a more detailed analysis of why the OL dictionary fails here.

2.3.4 Twitter Time Series Analysis

For Twitter data, we use the Gardenhose feed, a random 10% of the full stream. We store data on the Vermont Advanced Computing Core (VACC), and process the text first into hash tables (with approximately 8 million unique English words each day) and then into word vectors for each 15 minutes, for each sentiment dictionary tested. From this, we build sentiment time series for time resolutions of 15 minutes, 1 hour, 3 hours, 12 hours, and 1 day. In addition to the raw time series,

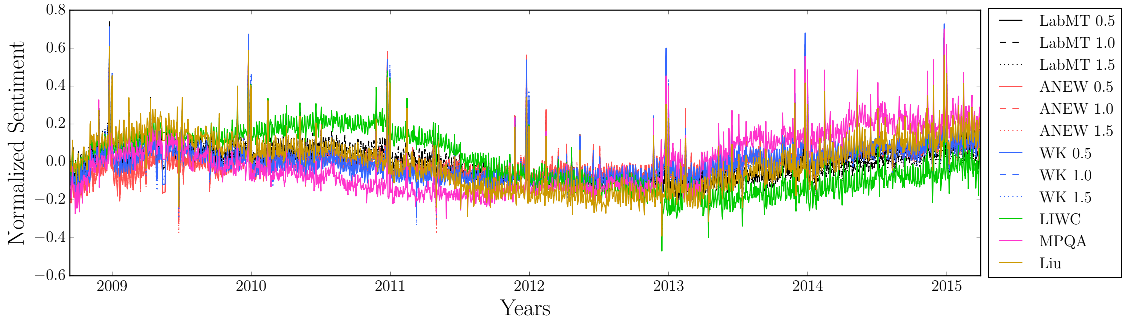


Figure 2.8: Normalized sentiment time series on Twitter using Δ_h of 1.0 for all dictionaries. To normalize the sentiment score, we subtract the mean and divide by the absolute range. The resolution is 1 day, and draws on the 10% gardenhose sample of public Tweets provided by Twitter. All of the dictionaries exhibit wide variation for very early Tweets, and from 2012 onward generally track together strongly with the exception of MPQA and LIWC. The LIWC and MPQA dictionaries show opposite trends: a rise until 2012 with a decline after 2012 from LIWC, and a decline before 2012 with a rise afterwards from MPQA. To analyze the trends we look at the words driving the movement across years using word shift Figures in S7 Appendix.

we compute correlations between each time series to assess the similarity of the ratings between dictionaries.

In Fig 2.8, we present a daily sentiment time series of Twitter processed using each of the dictionaries being tested. With the exception of LIWC and MPQA we observe that the dictionaries generally track well together across the entire range. A strong weekly cycle is present in all, although muted for ANEW.

We plot the Pearson's correlation between all time series in Fig 2.9, and confirm some of the general observations that we can make from the time series. Namely, the LIWC and MPQA time series disagree the most from the others, and even more so with each other. Generally, we see strong agreement within dictionaries with varying stop values Δh.

The time series from each sentiment dictionary exhibits increased variance at the start of the time frame, when Twitter volume is low in 2008 and into 2009. As more people join Twitter and the Tweet volume increases through 2010, we see that LIWC rates the text as happier, while the rest start a slow decline in rating that is led by MPQA in the negative direction. In 2010, the LIWC dictionary is more positive than the rest with words like "haha", "lol" and "hey" being used more frequently and swearing being less frequent than all years of Twitter put together. The other dictionaries with more coverage find a decrease in positive words to balance this increase, with the exception of MPQA which finds many negative words going up in frequency (see 2010 word shift

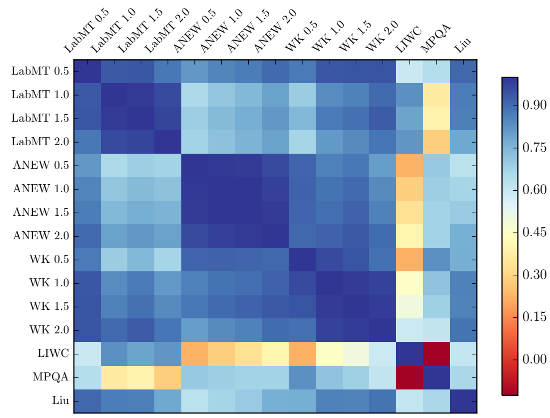


Figure 2.9: Pearson's r correlation between daily resolution Twitter sentiment time series for each sentiment dictionary. We see that there is strong agreement within dictionaries, with the biggest differences coming from the stop value of $\Delta h = 0.5$. The labMT and OL dictionaries do not strongly disagree with any of the others, while LIWC is the least correlated overall with other dictionaries. labMT and OL correlate strongly with each other, and disagree most with the ANEW, LIWC, and MPQA dictionaries. The two least correlated dictionaries are the LIWC and MPQA dictionaries. Again, since there is no publicly accessible ground truth for Twitter sentiment, we compare dictionaries against the others, and look at the words. With these criteria, we find the labMT dictionary to be the most useful.

Figure in Appendix S7). All of the dictionaries agree most strongly in 2012, all finding a lot of negative language and swearing that brings scores down (see 2012 word shift Figure in Appendix S7). From the bottom at 2012, LIWC continues to go downward while the others trend back up. The signal from MPQA jumps to the most positive, and LIWC does start trending back up eventually. We analyze the words in 2014 with a word shift against all 7 years of Tweets for each sentiment dictionary in each panel in the 2014 word shift Figure in Appendix S7: A. labMT scores 2014 as less happy with more negative language. B. ANEW finds it happier with a few positive words up. C. WK finds it happier with more negative words (like labMT). D. MPQA finds it more positive with less negative words. E. LIWC finds it less positive with more negative and less positive words. F. OL finds it to be of the same sentiment as the background with a balance in positive and negative word usage. From these word shift graphs, we can analyze which words cause MPQA and LIWC to disagree with the other dictionaries: the disagreement of MPQA is again marred by broad stem matches, and the disagreement of LIWC is due to a lack of coverage.

2.3.5 Brief Comparison to Machine Learning Methods

We implement a Naive Bayes (NB) classifier (sometimes harshly called idiot Bayes (Hand and Yu, 2001)) on the tagged movie review dataset, using 10% of the reviews for training and then testing performance on the rest. Following standard practice, we remove the top 30 ranked words ("stop words") from the 5000 most frequent words, and use the remaining 4970 words in our classifier for maximum performance (we observe a 0.5% improvement). Our implementation is analogous to those found in common Python natural language processing packages (see "NLTK" or "TextBlob" in (Bird, 2006)).

As we should expect, at the level of single review, NB outperforms the dictionary-based methods with a classification accuracy of 75.7% averaged over 100 trials. As the number of reviews is increased, the overlap from NB diminishes, and using our simple "fraction overlapping" metric, the error drops to 0 with more than 200 reviews. Interestingly, NB starts to do worse with more reviews, and with more than 500 of the 1000 reviews concatenated, it rates both the positive and negative reviews as positive (Figure in S8 Appendix).

The rating curves do not touch, and neither do the standard deviation error bars (indicating that the result is not statistically significant), but they both go very slightly above 0 (again, see Figure in S8 Appendix). Overall, with Naive Bayes we are able to classify a higher percentage of individual reviews correctly, but with more variance.

In the two Tables in S8 Appendix we compute the words which the NB classifier uses to classify all of the positive reviews as positive, and all of the negative reviews as positive. The Natural Language Toolkit (NLTK (Bird, 2006)) implements a method to obtain the "most informative" words, by taking the ratio of the likelihood of words between all available classes, and looking for the largest ratio:

$$\max_{\text{all words } w} \frac{P(w|c_i)}{P(w|c_j)} \tag{2.3}$$

for all combinations of classes c_i, c_j. This is possible because of the "naive" assumption that feature (word) likelihoods are independent, resulting in a classification metric that is linear for each feature. In S8 Appendix, we provide the derivation of this linearity structure.

We find that the trained NB classifier relies heavily on words that are very specific to the training set including the names of actors of the movies themselves, making them useful as classifiers but not in understanding the nature of the text. We report the top 10 words for both positive and negative classes using both the ratio and difference methods in Table in S8 Appendix. To classify a document using NB, we use the frequency of each word in the document in conjunction with the probability that that word occurred in each labeled class c_i. While steps can be taken to avoid this type of over-fitting, it is an ever-present danger that remains hidden without word shift graphs or similar.

We next take the movie-review-trained NB classifier and use it to classify the New York Times sections, both by ranking them and by looking at the words (the above ratio and difference weighted by the occurrence of the words). We ranked the sections 5 different times, and among those find the "Television" section both by far the happiest, and by far the least happy in independent tests. We show these rankings and report the top 10 words used to score the "Society" section in Table A.3.

We thus see that the NB classifier, a linear learning method, may perform poorly when assessing sentiment outside of the corpus on which it is trained. In general, performance will vary depending on the statistical dissimilarity of the training and novel corpora. Added to this is the inscrutability of black box methods: while susceptible to the aforementioned difficulty, nonlinear learning methods (unlike NB) also render detailed examination of how individual words contribute to a text's score more difficult.

2.4 Conclusion

We have shown that measuring sentiment in various corpora presents unique challenges, and that sentiment dictionary performance is situation dependent. Across the board, the ANEW dictionary performs poorly, and the continued use of this sentiment dictionary with clearly better alternatives is a questionable choice. We have seen that the MPQA dictionary does not agree with the other five dictionaries on the NYT corpus and Twitter corpus due to a variety of context, word sense, phrase, and stem matching issues, and we would not recommend using this sentiment dictionary. While the OL achieves the highest binary classification accuracy, in comparison to labMT, the WK, LIWC, and OL dictionaries fail to provide much detail in corpora where their coverage is lower, including all four corpora tested, the main goal of our analysis. Sufficient coverage is essential for producing meaningful word shift graphs and thereby enabling deeper understanding.

In each case, to analyze the output of the dictionary method, we rely on the use of word shift graphs. With this tool, we can produce a finer grained analysis of the lexical content, and we can also detect words that are used out of context and can mask them directly. It should be clear that using any of the dictionary-based sentiment detecting method without looking at how individual words contribute is indefensible, and analyses that do not use word shift graphs or similar tools cannot be trusted. The poor word shift performance of binary dictionaries in particular gravely limits their ability to reveal underlying stories.

In sum, we believe that dictionary-based methods will continue to play a powerful role—they are fast and well suited for web-scale data sets—and that the best instruments will be based on dictionaries with excellent coverage and continuum scores. To this end, we urge that all dictionaries should be regularly updated to capture changing lexicons, word usage, and demographics. Looking further ahead, a move from scoring words to scoring both phrases and words with senses should realize considerable improvement for many languages of interest. With phrase dictionaries, the resulting phrase shift graphs will allow for a more nuanced and detailed analysis of a corpus's sentiment score (Alajajian et al., 2016), ultimately affording clearer stories for sentiment dynamics.

2.5 References

Alajajian, S. E., J. R. Williams, A. J. Reagan, S. C. Alajajian, M. R. Frank, L. Mitchell, J. Lahne, C. M. Danforth, and P. S. Dodds (2016). The Lexicocalorimeter: Gauging public health through caloric input and output on social media. Available at http://arxiv.org/abs/1507.05098.

Baccianella, S., A. Esuli, and F. Sebastiani (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, Volume 10, pp. 2200–2204.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72. Association for Computational Linguistics.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*(1), 1–8.

Bradley, M. M. and P. J. Lang (1999). Affective norms for english words (ANEW): Stimuli, instruction manual and affective ratings. Technical report c-1, University of Florida, Gainesville, FL.

Cambria, E., D. Olsher, and D. Rajagopal (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, pp. 1515–1521. AAAI Press.

Chung, S. and S. Liu (2011). Predicting stock market fluctuations from Twitter. *Berkeley, California.*

De Smedt, T. and W. Daelemans (2012). Pattern for Python. *The Journal of Machine Learning Research 13*(1), 2063–2067.

Dodds, P. S., E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth (2015a). Human language reveals a universal positivity bias. *PNAS 112*(8), 2389–2394.

Dodds, P. S., E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth (2015b). Reply to garcia et al.: Common mistakes in measuring frequency-dependent word characteristics. *Proceedings of the National Academy of Sciences 112*(23), E2984–E2985.

Dodds, P. S. and C. M. Danforth (2009, July). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies 11*(4), 441–456.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011, 12). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE 6*(12), e26752.

Garcia, D., A. Garas, and F. Schweitzer (2015). The language-dependent relationship between word happiness and frequency. *Proceedings of the National Academy of Sciences 112*(23), E2983.

Golder, S. A. and M. W. Macy (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science Magazine 333*, 1878–1881.

Gonçalves, P., M. Araújo, F. Benevenuto, and M. Cha (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pp. 27–38. ACM.

Hand, D. J. and K. Yu (2001). Idiot's bayes—not so stupid after all? *International statistical review 69*(3), 385–398.

Hutto, C. J. and E. Gilbert (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*. AAAI Publications.

Kiritchenko, S., X. Zhu, and S. M. Mohammad (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research 50*, 723–762.

Levallois, C. (2013). Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, Volume 2, pp. 414–417.

Lin, Y., J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pp. 169–174. Association for Computational Linguistics.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing 2*, 627–666.

Liu, B. (2012, May). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science 331*(6014), 176–182.

Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth (2013, May). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE 8*(5), e64417.

Mohammad, S. M., S. Kiritchenko, and X. Zhu (2013, June). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

Nielsen, F. Å. (2011, May). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey (Eds.), *CEUR Workshop Proceedings*, Volume Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718, pp. 93–98.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pappas, N., G. Katsimpras, and E. Stamatatos (2013). Distinguishing the popularity between topics: A system for up-to-date opinion retrieval and mining in the web. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 197–209. Springer.

Pechenick, E. A., C. M. Danforth, and P. S. Dodds (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *arXiv preprint arXiv:1501.00960*.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001.

Poria, S., A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay (2013). Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems 28*(2), 31–38.

Rayner, J. M. V. (1985). Linear relations in biomechanics: the statistics of scaling functions. *J. Zool. Lond. (A) 206*, 415–439.

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto (2016, July). SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci. 5*(1), 23.

Ruiz, E. J., V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 513–522. ACM.

Sandhaus, E. (2008). The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng (2013). Exploiting topic based Twitter sentiment for stock prediction. In *ACL (2)*, pp. 24–29.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Volume 1631, pp. 1642. Citeseer.

Stone, P. J., D. C. Dunphy, and M. S. Smith (1966). The general inquirer: A computer approach to content analysis. *MIT Press*.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*(12), 2544–2558.

Warriner, A. B., V. Kuperman, and M. Brysbaert (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods 45*(4), 1191–1207.

Watson, D. and L. A. Clark (1999). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form: Manual for the positive and negative affect schedule-expanded form*. Ph. D. thesis, University of Iowa.

Whissell, C., M. Fournier, R. Pelland, D. Weir, and K. Makarec (1986). A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills 62*(3), 875–888.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Wojcik, S. P., A. Hovasapian, J. Graham, M. Motyl, and P. H. Ditto (2015). Conservatives report, but liberals display, greater happiness. *Science 347*(6227), 1243–1246.

Zhu, X., S. Kiritchenko, and S. M. Mohammad (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 443–447. Citeseer.

Chapter 3

The emotional arcs of stories are dominated by six basic shapes

Advances in computing power, natural language processing, and digitization of text now make it possible to study a culture's evolution through its texts using a "big data" lens. Our ability to communicate relies in part upon a shared emotional experience, with stories often following distinct emotional trajectories and forming patterns that are meaningful to us. Here, by classifying the emotional arcs for a filtered subset of 1,327 stories from Project Gutenberg's fiction collection, we find a set of six core emotional arcs which form the essential building blocks of complex emotional trajectories. We strengthen our findings by separately applying Matrix decomposition, supervised learning, and unsupervised learning. For each of these six core emotional arcs, we examine the closest characteristic stories in publication today and find that particular emotional arcs enjoy greater success, as measured by downloads.

3.1 Introduction

The power of stories to transfer information and define our own existence has been shown time and again (Pratchett et al., 2003; Campbell, 1949; Gottschall, 2013; Cave, 2013). We as people are fundamentally driven to find and tell stories, likened to *Pan Narrans* or *Homo Narrativus* (Dodds, 2013). Stories are encoded in art, language, and even in the mathematics of physics: We use equations to represent both simple and complicated functions that describe our observations of the real world. In science, we formalize the ideas that best fit our experience with principles such as Occam's Razor: The simplest story is the one we should trust. We tend to prefer stories that fit into the molds which are familiar, and reject narratives that do not align with our experience (Nickerson, 1998).

We seek here to better understand stories that are captured and shared in written form, a medium that since inception has radically changed how information flows (Gleick, 2011). Without evolved cues from tone, facial expression, or body language, written stories are forced to capture the entire transfer of experience on a page. An often integral part of a written story is the emotional experience that is evoked in the reader. Here, we use a simple, robust sentiment analysis tool to extract the reader-perceived emotional content of written stories as they unfold on the page.

We objectively test aspects of folkloristic theory (Propp, 1968; MacDonald, 1982), specifically the commonality of core stories within societal boundaries (Cave, 2013; da Silva and Tehrani, 2016). A major component of folkloristics is the study of society and culture through literary analysis. This is sometimes referred to as *narratology*, which at its core is "a series of events, real or fictional, presented to the reader or the listener" (Min and Park, 2016). In our present treatment, we consider the plot as the "backbone" of events that occur in a chronological sequence (more detail on previous theories of plot, and the framing we present next and adopt, are in Appendix B.1). While the plot captures the mechanics of a narrative and the structure encodes their delivery, in the present work we examine the emotional arc that is invoked through the words used. The emotional arc of a story does not give us direct information about the plot or the intended meaning of the story, but rather exists as part of the whole narrative (e.g., an emotional arc showing a fall in sentiment throughout a story may arise from very different plot and structure combinations). This distinction between the emotional arc and the plot of a story is one point of misunderstanding in other work that has drawn criticism from the digital humanities community (Jockers, 2014). Through the identification of motifs, narrative theories allow us to analyze, interpret, describe, and compare stories across cultures and regions of the world (Dundes, 1997; Dolby, 2008; Uther, 2011). We show that automated extraction of emotional arcs is not only possibly, but can test previous theories and provide new insights with the potential to quantify unobserved trends as the field transitions from data-scarce to data-rich (Kirschenbaum, 2007; Moretti, 2013).

The rejected master's thesis of Kurt Vonnegut—which he personally considered his greatest contribution—defines the *emotional arc* of a story on the "Beginning–End" and "Ill Fortune–Great Fortune" axes (Vonnegut, 1981). Vonnegut finds a remarkable similarity between Cinderella and the origin story of Christianity in the Old Testament (see Fig. B.1 in Appendix B.2), leading us to search for all such groupings. In a recorded lecture available on YouTube (Vonnegut, 1995), Vonnegut asserted:

> "There is no reason why the simple shapes of stories can't be fed into computers, they are beautiful shapes."

For our analysis, we apply three independent tools: Matrix decomposition by Singular Value Decomposition (SVD), supervised learning by agglomerative (hierarchical) clustering with Ward's method, and unsupervised learning by a Self Organizing Map (SOM, a type of neural network). Each tool has different strengths: the SVD finds the underlying basis of all of the emotional arcs, the clustering classifies the emotional arcs into distinct groups, and the SOM generates arcs from noise which are similar to those in our corpus using a stochastic process. By considering the results of each tool independently, we are able to confirm our findings of broad support.

We proceed as follows. We first introduce our methods in Section 3.2, we then discuss the combined results of each method in Section 3.3, and we present our conclusions in Section 3.4. A graphical outline of the methodology and results can be found as Fig. B.2 in Appendix B.2.

3.2 Methods

3.2.1 Emotional arc construction

To generate emotional arcs, we analyze the sentiment of 10,000 word windows, which we slide through the text (see Fig. 3.1). We rate the emotional content of each window using our Hedonometer with the labMT dataset, chosen for lexical coverage and its ability to generate meaningful word shift graphs, specifically using 10,000 words as a minimum necessary to generate meaningful sentiment

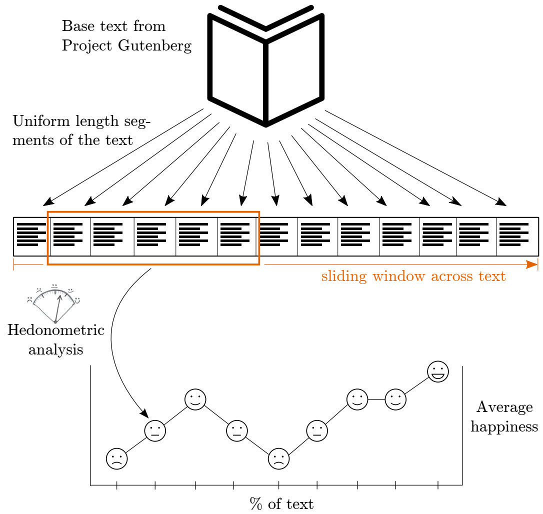


Figure 3.1: Schematic of how we compute emotional arcs. The indicated uniform length segments (gap between samples) taken from the text form the sample with fixed window size set at $N_w = 10,000$ words. The segment length is thus $N_s = (N - (N_w + 1))/n$ for N the length of the book in words, and n the number of points in the time series. Sliding this fixed size window through the book, we generate n sentiment scores with the Hedonometer, which comprise the emotional arc (Dodds et al., 2011).

scores (Reagan et al., 2015; Ribeiro et al., 2016). We emphasize that dictionary-based methods for sentiment analysis usually perform worse than random on individual sentences (Reagan et al., 2015; Ribeiro et al., 2016), and although this issue can be resolved by using a rolling average of sentences scores, it betrays a basic misunderstanding of similar efforts (Jockers, 2014). In Fig. 3.2, we show the emotional arc of *Harry Potter and the Deathly Hallows*, the final book in the popular Harry Potter series by J.K. Rowling. While the plot of the book is nested and complicated, the emotional arc associated with each sub-narrative is clearly visible. We analyze the emotional arcs corresponding to complete books, and to limit the conflation of multiple core emotional arcs, we restrict our analysis to shorter books by selecting a maximum number of words when building our filter. Further details of the emotional arc construction can be found in Appendix B.3.

3.2.2 Project Gutenberg Corpus

For a suitable corpus we draw on the open access Project Gutenberg data set (Various, Various). We apply rough filters to the collection (roughly 50,000 books) in an attempt to obtain a set of books that represent English works of fiction. We start by selecting for only English books, with

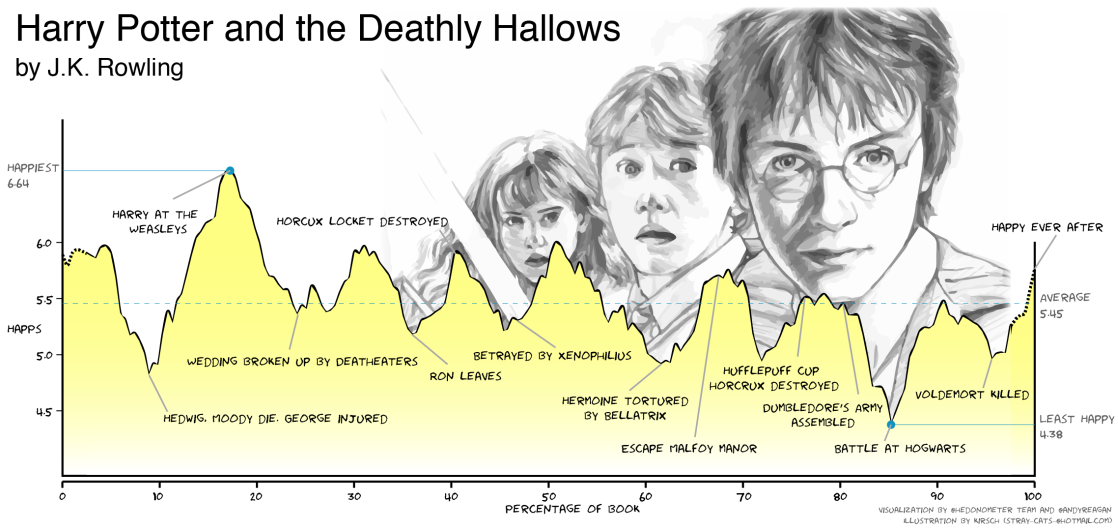


Figure 3.2: Annotated emotional arc of Harry Potter and the Deathly Hallows, by J.K. Rowling, inspired by the illustration made by Medaris for The Why Files (Tenenbaum et al., 2015). The entire seven book series can be classified as a "Kill the monster" plot (Booker, 2006), while the many sub plots and connections between them complicate the emotional arc of each individual book: this plot could not be readily inferred from the emotional arc alone. The emotional arc shown here captures the major highs and lows of the story, and should be familiar to any reader well acquainted with Harry Potter. Our method does not pick up emotional moments discussed briefly, perhaps in one paragraph or sentence (e.g., the first kiss of Harry and Ginny). We provide interactive visualizations of all Project Gutenberg books at http://hedonometer.org/books/v3/1/ and a selection of classic and popular books at http://hedonometer.org/books/v1/.

total words between 20,000 and 100,000, with more than 40 downloads from the Project Gutenberg website, and with Library of Congress Class corresponding to English fiction[1]. To ensure that the 40-download limit is not influencing the results here, we repeat the entire analysis for each method with 10, 20, 40, and 80 download thresholds, in each case confirming the 40 download findings to be qualitatively unchanged. Next, we remove books with any word in the title from a list of keywords (e.g., "poems" and "collection", full list in Appendix B.3). From within this set of books, we remove the front and back matter of each book using regular expression pattern matches that match on 98.9% of the books included. Two slices of the data for download count and the total word count are shown in Appendix B.3 Fig. B.4. We provide a list of the book ID's which are included for download in the Online Appendices at http://compstorylab.org/share/papers/reagan2016b/, the books are listed in Table B.1 in Appendix B.4, and we attempt to provide the Project Gutenberg ID when we mention a book by title herein. Given the Project Gutenberg ID n, the raw ebook is available online from Project Gutenberg at http://www.gutenberg.org/ebooks/n, e.g., *Alice's Adventures in Wonderland* by Lewis Carroll, has ID 11 and is available at http://www.gutenberg.org/ebooks/11. We also provide an online, interactive version of the emotional arc for each book indexed by the ID, e.g., *Alice's Adventures in Wonderland* is available at http://hedonometer.org/books/v3/11/.

3.2.3 Principal Component Analysis (SVD)

We use the standard linear algebra technique Singular Value Decomposition (SVD) to find a decomposition of stories onto an orthogonal basis of emotional arcs. Starting with the emotional arc (sentiment time series) for each book b_i as row i in the matrix A, we apply the SVD to find

$$A = U\Sigma V^T = WV^T, \tag{3.1}$$

where U contains the projection of each sentiment time series onto each of the right singular vectors (rows of V^T, eigenvectors of A^TA), which have singular values given along the diagonal of Σ, with $W = U\Sigma$. Different intuitive interpretations of the matrices U, Σ, and V^T are useful in the various domains in which the SVD is applied; here, we focus on right singular vectors as an orthonormal basis for the sentiment time series in the rows of A, which we will refer to as the *modes*. We combine

[1]The specific classes have labels PN, PR, PS, and PZ.

Σ and U into the single coefficient matrix W for clarity and convenience, such that W now represents the mode coefficients.

3.2.4 Hierarchical Clustering

We use Ward's method to generate a hierarchical clustering of stories, which proceeds by minimizing variance between clusters of books (Ward Jr, 1963). We use the mean-centered books and the distance function

$$D(b_i, b_j) = l^{-1} \sum_{t=1}^{l} |b_i(t) - b_j(t)|. \tag{3.2}$$

for t indexing the window in books b_i, b_j to generate the distance matrix.

3.2.5 Self Organizing Map (SOM)

We implement a Self Organized Map (SOM), an unsupervised machine learning method (a type of neural network) to cluster emotional arcs (Kohonen, 1990). The SOM works by finding the most similar emotional arc in a random collection of arcs. We use an 8x8 SOM (for 64 nodes, roughly 5% of the number of books), connected on a square grid, training according to the original procedure (with winner take all, and scaling functions across both distance and magnitude). We take the neighborhood influence function at iteration i as

$$\text{Nbd}_k(i) = \left[j \in \mathcal{N} \mid D(k, j) < \sqrt{N} \cdot (i + 1)^{\alpha} \right] \tag{3.3}$$

for a node k in the set of nodes $\mathcal{N}$, with distance function D given above and total number of nodes N. For results shown here we take $\alpha = -0.15$. We implement the learning adaptation function at training iteration i as $f(i) = (i + 1)^{\beta}$, again with $\beta = -0.15$, a standard value for the training hyper-parameters.

3.3 Results

We obtain a collection of 1,327 books that are mostly, but not all, fictional stories by using metadata from Project Gutenberg to construct a rough filter. We find broad support for the following six emotional arcs:

- "Rags to riches" (rise).
- "Tragedy", or "Riches to rags" (fall).
- "Man in a hole" (fall-rise).
- "Icarus" (rise-fall).
- "Cinderella" (rise-fall-rise).
- "Oedipus" (fall-rise-fall).

Importantly, we obtain these same six emotional arcs from all possible arcs by observing them as the result of three methods: As modes from a matrix decomposition by SVD, as clusters in a hierarchical clustering using Ward's algorithm, and as clusters using unsupervised machine learning. We examine each of the results in this section.

3.3.1 Principal Component Analysis (SVD)

In Fig. 3.3 we show the leading 12 modes in both the weighted (dark) and un-weighted (lighter) representation. In total, the first 12 modes explain 80% and 94% of the variance from the mean centered and raw time series, respectively. The modes are from mean-centered emotional arcs, such that the first SVD mode need not extract the average from the labMT scores nor the positivity bias present in language (Dodds et al., 2015). The coefficients for each mode within a single emotional arc are both positive and negative, so we need to consider both the modes and their negation. We can immediately recognize the familiar shapes of core emotional arcs in the first four modes, and compositions of these emotional arcs in modes 5 and 6. We observe "Rags to riches" (mode 1, positive), "Tragedy" or "Riches to rags" (mode 1, negative), Vonnegut's "Man in a hole" (mode

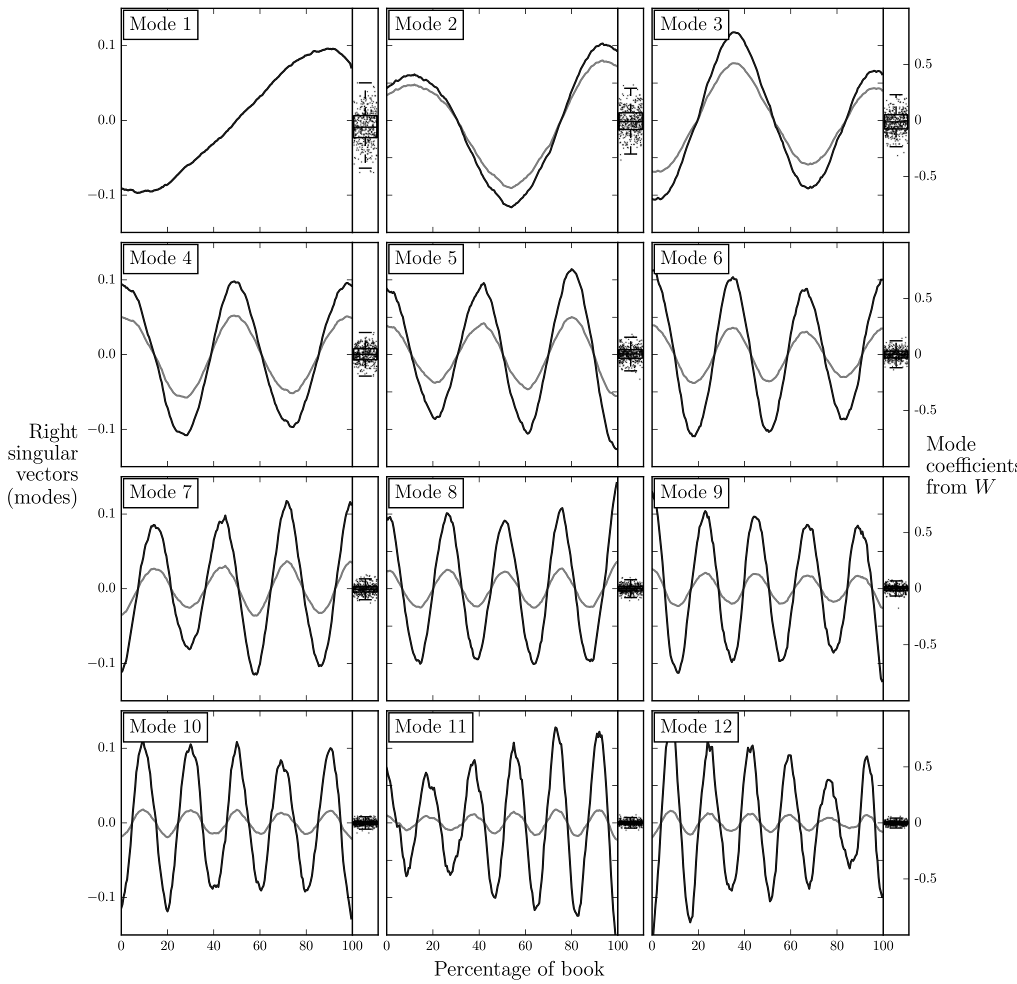


Figure 3.3: Top 12 modes from the Singular Value Decomposition of 1,327 Project Gutenberg books. We show in a lighter color modes weighted by their corresponding singular value, where we have scaled the matrix Σ such that the first entry is 1 for comparison (for reference, the largest singular value is 34.5). The mode coefficients normalized for each book are shown in the right panel accompanying each mode, in the range -1 to 1, with the "Tukey" box plot.

2, positive), "Icarus" (mode 2, negative), "Cinderella" (mode 3, positive), "Oedipus" (mode 3, negative). We choose to include modes 7–12 only for completeness, as these high frequency modes have little contribution to variance and do not align with core emotional arc archetypes from other methods (more below).

We emphasize that by definition of the SVD, the mode coefficients in W can be either positive and negative, such that the modes themselves explain variance with both the positive and negative version. In the right panels of each mode in Fig. 3.3 we project the 1,327 stories onto each of first six modes and show the resulting coefficients. While none are far from 0 (as would be expected), mode 1 has a mean slightly above 0 and both modes 3 and 4 have means slightly below 0. To sort the books by their coefficient for each mode, we normalize the coefficients within each book in the rows of W to sum to 1, accounting for books with higher total energy, and these are the coefficients shown in the right panels of each mode in Fig. 3.3. In Appendix B.5, we provide supporting, intuitive details of the SVD method, as well as example emotional arc reconstruction using the modes (see Figs. B.5–B.7). As expected, less than 10 modes are enough to reconstruct the emotional arc to a degree of accuracy visible to the eye.

We show labeled examples of the emotional arcs closest to the top 6 modes in Figs. 3.4 and B.8.

We present both the positive and negative modes, and the stories closest to each by sorting on the coefficient for that mode. For the positive stories, we sort in ascending order, and vice versa. Mode 1, which encompasses both the "Rags to riches" and "Tragedy" emotional arcs, captures 30% of the variance of the entire space. We examine the closest stories to both sides of modes 1–3, and direct the reader to Fig. B.8 for more details on the higher order modes. The two stories that have the most support from the "Rags to riches" mode are *The Winter's Tale* (1539) and *Oscar Wilde, Art and Morality: A Defence of "The Picture of Dorian Gray"* (33689). Among the most categorical tragedies we find *Lady Susan* (946) and *Warlord of Kor* (17958). Number 8 in the sorted list of tragedies is perhaps the most famous tragedy: *Romeo and Juliet* by William Shakespeare. Mode 2 is the "Man in a hole" emotional arc, and we find the stories which most closely follow this path to be *The Magic of Oz* (419) and *Children of the Frost* (10736). The negation of mode 2 most closely resembles the emotional arc of the "Icarus" narrative. For this emotional arc, the most characteristic stories are *Shadowings* (34215) and *Battle-Pieces and Aspects of the War* (12384). Mode 3 is the "Cinderella" emotional arc, and includes *Mystery of the Hasty Arrow* (17763) and *Through the Magic Dorr* (5317). The negation of Mode 3, which we refer to as "Oedipus", is found most characteristically in *This World is Taboo* (18172), *Old Indian Days* (339), and *The Evil Guest* (10377). We also note that the spread of the stories from their core mode increases strongly for the higher modes.

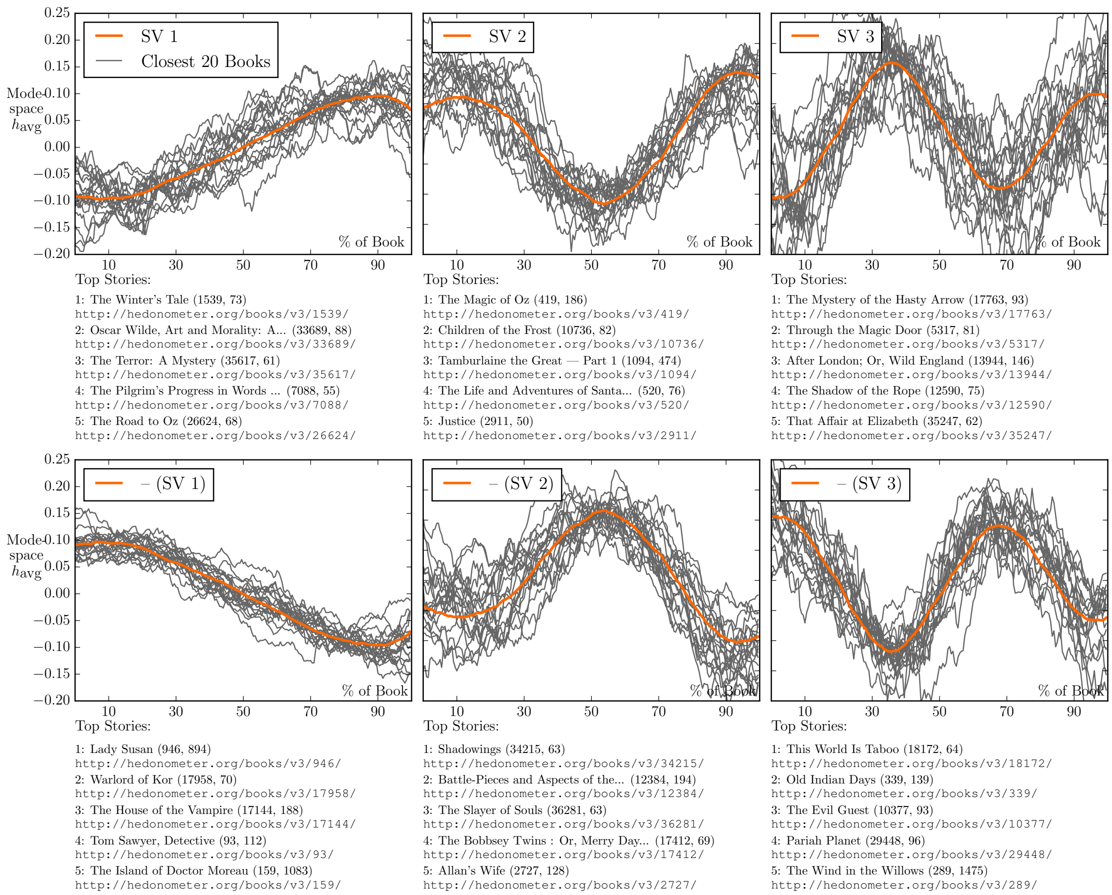


Figure 3.4: First 3 SVD modes and their negation with the closest stories to each. To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W. In parentheses for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

3.3.2 Hierarchical Clustering

We show a dendrogram of the 60 clusters with highest linkage cost in Fig. 3.5. The average silhouette coefficient is shown on the bottom of Fig. 3.5, and the distributions of silhouette values within each cluster (see Figs. B.17–B.18) can be used to analyze the appropriate number of clusters (Rousseeuw, 1987). A characteristic book from each cluster is shown on the leaf nodes by sorting the books within each cluster by the total distance to other books in the cluster (e.g., considering each intra-cluster collection as a fully connected weighted network, we take the most central node), and in parenthesis the number of books in that cluster. In other words, we label each cluster by considering the network centrality of the fully connected cluster with edges weighted by the distance between stories. By cutting the dendrogram in Fig. 3.5 at various linkage costs we are able to extract clusters of the desired granularity. For the cuts labeled C2, C4, and C8, we show these clusters in Figs. B.9, B.11, and B.15. We find the first four of our final six arcs appearing among the eight most different clusters (Fig. B.15).

The clustering method groups stories with a “Man in a hole” emotional arc for a range of different variances, separate from the other arcs. In total these clusters (Panel A, E, and I of Fig. B.16) account for 30% of the Gutenberg corpus. The remainder of the stories have emotional arcs that are clustered among the “Tragedy” arc (32%), “Rags to riches” arc (5%), and the “Oedipus” arc (31%). A more detailed analysis of the results from hierarchical clustering can be found in Appendix B.6, and this result generally agrees with other attempts that use only hierarchical clustering (Jockers, 2015).a

3.3.3 Self Organizing Map (SOM)

Finally, we apply Kohonen's Self-Organizing Map (SOM) and find core arcs from unsupervised machine learning on the emotional arcs. On the two dimensional component plane, the prescribed network topology, we find seven spatially coherent groups, with five emotional arcs. These spatial groups are comprised of stories with core emotional arcs of differing variance.

In Fig. 3.6 we see both the B-Matrix to demonstrate the strength of spatial clustering and a heat-map showing where we find the winning nodes. The A–I labels refer to the individual nodes

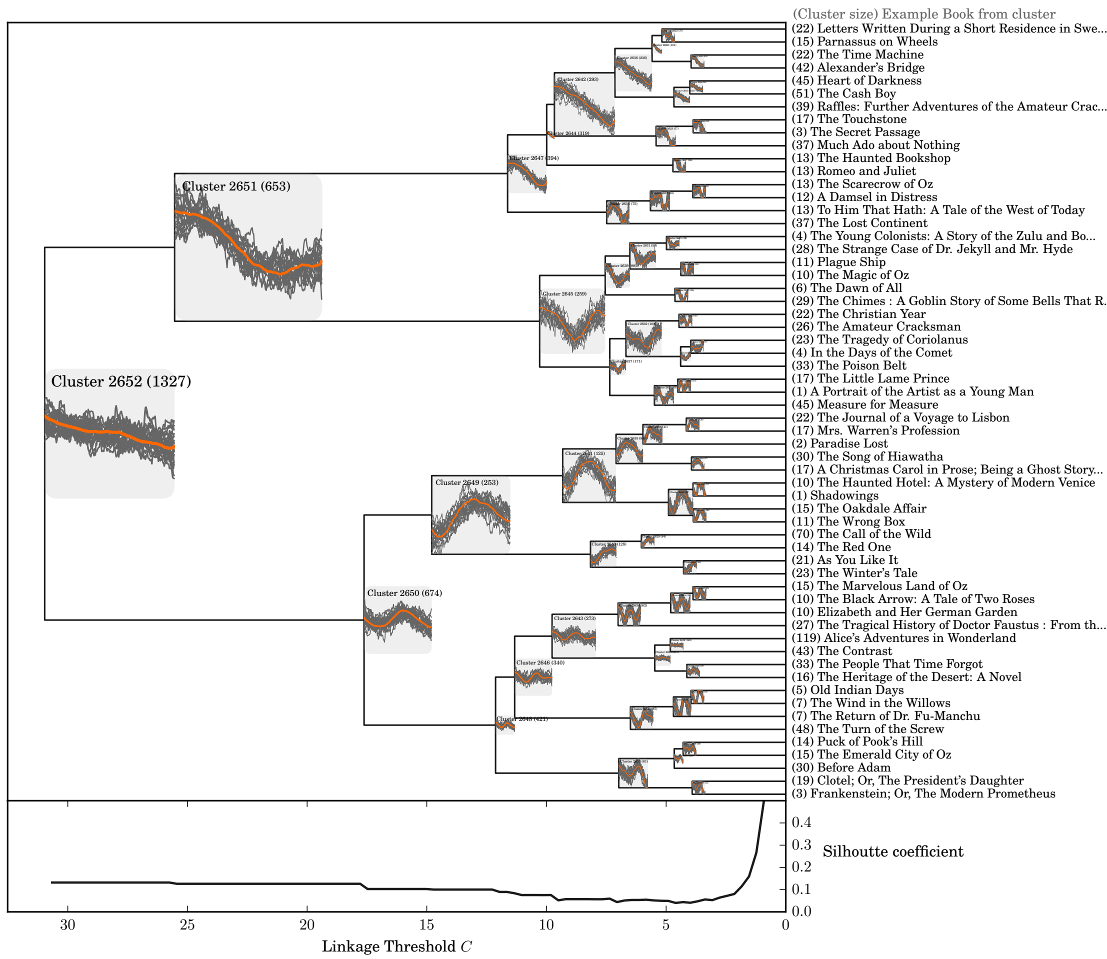


Figure 3.5: Dendrogram from the hierarchical clustering procedure using Ward's minimum variance method. For each cluster, a selection of the 20 most central books to a fully-connected network of books are shown along with the average of the emotional arc for all books in the cluster, along with the cluster ID and number of books in each cluster (shown in parenthesis). The cluster ID is given by numbering the clusters in order of linkage starting at 0, with each individual book representing a cluster of size 1 such that the final cluster (all books) has the ID $2(N-1)$ *for the* $N = 1,327$ *books. At the bottom, we show the average Silhouette value for all books, with higher value representing a more appropriate number of clusters. For each of the 60 leaf nodes (right side) we show the number of books within the cluster and the most central book to that cluster's book network.*

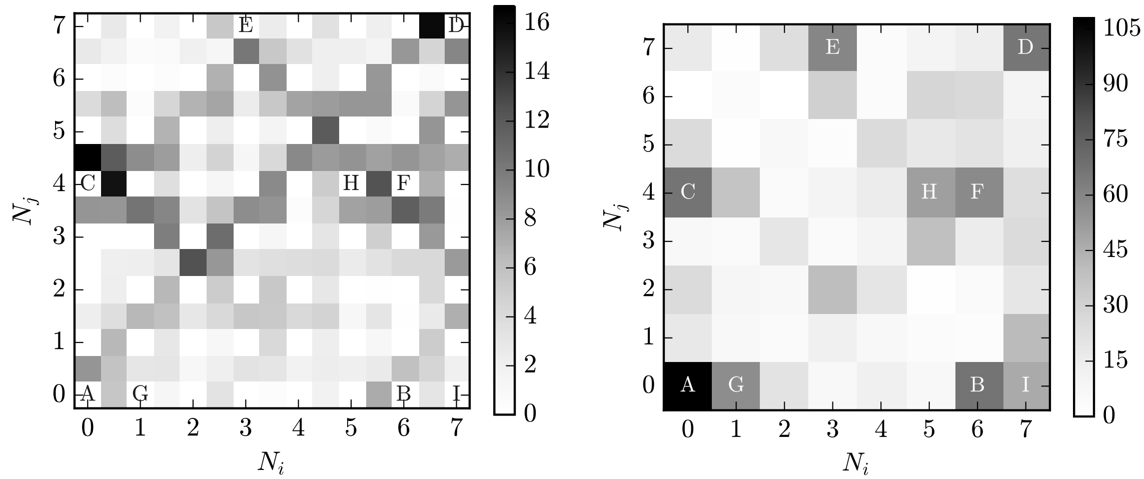


Figure 3.6: Results of the SOM applied to Project Gutenberg books. Left panel: Nodes on the 2D SOM grid are shaded by the number of stories for which they are the winner. Right panel: The B-Matrix shows that there are clear clusters of stories in the 2D space imposed by the SOM network.

shown in Fig. B.19, and we observe seven spatial groups in both panels of Fig. 3.6: (1) A and G, (2) B and I, (3) C, (4) D, (5) E, and (6) H, and (7) F. These spatial clusters reinforce the visible similarity of the winning node arcs, given that nodes H and F are close spatially but separated by the B-Matrix and contain very distinct arcs. We show the winning node emotional arcs and the arcs of books for which they are the winners in Fig. B.19. The legend shows the node ID, numbers the cluster by size, and in parentheses indicates the size of the cluster on that individual node. In Panels A and G we see varying strengths of the "Man in a hole" emotional arc. In Panels B and I, the second largest individual cluster consists of the "Rags to riches" arcs. In Panel C, and in Panel F, we find the "Oedipus" emotional arc, with a more pronounced positive start and decline in Panel C. In Panel D we see the "Icarus" arc, and in Panel E and Panel H we see the "Tragedy" arc. Each of these top stories are all readily identifiable, yet again demonstrating the universality of these story types.

3.3.4 Null comparison

There are many possible emotional arcs in the space that we consider. To demonstrate that these specific arcs are uniquely compelling as stories written by and for *homo narrativus*, we consider the true emotional arcs in relation to their most suitable comparison: the book with randomly shuffled words ("word salad") and the resulting text from a 2-gram Markov model trained on the individual book itself ("nonsense"). We chose to compare to "word salad" and "nonsense" versions as they are more representative of a null model: written stories that are without coherent plot or structure to generate a coherent emotional arc, which is not true of a stochastic process (e.g., a random walk model or noise). Examples of the emotional arc and null emotional arcs for a single book are shown in Fig. B.20, with 10 "word salad" and "nonsense" versions. Sampled text using each method is given in Appendix B.3. We re-run each method on the English fiction Gutenberg Corpus with the null versions of each book and verify that the emotional arcs of real stories are not simply an artifact. The singular value spectrum from the SVD is flatter, with higher-frequency modes appearing more quickly, and in total representing 45% of the total variance present in real stories (see Figs. B.22 and B.25). Hierarchical clustering generates less distinct clusters with considerably lower linkage cost (final linkage cost 1400 vs 7000) for the emotional arcs from nonsense books, and the winning node vectors on a self-organizing map lack coherent structure (see Figs. B.26 and B.29 in Appendix B.8).

3.3.5 The Success of Stories

To examine how the emotional trajectory impacts success, in Fig. 3.7 we examine the downloads for all of the books that are most similar to each SVD mode (for additional modes, see Fig. B.3 in Appendix B.2). We find that the first four modes, which contain the greatest total number of books, are not the most popular. Along with the negative of mode 2, both polarities of modes 3 and 4 have markedly higher median downloads, while we discount the importance of the mean with the high variance. The success of the stories underlying these emotional arcs suggests that the emotional experience of readers strongly affects how stories are shared. We find "Icarus" (-SV 2), "Oedipus" (-SV 3), and two sequential "Man in a hole" arcs (SV 4), are the three most successful emotional arcs. These results are influenced by individual books within each mode which have high numbers of downloads, and we refer the reader to the download-sorted tables for each mode in Appendix B.5.

Mode	Mode Arc	N_m	N_m/N	DL Median ▼	DL Mean ▽	DL Variance	% > Average	Download Distribution
SV 1		133	10.0%	80.0	296.0	826779	17.3%	
- SV 1		407	30.7%	83.0	255.2	477221	14.5%	
SV 2		148	11.2%	76.0	240.9	319929	12.2%	
- SV 2		171	12.9%	97.0	251.6	252737	18.7%	
SV 3		73	5.5%	89.0	221.4	297604	12.3%	
- SV 3		139	10.5%	94.0	361.5	1280847	16.5%	
SV 4		66	5.0%	105.5	496.9	1937690	18.2%	
- SV 4		50	3.8%	90.0	195.6	107131	14.0%	
SV 5		46	3.5%	86.0	597.8	6462567	19.6%	

Figure 3.7: Download statistics for stories whose SVD Modes comprise more than 2.5% of books, for N the total number of books and N_m the number corresponding to the particular mode. Modes SV 3 through -SV 4 (both polarities of modes 3 and 4) exhibit a higher average number of downloads and more variance than the others. Mode arcs are rows of V^T and the download distribution is show in $\log_{10}$ space from 20 to 30,000 downloads.

3.4 Conclusion

Using three distinct methods, we have demonstrated that there is strong support for six core emotional arcs. Our methodology brings to bear a cross section of data science tools with a knowledge of the potential issues that each present. We have also shown that consideration of the emotional arc for a given story is important for the success of that story. Of course, downloads are only a rough proxy for success, and this work may provide an outline for more detailed analysis of the factors that impact meaningful measures of success, i.e., sales or cultural influence.

Our approach could be applied in the opposite direction: namely by beginning with the emotional arc and aiding in the generation of compelling stories (Li et al., 2013). Understanding the emotional arcs of stories may be useful to aid in constructing arguments (Bex and Bench-Capon, 2010) and teaching common sense to artificial intelligence systems (Riedl and Harrison, 2015).

Extensions of our analysis that use a more curated selection of full-text fiction can answer more detailed questions about which stories are the most popular throughout time, and across regions (da Silva and Tehrani, 2016). Automatic extraction of character networks would allow a more detailed analysis of plot structure for the Project Gutenberg corpus used here (Bost et al., 2016; Prado et al., 2016; Min and Park, 2016). Bridging the gap between the full text stories (Nenkova and McKeown, 2012) and systems that analyze plot sequences will allow such systems to undertake studies of this scale (Winston, 2011). Place could also be used to consider separate character networks through time, and to help build an analog to Randall Munroe's Movie Narrative Charts (Munroe, 2009).

We are producing data at an ever increasing rate, including rich sources of stories written to entertain and share knowledge, from books to television series to news. Of profound scientific interest will be the degree to which we can eventually understand the full landscape of human stories, and data driven approaches will play a crucial role.

PSD and CMD acknowledge support from NSF Big Data Grant #1447634.

3.5 References

Bex, F. J. and T. J. Bench-Capon (2010). Persuasive stories for multi-agent argumentation. In *AAAI Fall Symposium: Computational Models of Narrative*, Volume 10, pp. 04.

Booker, C. (2006). *The Seven Basic Plots: Why We Tell Stories*. New York: Bloomsbury Academic.

Bost, X., V. Labatut, and G. Linarès (2016). Narrative smoothing: dynamic conversational network for the analysis of tv series plots.

Campbell, J. (1949). *The Hero with a Thousand Faces* (third ed.). California: New World Library.

Cave, S. (2013, Jul). The 4 stories we tell ourselves about death. http://www.ted.com/talks/stephen_cave_the_4_stories_we_tell_ourselves_about_death.

da Silva, S. G. and J. J. Tehrani (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science 3*(1).

Dodds, P. S. (2013). Homo Narrativus and the trouble with fame. Nautilus Magazine. http://nautil.us/issue/5/fame/homo-narrativus-and-the-trouble-with-fame.

Dodds, P. S., E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth (2015). Human language reveals a universal positivity bias. *PNAS 112*(8), 2389–2394.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011, 12). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE 6*(12), e26752.

Dolby, S. K. (2008). *Literary Folkloristics and the Personal Narrative*. Indiana: Trickster Press.

Dundes, A. (1997). The motif-index and the tale type index: A critique. *Journal of Folklore Research*, 195–202.

Gleick, J. (2011). *The Information: A History, A Theory, A Flood*. New York: Pantheon.

Gottschall, J. (2013). *The Storytelling Animal: How Stories Make Us Human*. New York, NY: Mariner Books.

Jockers, M. (2014, Jun). A novel method for detecting plot. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

Jockers, M. (2015, Feb). The rest of the story. http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/.

Kirschenbaum, M. G. (2007). The remaking of reading: Data mining and the digital humanities. In *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Maryland*.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

Li, B., S. Lee-Urban, G. Johnston, and M. Riedl (2013). Story generation with crowdsourced plot graphs. In *AAAI*.

MacDonald, M. R. (1982). *Storytellers Sourcebook: A Subject, Title, and Motif Index to Folklore Collections for Children*. Michigan: Gale Group.

Min, S. and J. Park (2016). Narrative as a complex network: A study of Victor Hugo's les misérables. In *Proceedings of HCI Korea*.

Moretti, F. (2013). *Distant Reading*. New York: Verso.

Munroe, R. (2009, 11). Movie narrative charts. http://xkcd.com/657/.

Nenkova, A. and K. McKeown (2012). A survey of text summarization techniques. In *Mining text data*, pp. 43–76. Berlin, Germany: Springer.

Nickerson, R. S. (1998). Confirmation Bias; A ubiquitous phenomenon in many guises. *Review of General Psychology 2*, 175–220.

Prado, S. D., S. R. Dahmen, A. L. C. Bazzan, P. M. Carron, and R. Kenna (2016). Temporal network analysis of literary texts.

Pratchett, T., I. Stewart, and J. Cohen (2003). *The Science of Discworld II: The Globe*. London, UK: Ebury Press.

Propp, V. (1968). *Morphology of the Folktale. 1928*. Texas: Texas University Press.

Reagan, A., B. Tivnan, J. R. Williams, C. M. Danforth, and P. S. Dodds (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. Preprint available at https://arxiv.org/abs/1512.00531.

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto (2016, July). SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci. 5*(1), 23.

Riedl, M. O. and B. Harrison (2015). Using stories to teach human values to artificial agents.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Tenenbaum, D. J., K. Barrett, S. Medaris, and T. Devitt (2015, February). In 10 languages, happy words beat sad ones. http://whyfiles.org/2015/in-10-languages-happy-words-beat-sad-ones/.

Uther, H.-J. (2011). *The Types of International Folktales. A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson. Part I. Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction (FF Communications, 284)*. Helsinki, Finland: Finnish Academy of Science and Letters.

Various. Project Gutenberg.

Vonnegut, K. (1981). *Palm Sunday*. New York: RosettaBooks LLC.

Vonnegut, K. (1995). Shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Winston, P. H. (2011). The strong story hypothesis and the directed perception hypothesis.

Chapter 4

Selected contributions to published work

Throughout the course of my studies at the University of Vermont, I have enjoyed the collaborative research environment afforded by the Computational Story Lab [1]. As a result of these collaborations, I have assisted in the preparation of 10 other research papers. I have variously done data visualization work, curated data from our Twitter database, built interactive online appendices, and assisted in performing mathematical analysis. In this Chapter, I detail my contributions to each of these 10 papers, beginning with the paper abstract and then discussing my personal contribution.

[1]In this Chapter I use the singular first person noun in place of the plural pronoun to discuss my individual contributions

4.1 Collective Philanthropy: Describing and Modeling the Ecology of Giving

The first paper is *Collective Philanthropy: Describing and Modeling the Ecology of Giving* by William L. Gottesman, Andrew James Reagan, and Peter Sheridan Dodds, cited as Gottesman et al. (2014).

4.1.1 Abstract

Reflective of income and wealth distributions, philanthropic gifting appears to follow an approximate power-law size distribution as measured by the size of gifts received by individual institutions. We explore the ecology of gifting by analyzing data sets of individual gifts for a diverse group of institutions dedicated to education, medicine, art, public support, and religion. We find that the detailed forms of gift-size distributions differ across but are relatively constant within charity categories. We construct a model for how a donor's income affects their giving preferences in different charity categories, offering a mechanistic explanation for variations in institutional gift-size distributions. We discuss how knowledge of gift-sized distributions may be used to assess an institution's gift-giving profile, to help set fund-raising goals, and to design an institution-specific giving pyramid.

4.1.2 Contribution

In this paper I prepared final versions of each visualization in the paper, working from the initial designs from both Professor Dodds and Bill Gottesman, and working closely with Professor Dodds in their preparation. Additionally and at the request of the reviewers, I performed the statistical tests for support of power law distributions discussed in the paper, and included in the Appendix. In addition to testing for support of power law distributions using the MLE estimator Clauset et al. (2009), I ran likelihood comparison tests across many distributions, which we argue in the manuscript are potentially more applicable here to determine the most appropriate distribution. In Figure The parameters for the various distributions mentioned in the paper are written using LaTeX variables,

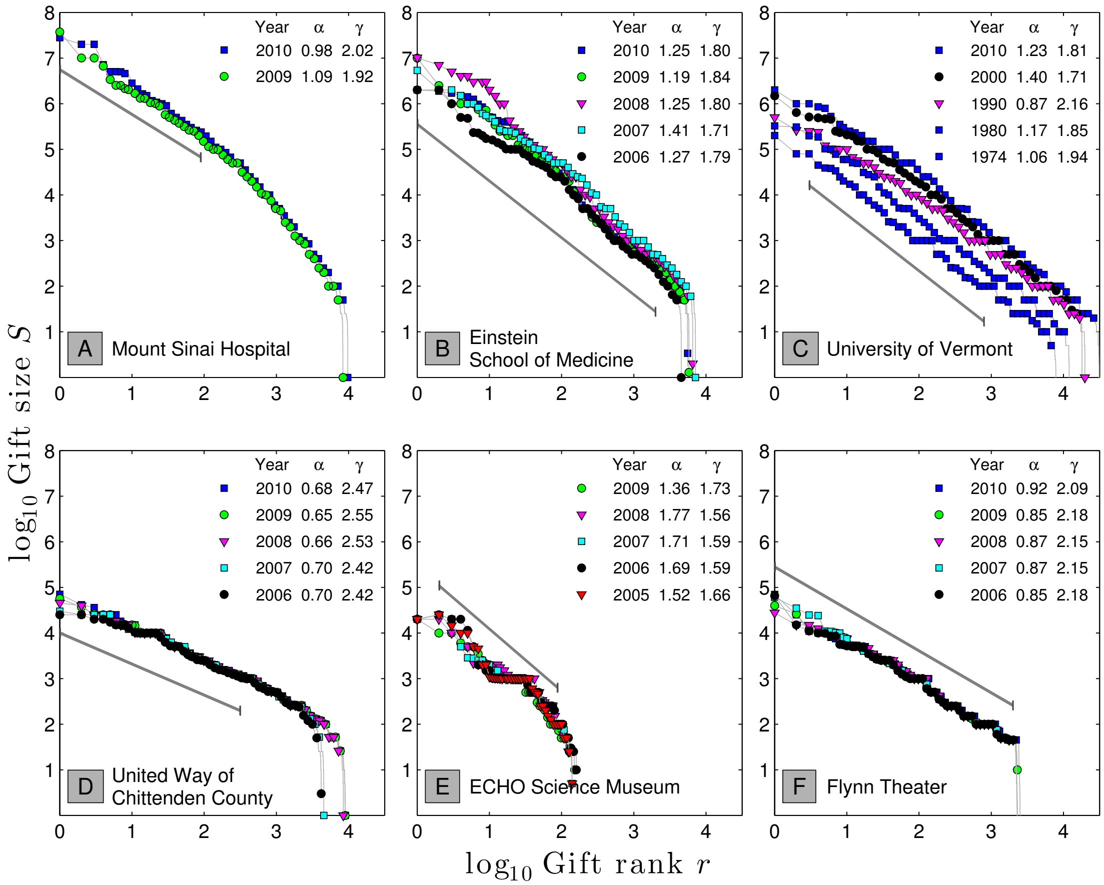


Figure 4.1: A reprint of Figure 1 from Gottesman et al. (2014), part of the caption is as follows: "Gift size distributions for a range of institutions. The reported α and γ were fitted to the region indicated by solid gray line, and the 95% CI of this fit, as well as year for which the fit is plotted, are included for each organization. The ranges over which the data were fit was chosen empirically; other approaches were found to be inconsistent (see Supplementary)."

written in a .tex file by the MATLAB and Python scripts that perform the statistical procedures. To the extend possible, all figures and analysis can be reproduced by running a single script. In this Section we include a reprint of Figure 1, Figure S1, and the power law fit tables from the paper. The codebase for creating the figures and performing the statistical procedures is available at https://github.com/andyreagan/philanthropy-distributions-code.

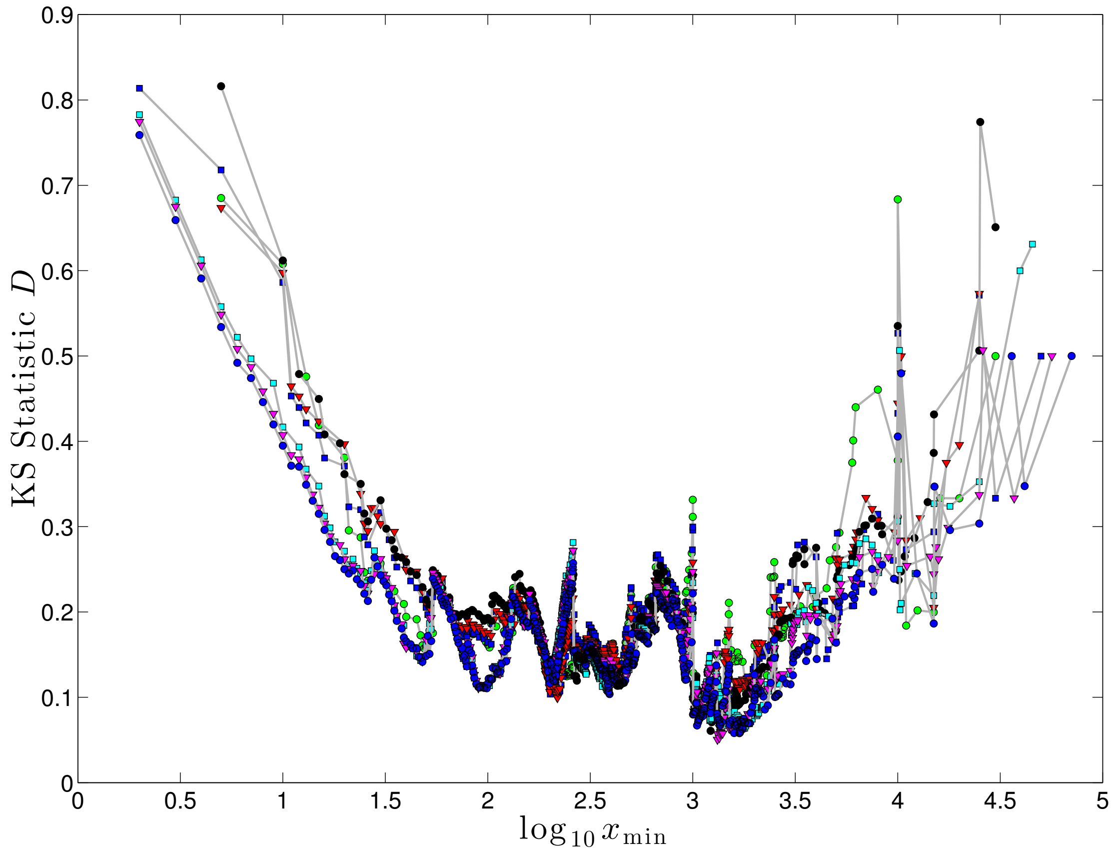


Figure 4.2: A reprint of Figure S1 from Gottesman et al. (2014), part of the caption is as follows: "The Kolmogorov-Smirnoff statistic D plotted over the log of $x_{\min}$, the minimum value fit for power law behavior, for the United Way of Chittenden County over the years 2006-2010. D is generated from the ML estimate. Existence of multiple minima in our data indicate that there are multiple possible fitting regions for which the KS statistic details a good fit. The variability of this value over each year plotted produced widely varying scaling parameters γ, and thus cannot be used without actually looking at the data."

Institution	**Year**	$\langle \mathbf{x} \rangle$	σ	$\mathbf{x}_{\max}$	γ	Range	**D**	**p**
Mount Sinai Hospital	2009	17618.40	450408.65	37259947	1.92 ± 0.08	1 to 90	0.12	0.00
	2010	19348.18	429587.88	27885708	2.02 ± 0.10	1 to 90	0.10	0.00
Einstein School of Medicine	2006	3247.30	46940.29	2000000	1.79 ± 0.02	1 to 2000	0.11	0.00
	2007	4768.09	78762.48	5350000	1.71 ± 0.01	1 to 2000	0.15	0.00
	2008	10385.80	199751.68	10200000	1.80 ± 0.01	1 to 2000	0.21	0.00
	2009	5212.92	139468.89	10000000	1.84 ± 0.01	1 to 2000	0.15	0.00
	2010	4917.94	61893.49	2000000	1.80 ± 0.06	1 to 2000	0.15	0.00
Univeristy of Vermont	1974	155.76	2811.94	200000	1.94 ± 0.01	3 to 794	0.18	0.00
	1980	284.31	5284.36	326000	1.85 ± 0.03	3 to 794	0.11	0.00
	1990	350.23	5382.45	500000	2.16 ± 0.01	3 to 794	0.38	0.00
	2000	805.33	15120.53	1488000	1.71 ± 0.03	3 to 794	0.09	0.00
	2010	741.40	17029.10	2000000	1.81 ± 0.05	3 to 794	0.13	0.00
United Way, Chittendon County	2004	441.71	1133.02	30000	2.77 ± 0.04	1 to 316	0.21	0.00
	2005	464.47	1444.26	50000	2.58 ± 0.22	1 to 316	0.13	0.00
	2006	456.86	1199.92	25000	2.42 ± 0.05	1 to 316	0.07	0.00
	2007	456.16	1279.14	30000	2.42 ± 0.14	1 to 316	0.07	0.00
	2008	287.53	1089.92	45460	2.53 ± 0.00	1 to 316	0.14	0.00
	2009	278.93	1122.44	56500	2.55 ± 0.08	1 to 316	0.12	0.00
	2010	287.58	1271.10	70518	2.47 ± 0.09	1 to 316	0.08	0.00
ECHO Science Museum	2005	977.77	3153.41	25000	1.66 ± 0.03	2 to 88	0.20	0.00
	2006	951.16	3415.22	25000	1.59 ± 0.02	2 to 88	0.28	0.00
	2007	941.61	3161.08	25000	1.59 ± 0.07	2 to 88	0.31	0.00
	2008	956.88	2688.31	20000	1.56 ± 0.01	2 to 88	0.26	0.00
	2009	676.84	2098.96	20000	1.73 ± 0.15	2 to 88	0.17	0.00
Flynn Theater	2006	241.87	1528.82	65065	2.18 ± 0.04	1 to 2000	0.26	0.00
	2007	268.54	1732.33	60000	2.15 ± 0.05	1 to 2000	0.25	0.00
	2008	248.00	1015.39	27500	2.15 ± 0.00	1 to 2000	0.22	0.00
	2009	242.90	1212.42	40000	2.18 ± 0.04	1 to 2000	0.23	0.00
	2010	246.13	1606.43	70000	2.09 ± 0.05	1 to 2000	0.22	0.00

Table 4.1: Summary statistics of all of the donation data is presented. The reported γ and range are fit with the MLE method, and the x_{min} which was found to minimize the Kolmogorov-Smirnoff statistc **D** *is reported along with* **D** *itself. In this case, lower values of* **D** *indicate a better fit.*

Institution	Year	p	Log-Normal LR	Log-Normal p	Exponential LR	Exponential p	Stretched Exp. LR	Stretched Exp. p	Cutoff Power Law LR	Cutoff Power Law p
Mount Sinai Hospital	2009	0.00	-0.21	0.67	31.80	**0.01**	-0.19	0.82	-0.53	0.30
	2010	0.00	-0.00	0.99	47.31	**0.00**	0.46	0.60	-0.23	0.50
Einstein School of Medicine	2006	0.00	-6.22	**0.03**	378.82	**0.00**	-7.06	**0.03**	-8.31	**0.00**
	2007	0.00	-0.30	0.59	17.65	**0.01**	-0.35	0.61	-0.67	0.25
	2008	0.00	-1.03	0.37	1235.22	**0.00**	0.71	0.81	-2.85	**0.02**
	2009	0.00	-2.48	0.13	578.27	**0.00**	-2.75	0.22	-5.82	**0.00**
	2010	0.00	-1.52	0.22	842.87	**0.00**	-0.64	0.80	-5.19	**0.00**
Univeristy of Vermont	1974	0.00	-0.39	0.54	20.93	**0.00**	-0.49	0.54	-1.17	0.13
	1980	0.00	-0.72	0.41	82.27	**0.00**	-0.81	0.47	-1.82	**0.06**
	1990	0.00	-0.94	0.36	23.05	**0.01**	-1.11	0.34	-1.79	**0.06**
	2000	0.00	-0.65	0.45	30.59	**0.00**	-0.78	0.44	-1.52	**0.08**
	2010	0.00	-inf	nan	7.75	**0.02**	0.39	0.34	-0.00	0.94
United Way, Chittendon County	2004	0.00	-0.46	0.47	28.75	**0.00**	-0.53	0.55	-1.29	0.11
	2005	0.00	-0.08	0.77	54.69	**0.00**	0.36	0.74	-0.69	0.24
	2006	0.00	-0.12	0.71	68.71	**0.00**	0.44	0.71	-0.85	0.19
	2007	0.00	-0.61	0.43	48.21	**0.00**	-0.65	0.57	-1.64	**0.07**
	2008	0.00	-0.13	0.72	46.52	**0.00**	0.14	0.90	-0.71	0.23
	2009	0.00	-0.35	0.55	48.39	**0.00**	-0.28	0.80	-1.15	0.13
	2010	0.00	-0.32	0.58	35.25	**0.00**	-0.30	0.77	-0.90	0.18
ECHO Science Museum	2005	0.00	-2.47	0.25	31.43	**0.04**	-3.04	0.21	-3.56	**0.01**
	2006	0.00	-0.20	0.69	1.42	0.57	-0.28	0.68	-0.53	0.30
	2007	0.00	-inf	nan	4.56	**0.03**	0.20	0.35	0.00	1.00
	2008	0.00	-inf	nan	4.28	**0.00**	0.29	0.19	0.00	1.00
	2009	0.00	-0.87	0.47	31.48	**0.01**	-1.23	0.44	-2.51	**0.03**
Flynn Theater	2006	0.00	-0.52	0.46	272.93	**0.00**	0.32	0.87	-2.80	**0.02**
	2007	0.00	-0.06	0.80	4.53	0.14	-0.08	0.86	-0.26	0.47
	2008	0.00	-0.56	0.45	303.73	**0.00**	0.38	0.86	-3.35	**0.01**
	2009	0.00	-0.25	0.63	281.34	**0.00**	1.11	0.59	-2.16	**0.04**
	2010	0.00	-3.96	**0.07**	129.19	**0.00**	-4.61	**0.06**	-6.78	**0.00**

Table 4.2: The results of the Likelihood-Ratio and its associated **p***-value are reported for different distributions. Here, positive values lend support to the Power Law and negative values to the other stated distribution. The significance of the* **LR** *is* **p***, where low values of* **p** *indicate a trustworthy* **LR***. Values for which* $\mathbf{p} < 0.05$ *are bolded.*

4.2 Shadow networks: Discovering hidden nodes with models of information flow

Paper number two is *Shadow networks: Discovering hidden nodes with models of information flow* by James P. Bagrow, Suma Desu, Morgan R. Frank, Narine Manukyan, Lewis Mitchell, Andrew Reagan, Eric E. Bloedorn, Lashon B. Booker, Luther K. Branting, Michael J. Smith, Brian F. Tivnan, Christopher M. Danforth, Peter S. Dodds, and Joshua C. Bongard, cited as Bagrow et al. (2014).

4.2.1 Abstract

Complex, dynamic networks underlie many systems, and understanding these networks is the concern of a great span of important scientific and engineering problems. Quantitative description is crucial for this understanding yet, due to a range of measurement problems, many real network datasets are incomplete. Here we explore how accidentally missing or deliberately hidden nodes may be detected in networks by the effect of their absence on predictions of the speed with which information flows through the network. We use Symbolic Regression (SR) to learn models relating information flow to network topology. These models show localized, systematic, and non-random discrepancies when applied to test networks with intentionally masked nodes, demonstrating the ability to detect the presence of missing nodes and where in the network those nodes are likely to reside.

4.2.2 Contribution

This paper is the result of a multi-day intensive collaboration called a Flash Mob Research Event. The format is one or two days of everyone in the same room, brain storming how to tackle an important open question. An outline of the paper is written, and after the event each member works to complete their part in carrying out the research idea. My responsibility was to build reciprocal reply networks from Twitter data, in an effort to measure information flow over the network. The network construction proceeded in three steps: (1) build a network using replies, (2) measure information flow over this reciprocal reply network, and (3) collect edges in the network for the actual information flow. Each step of the construction would be carried out over a number of days, and using a single note on the VACC, we were able to build networks in memory for a total of 9 days. These 9 days were considered for combinations 3/3/3 or 4/4/1 days, respectively. These data were used in a real world test, to accompany testing of simulated data.

4.3 Human language reveals a universal positivity bias

Paper number three is *Human language reveals a universal positivity bias* by Peter Sheridan Dodds, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew T. McMahon, Brian F. Tivnan, and Christopher M. Danforth, cited as Dodds et al. (2015a).

4.3.1 Abstract

Using human evaluation of 100,000 words spread across 24 corpora in 10 languages diverse in origin and culture, we present evidence of a deep imprint of human sociality in language, observing that (1) the words of natural human language possess a universal positivity bias; (2) the estimated emotional content of words is consistent between languages under translation; and (3) this positivity bias is strongly independent of frequency of word usage. Alongside these general regularities, we describe inter-language variations in the emotional spectrum of languages which allow us to rank corpora. We also show how our word evaluations can be used to construct physical-like instruments for both real-time and offline measurement of the emotional content of large-scale texts.

4.3.2 Contribution

In this paper I built the online appendices and performed additional tests of our method for building the sentiment timeseries for books (measuring their emotional arcs). This included building a fully interactive version of an application of this dataset to analyze the emotional arcs of stories, which was done for a selection of the Western Canon and Project Gutenberg books. In particular, we analyzed the emotional arc for these books in their original language, providing translations of the word shifts graphs into English. The translations relied upon the translations of Google Translate, as curated by Eric Clark. The additional statistical tests amounted to randomly shuffling the words in each book which we showcased, to demonstrate that the emotional arcs were meaningful.

4.4 Climate change sentiment on Twitter: An unsolicited public opinion poll

Paper number four is *Climate change sentiment on Twitter: An unsolicited public opinion poll* by Emily M. Cody, Andrew J. Reagan, Lewis Mitchell, Peter Sheridan Dodds, and Christopher M. Danforth, cited as Cody et al. (2015).

4.4.1 Abstract

The consequences of anthropogenic climate change are extensively debated through scientific papers, newspaper articles, and blogs. Newspaper articles may lack accuracy, while the severity of findings in scientific papers may be too opaque for the public to understand. Social media, however, is a forum where individuals of diverse backgrounds can share their thoughts and opinions. As consumption shifts from old media to new, Twitter has become a valuable resource for analyzing current events and headline news. In this research, we analyze tweets containing the word "climate" collected between September 2008 and July 2014. Through use of a previously developed sentiment measurement tool called the Hedonometer, we determine how collective sentiment varies in response to climate change news, events, and natural disasters. We find that natural disasters, climate bills, and oil-drilling can contribute to a decrease in happiness while climate rallies, a book release, and a green ideas contest can contribute to an increase in happiness. Words uncovered by our analysis suggest that responses to climate change news are predominantly from climate change activists rather than climate change deniers, indicating that Twitter is a valuable resource for the spread of climate change awareness.

4.4.2 Contribution

In this paper I was responsible for the data curation. This amounted to searching the Twitter database on the VACC for a variety of keywords, storing those results, and processing them into useful formats for analysis. Weighing at approximately 37TB of compressed JSON files, the Twitter database is difficult to search quickly over the GPFS architecture of the VACC, and only possible through the use of many short runtime (less than 2 hour) jobs. Given all of this, a single search of the database takes approximately 2 days if everything is running smoothly.

4.5 Reply to Garcia et al.: Common mistakes in measuring frequency dependent word characteristics

The fifth paper is *Reply to Garcia et al.: Common mistakes in measuring frequency dependent word characteristics* by P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth, cited as Dodds et al. (2015b).

4.5.1 Abstract

We demonstrate that the concerns expressed by Garcia et al. are misplaced, due to (1) a misreading of our findings in Dodds et al. (2015a); (2) a widespread failure to examine and present words in support of asserted summary quantities based on word usage frequencies; and (3) a range of misconceptions about word usage frequency, word rank, and expert-constructed word lists. In particular, we show that the English component of our study compares well statistically with two related surveys, that no survey design influence is apparent, and that estimates of measurement error do not explain the positivity biases reported in our work and that of others. We further demonstrate that for the frequency dependence of positivity —of which we explored the nuances in great detail in Dodds et al. (2015a) —Garcia et al did not perform a reanalysis of our data— they instead

carried out an analysis of a statistically improper data set and introduced a nonlinearity before performing linear regression.

4.5.2 Contribution

For this paper I built a new online appendix, performed tests of the claims made by Garcia *et al.*(including re-making their visualizations), and built visualizations for the extended version of the reply (e.g. Table I and Figure 1 in the arXiv version). Below, we include a reprint of the aforementioned Figure 1 and reproduction of the Figure from Garcia *et al.*:

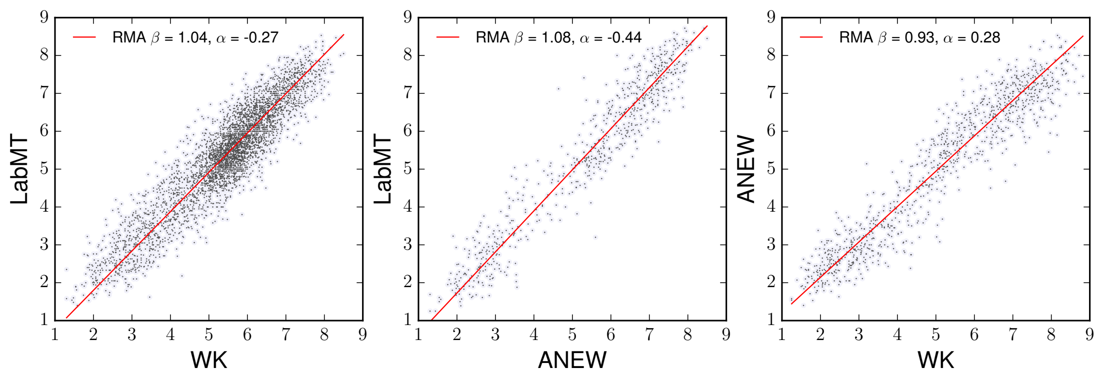


Figure 4.3: Reprint of Figure 1 from Dodds et al. (2015b), with the caption as follows: "Comparison of word ratings for three studies for overlapping words: labMT (Dodds et al., 2011), ANEW (Bradley and Lang, 1999), and Warriner and Kuperman (Warriner et al., 2013) Reduced major axis regression (Rayner, 1985) yield the fits $h'_{\text{avg}} = \beta h_{\text{avg}} + \alpha$."

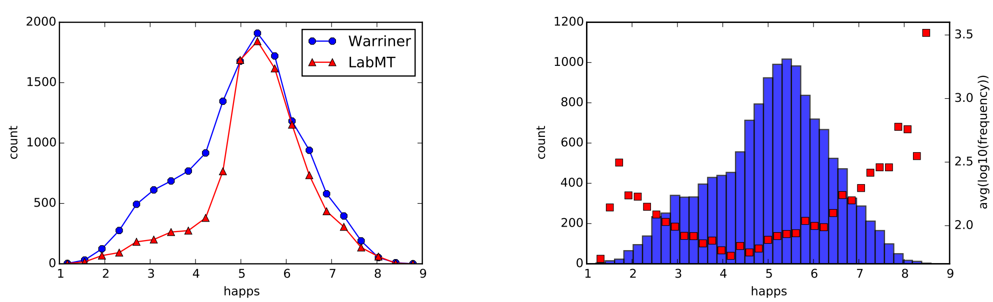


Figure 4.4: A reproduction of the Figure 1A and 1B from Garcia et al. (2015).

4.6 The game story space of professional sports: Australian Rules Football

Paper number six is *The game story space of professional sports: Australian Rules Football* by D. P. Kiley, A. J. Reagan, L. Mitchell, C. M. Danforth, and P. S. Dodds, cited as Kiley et al. (2016).

4.6.1 Abstract

Sports are spontaneous generators of stories. Through skill and chance, the script of each game is dynamically written in real time by players acting out possible trajectories allowed by a sport's rules. By properly characterizing a given sport's ecology of 'game stories', we are able to capture the sport's capacity for unfolding interesting narratives, in part by contrasting them with random walks. Here, we explore the game story space afforded by a data set of 1,310 Australian Football League (AFL) score lines. We find that AFL games exhibit a continuous spectrum of stories rather than distinct clusters. We show how coarse-graining reveals identifiable motifs ranging from last minute comeback wins to one-sided blowouts. Through an extensive comparison with biased random walks, we show that real AFL games deliver a broader array of motifs than null models, and we provide consequent insights into the narrative appeal of real games.

4.6.2 Contribution

For this paper I consulted with lead author Dilan Kiley on the statistical methods used, and assisted in performing the statistical analysis by leveraging the computational resources of the VACC.

4.7 The Lexicocalorimeter: Gauging public health through caloric input and output on social media

Paper number seven is *The Lexicocalorimeter: Gauging public health through caloric input and output on social media* by S. E. Alajajian, J. R. Williams, A. J. Reagan, S. C. Alajajian, M. R. Frank, L. Mitchell, J. Lahne, C. M. Danforth, and P. S. Dodds, cited as Alajajian et al. (2016).

4.7.1 Abstract

We propose and develop a Lexicocalorimeter: an online, interactive instrument for measuring the "caloric content" of social media and other large-scale texts. We do so by constructing extensive yet improvable tables of food and activity related phrases, and respectively assigning them with sourced estimates of caloric intake and expenditure. We show that for Twitter, our naive measures of "caloric input", "caloric output", and the ratio of these measures are all strong correlates with health and well-being measures for the contiguous United States. Our caloric balance measure in many cases outperforms both its constituent quantities, is tunable to specific health and well-being measures such as diabetes rates, has the capability of providing a real-time signal reflecting a population's health, and has the potential to be used alongside traditional survey data in the development of public policy and collective self-awareness. Because our Lexicocalorimeter is a linear superposition of principled phrase scores, we also show we can move beyond correlations to explore what people talk about in collective detail, and assist in the understanding and explanation of how population-scale conditions vary, a capacity unavailable to black-box type methods.

4.7.2 Contribution

For this paper I built an extensive online appendix and the accompanying website. The online appendix at http://www.uvm.edu/storylab/share/papers/alajajian2015a/ features an interactive dashboard provided at http://panometer.org. In addition to this tool, we provide searchable maps for all food and activity words used in the study. Next, we show snapshots of the various visualizations available on the website, in Figures 4.5–4.8.

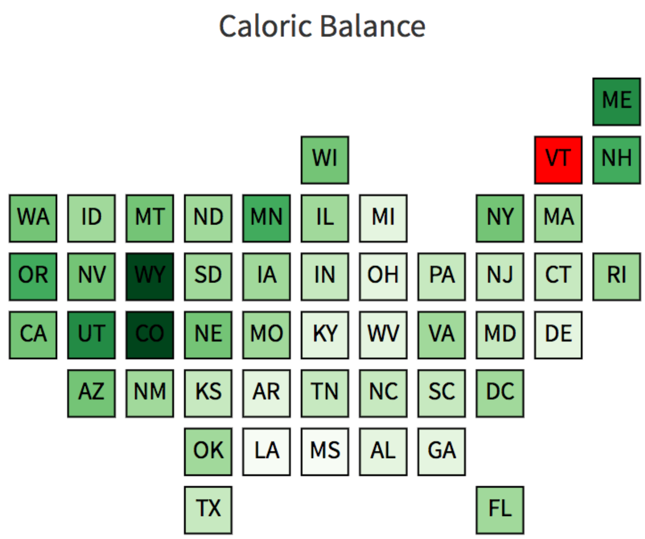


Figure 4.5: Lexicocalorimeter map, using square states to control for the disproportionate area and population of US States. Here, Vermont is highlighted by a hover.

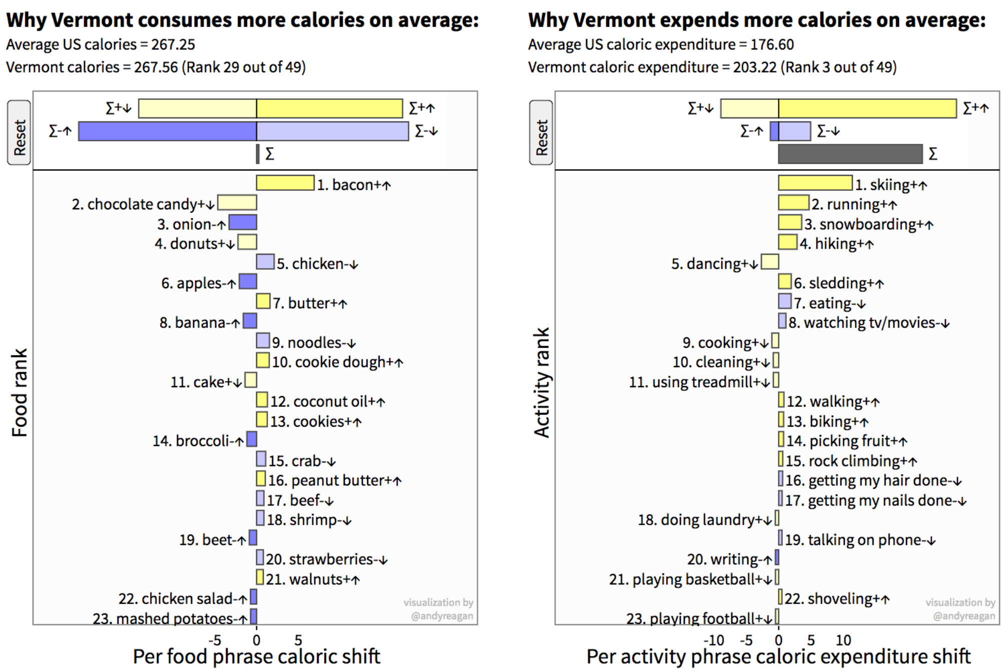


Figure 4.6: Lexicocalorimeter food and activity shifts. Here we see which foods and which activities contribute to Vermont's difference in caloric intake and expenditure from the US as a whole. We see that Bacon contributes most to caloric intake in Vermont relative to the average US intake, and overall Vermont is a middle-of-the-pack state (29th out of 49). On the right, Tweets from Vermont expend more calories than the US average with activities such as skiing, running, snowboarding, hiking, and sledding, giving the outdoorsy Vermont Twitter population the 3rd highest expenditure.

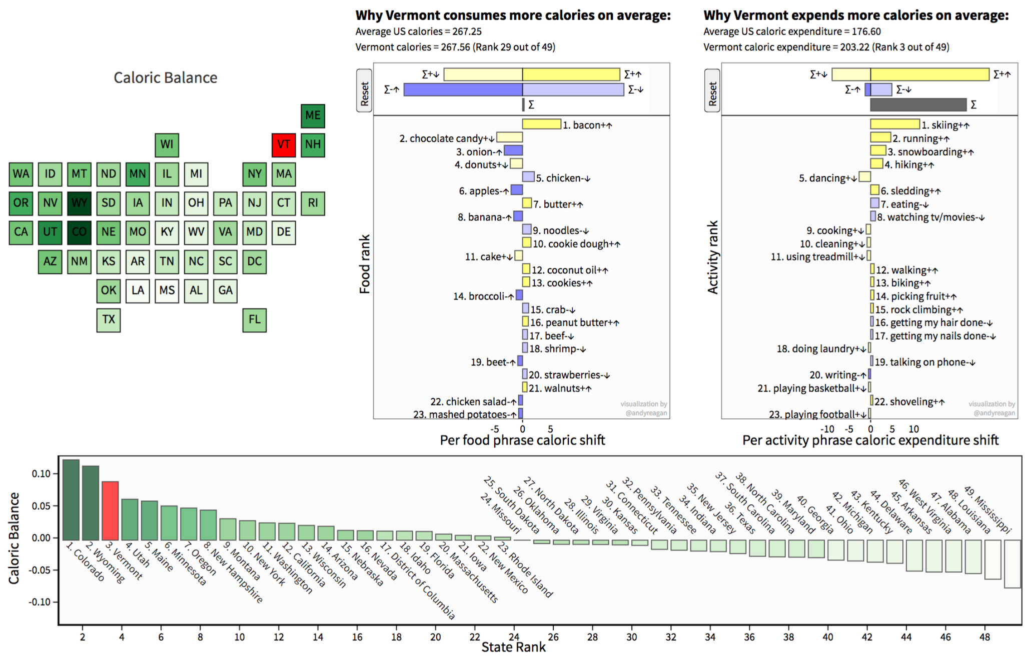


Figure 4.7: Overview of the Lexicocalorimeter dashboard. Each view is linked by hovering, and we can explore details of the caloric difference balances between states.

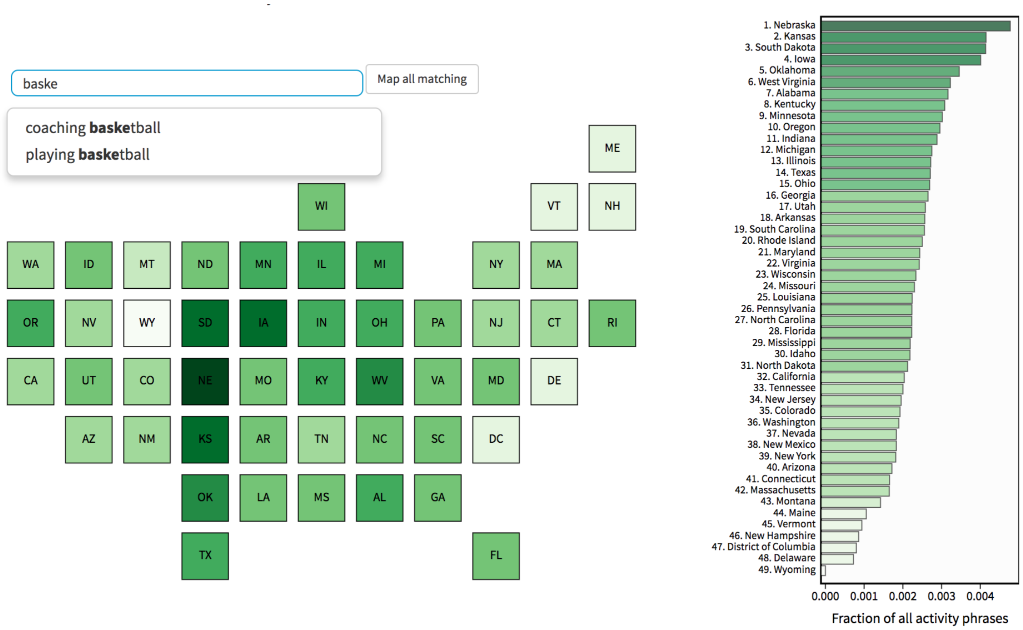


Figure 4.8: Snapshot of the Lexicocalorimeter activity search page. A similar page exists for foods. Here, we submit the query for "basketball", seeing that Nebraskans Tweet more about basketball relative to other activities than other US States.

4.8 Tracking the Teletherms: The spatiotemporal dynamics of the hottest and coldest days of the year

Paper number eight is *Tracking the Teletherms: The spatiotemporal dynamics of the hottest and coldest days of the year* by Peter Sheridan Dodds, Lewis Mitchell, Andrew J. Reagan, and Christopher M. Danforth, cited as Dodds et al. (2016).

4.8.1 Abstract

Instabilities and long term shifts in seasons, whether induced by natural drivers or human activities, pose great disruptive threats to ecological, agricultural, and social systems. Here, we propose, measure, and explore two fundamental markers of location-sensitive seasonal variations: the Summer and Winter Teletherms — the on-average annual dates of the hottest and coldest days of the year. We analyze daily temperature extremes recorded at 1218 stations across the contiguous United States from 1853–2012, and observe large regional variation with the Summer Teletherm falling up to 90 days after the Summer Solstice, and 50 days for the Winter Teletherm after the Winter Solstice. We show that Teletherm temporal dynamics are substantive with clear and in some cases dramatic shifts reflective of system bifurcations. We also compare recorded daily temperature extremes with output from two regional climate models finding considerable though relatively unbiased error. Our work demonstrates that Teletherms are an intuitive, powerful, and statistically sound measure of local climate change, and that they pose detailed, stringent challenges for future theoretical and computational models.

4.8.2 Contribution

For this paper, I built the online appendices and transformed the visualizations into online, interactive versions at http://teletherm.org/ using D3 Javascript (Bostock et al., 2011). The online appendices are available at http://compstorylab.org/share/papers/dodds2015c/index.html. Maps of the United States are shown in Figure 4.9, with Voronoi cells for each station colored in addition

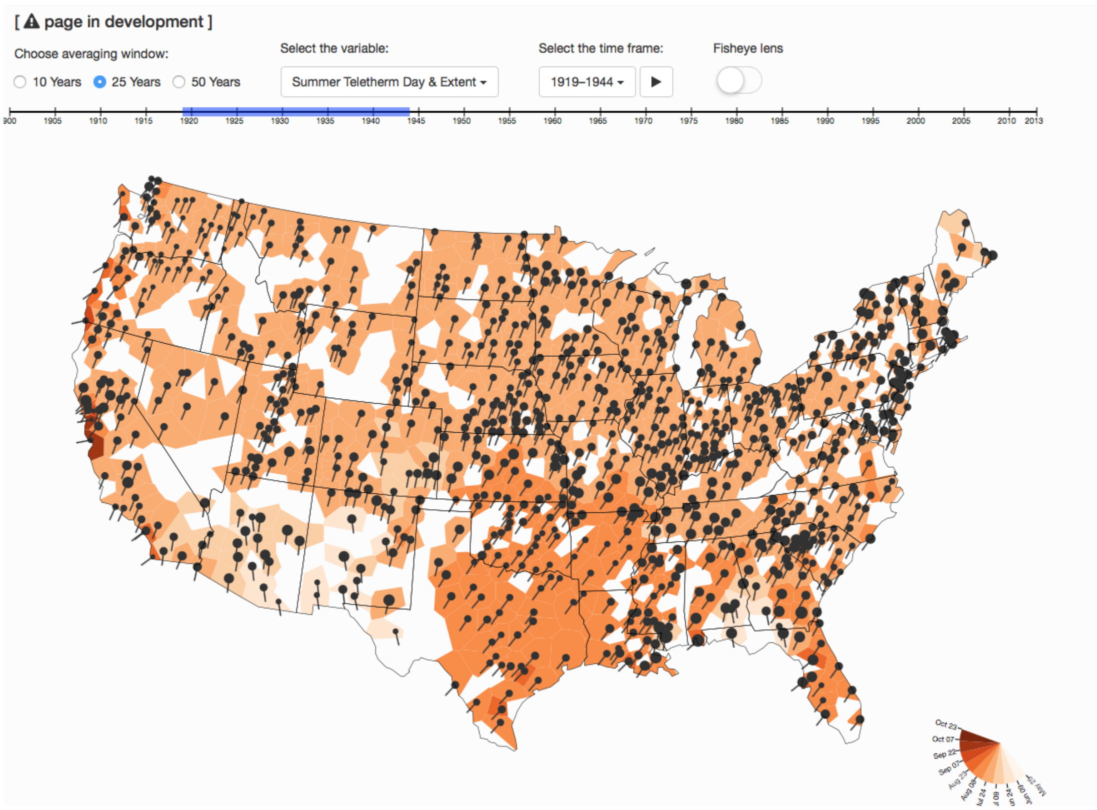


Figure 4.9: Interactive teletherm map with time and variable controls. Select between the teletherm day & extent and teletherm temperature, the averaging window to compute the teletherms, and the time to show on the map. A linear color scale, "oranges", is shown for teletherm day and extent. A diverging color scale is shown for temperatures, inspired by https://darksky.net. For each weather station, a tooltip hover shows details on demand.

to the direction and color of the arrows used in the static maps. Other features of these online maps include the ability to animate through time, select a fisheye lens for inspecting the map, and toggle between the various indicators (Summer/Winter Teletherm day and temperature).

To realize the goals of this research, the website is designed to communicate the patterns of Teletherm dynamics at both a local and a regional level. In addition to building interactive versions of the US maps, I worked with Professor Dodds to design novel visualizations for the individual station teletherm dynamics. These plots are shown in Figure 4.10, and accompany visualizations of the time dynamics of Teletherm days, extends, and temperatures. The online source code repository is publicly available at https://github.com/andyreagan/teletherm.org.

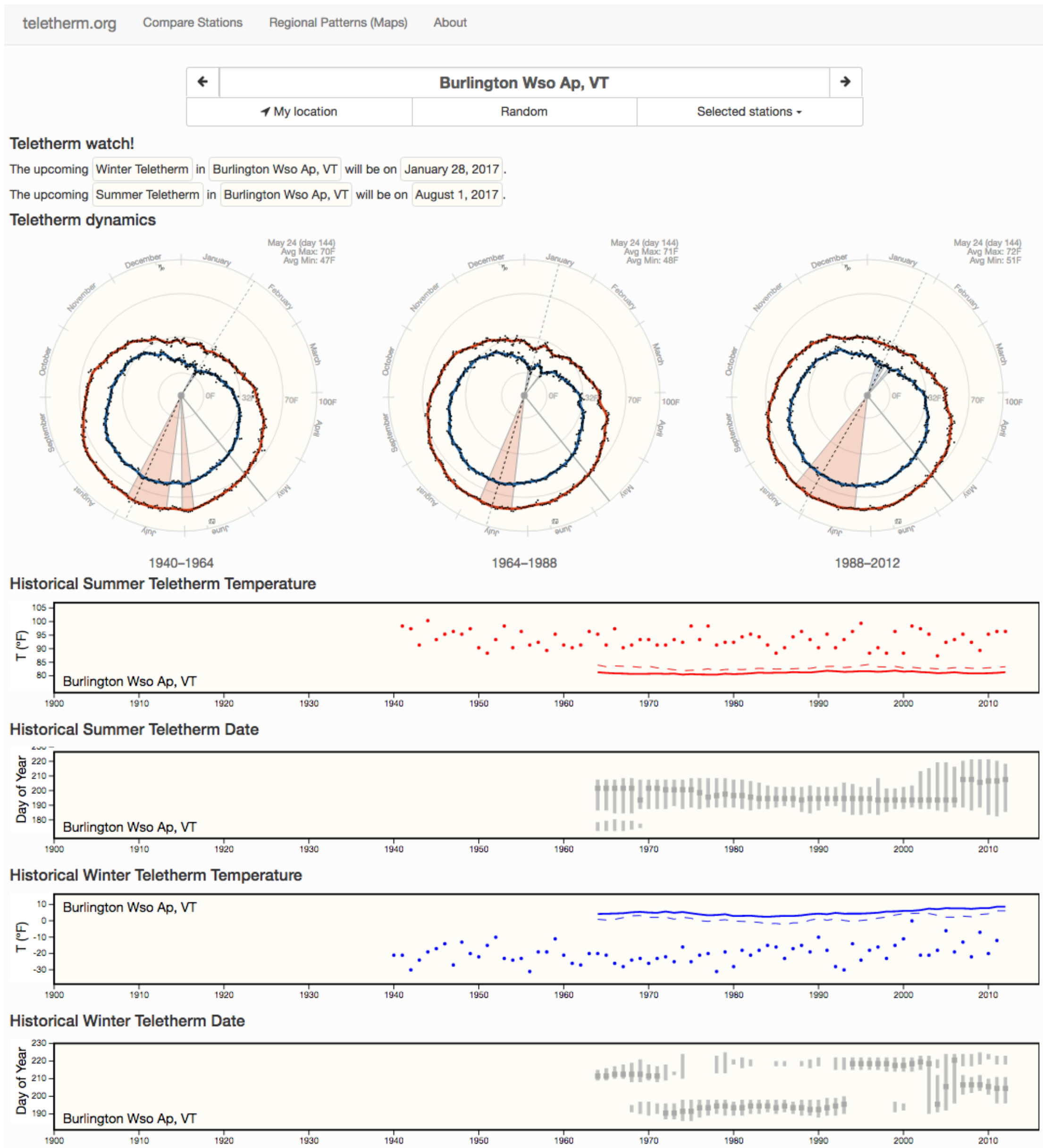


Figure 4.10: Teletherm dials shows the yearly temperature dynamics for a single location over a period of time, and time series below show the trends for both temperature extremes and teletherm dates. The min and max temperature for each day of the year are smoothed over three 25 year windows, one for each dial, and show in blue and red, respectively. As in the paper, the smoothed temperature is computed with a Gaussian kernel smoothing over the average min/max over days of the year. To avoid issues with the boundary, to compute the Gaussian kernel the temperature is wrapped on both ends of the year (with the same data). Summer and winter solstice are shown with icons, and the details of the day of year are shown in the upper right of each dial (over which the hover is linked between each dial—they all move together).

4.9 Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter

Paper number 10 is *Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter* by Ryan J. Gallagher, Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds, cited as Gallagher et al. (2016).

4.9.1 Abstract

Since the shooting of Black teenager Michael Brown by White police officer Darren Wilson in Ferguson, Missouri, the protest hashtag #BlackLivesMatter has amplified critiques of extrajudicial killings of Black Americans. In response to #BlackLivesMatter, other Twitter users have adopted #AllLivesMatter, a counter-protest hashtag whose content argues that equal attention should be given to all lives regardless of race. Through a multi-level analysis, we study how these protests and counter-protests diverge by quantifying aspects of their discourse. In particular, we introduce methodology that not only quantifies these divergences, but also reveals whether they are from widespread discussion or a few popular retweets within these groups. We find that #BlackLivesMatter exhibits many information rich conversations, while those within #AllLivesMatter are more muted and susceptible to hijacking. We also show that the discussion within #BlackLivesMatter is more likely to center around the deaths of Black Americans, while that of #AllLivesMatter is more likely to sympathize with the lives of police officers and express politically conservative views.

4.9.2 Contribution

My main contribution to this paper was working closely with lead author Ryan Gallagher to collect the data from our Twitter database on the VACC. We collected data for a number of hashtags, specifically all of the following:

```
keywords = [{"re": re.compile(r"#blacklivesmatter\b",flags=re.IGNORECASE)}},
```

```
{"re": re.compile(r"#alllivesmatter\b",flags=re.IGNORECASE)}},
{"re": re.compile(r"#bluelivesmatter\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#policelivesmatter\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#michaelbrown\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#ferguson\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#freddiegray\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#ericgarner\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#icantbreathe\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#sarahbland\b",flags=re.IGNORECASE)},
{"re": re.compile(r"#templeton\b",flags=re.IGNORECASE)},]
```

After collect the Tweets for these hashtags, they were reorganized by user, and then collected into a `sqlite` database using Django, a Python web framework. This web framework was then used to go back and collect the most recent 3,200 Tweets from each public Twitter account that we had found in our initial search. The collection ended on Nov 25th, 2015, so these Tweets were the 3,200 most recent as of that date. From this data, we were able to construct the social networks for analysis of the dynamics of these online communities.

Chapter 5

Conclusion

5.1 Future directions

First we take a look to the future research around sentiment analysis, emotional arcs, and the related projects we covered in Chapter 4.

5.1.1 Sentiment analysis

Our work looked in detail at dictionary-based sentiment analysis methodology, focusing on the use of these methods in qualitative and quantitative analysis. Immediate directions for the extension of dictionary based methods can examine the creation and use of dictionaries that offer (1) many emotions (Section 1.2.1), (2) MWEs (Section 1.2.4), (3) multiple word senses (Section 1.2.4), and (4) corpus-specific tuning. We reviewed automated methods to build corpus-specific dictionaries in Section 1.2.5, and while most approaches are low precision, we identified directions for that provide the highest precision and recall. Combining automated (machine learning, propagation-based) approaches with MWEs, word senses, and many emotions will provide many opportunities for the study of the sentiment properties of language and the improvement of sentiment analysis.

In addition to the improvement of the dictionaries, many unanswered questions remain around the visualization of sentiment analysis measures. We reviewed some approaches in Section 1.2.6 and reiterate that future work can (1) incorporate task-specific usability testing (Munzner, 2014), (2) visualize non-linear features (Ribeiro et al., 2016), and (3) continue to build more tools that enable other researchers to make use of visualization.

5.1.2 Emotional arcs

Here we enumerate some directions for research on emotional arcs in addition those mentioned at the end of Chapter 3 (see Section 3.4).

The emotional arcs of movies could be considered as a feature driving once controversial movies towards normalization over time, a closer examination of the trend presented by Amendola et al. (2015). Various studies have examined the changes in the valence of language over time, and in a similar fashion this will be possible to see how the emotional trajectories of stories has changed.

The emotional arc of a book can be used to predict the Library of Congress classification, using fiction and non-fiction separately to demonstrate the applicability of emotional arcs. In particular, one could feed the coefficient vector from the SVD projection for the first n modes into a predictor and see how much predictive power is contained in each mode, and exploring n can provide additional testing of how explanatory the first 6 modes are. Clustering on the emotional arc embedding vector would show whether these groups can be separated in a purely unsupervised manner.

Extending the approach of Bamman (2015) and the validation shown in Figure 1.8, it will remain important to keep people in the loop of the analysis of emotional arcs, since it is our reaction to stories that is being measured. A follow-up project to our work on emotional arcs could build a more complete user study to examine the human aspect of emotion in narrative more directly.

We broadly examined the the forefront of NLP research (Section 1.2.4), and can use the advancing methods to answer such questions as “is a character good or bad?”. The analysis of character networks (Section 1.3.3) will continue to improve with identification of the nature of relationships, and the events for particular characters (e.g., birth, marriage, death, and the associated sentiments).

Connecting the scripts, frames, and SIG-like approaches (see Section 1.3.1 and Section 1.3.4) to narrative more directly to the emotional arcs will be provide a finer-grained emotional arc representation, connected to the events in a narrative. This approach will in-part realize the jump from a bag-of-words to a bag-of-stories approach to natural language. As neural network approaches pust the state-of-the-art in NLP, there may be utility to considering architectures that have an explicit representations of abstraction levels. This approach is analagous to the Convolution Neural Network (CNN) architecture that has proven successful in image recognition tasks. An example structure to build upon is the Historical Thesaurus of English (Kay et al., 2009), as is done by Alexander et al.

(2015). In contrast to this proposed approach, the "automatic" feature selection (magic) of neural networks remains powerful (Radford et al., 2017).

5.1.3 Other projects

We have shown that it is possible to build population scale measures of well-being and public health. The Hedonometer and the Lexicocalorimeter can be utilized as only two of many broad measures that extend our dashboard of societal indicators; such additional "meters" of general interest that the Computational StoryLab has considered include such tools as an "insomniometer". Considering the Lexicocalorimeter, taking these lexical meters from snapshot-in-time analysis to real-time feeds remains a difficult challenge that has been accomplished with http://hedonometer.org/ and can be extended to additional meters.

There are many improvements possible for the visualizations hosted online at http://teletherm.org/. The teletherm animations can be improved through the use of the `d3.timer` module for smoother animation. Voronoi cells on the map are clipped at the boundary of the contiguous United States using a clipping mask that contains all 50 states as individual paths, and this does not work reliably in Google Chrome. More issues for improvement are noted in the "issues" tab of the online source code repository at https://github.com/andyreagan/teletherm.org. In addition, it will be possible to extend the teletherm project to incorporate temperature data from across the world.

5.2 Parting thoughts

Narratives are not unique in their explanation of causal links between events, and often the "adjacent narratives" are in direct competition. We saw in Section 1.3.4 that the the disambiguation of competing event chains is an active area of NLP research. This is identified as one factor contributing to information overload on the Internet (Orman, 2015), and participating in a collective cognitive denial of service attack (King et al., 2016). We are biased to seeing the world through narratives that have the most support from our existing experiences. Embodied in the principle of Occam's

Razor, we often prefer stories that are the simplest. This premise is explored anecdotally (Storr, 2014), and the competition between competing narratives is a new avenue for computation study.

The use of narratives in science belies an understanding of natural phenomena through metaphor, the consequences of stories in science has been examined by Mahoney and Goertz (2006); Levy (2008); Collier (2011); Gelman and Basbøll (2014). Narrative itself has been in the spotlight, being put forth to frame the decisions of economists in times of crisis and related to the political functions of democratic elections (Shriller, 2017).

Every-day causality and personal narrative build upon a fundamental assumption of personal agency and free will. Post-hoc rationalization is only useful to explain behavior that was intentional. Deterministic laws of physics are at odds with this worldview, but the science of complex systems has shown us that systems at different levels can exhibit emergent behavior that cannot be predicted from lower level interactions (Anderson, 1972). Applying computational thinking to the human concepts of metaphor and narrative can force us to further elucidate these distinctions and provide us with a deeper understanding of the world around us as we see it.

Bibliography

Abbott, H. P. (2008). *The Cambridge introduction to narrative.* Massachusetts: Cambridge University Press.

Abney, S. (1997). Part-of-speech tagging and partial parsing. In *Corpus-based methods in language and speech processing*, pp. 118–136. Springer.

Agarwal, A., B. Xie, I. Vovsha, O. Rambow, and R. Passonneau (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pp. 30–38. Association for Computational Linguistics.

Alajajian, S. E., J. R. Williams, A. J. Reagan, S. C. Alajajian, M. R. Frank, L. Mitchell, J. Lahne, C. M. Danforth, and P. S. Dodds (2016). The Lexicocalorimeter: Gauging public health through caloric input and output on social media. Available at http://arxiv.org/abs/1507.05098.

Alexander, M., F. Dallachy, S. Piao, A. Baron, and P. Rayson (2015). Metaphor, popular science, and semantic tagging: Distant reading with the historical thesaurus of english. *Digital Scholarship in the Humanities 30*(suppl 1), i16–i27.

Amendola, L., V. Marra, and M. Quartin (2015, Jul). The evolving perception of controversial movies. *Palgrave Communication* (1), 15038.

Amir, S., R. Astudillo, W. Ling, P. C. Carvalho, and M. J. Silva (2016). Expanding subjective lexicons for social media mining with embedding subspaces. *arXiv preprint arXiv:1701.00145*.

Anderson, P. W. (1972). More is different. *Science 177*(4047), 393–396.

Andor, D., C. Alberti, D. Weiss, A. Severyn, A. Presta, K. Ganchev, S. Petrov, and M. Collins (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Awad, H. (2013). *Culturally based story understanding.* Ph. D. thesis, Citeseer.

Baccianella, S., A. Esuli, and F. Sebastiani (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, Volume 10, pp. 2200–2204.

Bagrow, J., S. Desu, M. R. Frank, N. Manukyan, L. Mitchell, A. Reagan, E. Bloedorn, L. K. Booker, L. B. Branting, M. J. Smith, B. F. Tivnan, C. M. Danforth, P. S. Dodds, and J. C. Bongard (2014). Shadow networks: Discovering hidden nodes with models of information flow. Preprint available at http://arxiv.org/abs/1312.6122.

Baker, C. F., C. J. Fillmore, and J. B. Lowe (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pp. 86–90. Association for Computational Linguistics.

Bamman, D. (2015, Apr). Validity. http://www.davidbamman.com/?p=52.

Bamman, D., B. O'Connor, and N. A. Smith (2014). Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 352.

Bamman, D., T. Underwood, and N. A. Smith (2014). A bayesian mixed effects model of literary character. In *ACL (1)*, pp. 370–379.

Bar-Haim, R., E. Dinur, R. Feldman, M. Fresko, and G. Goldstein (2011). Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1310–1319. Association for Computational Linguistics.

Bestgen, Y. et al. (2008). Building affective lexicons from specific corpora for automatic sentiment analysis. In *LREC*. Citeseer.

Bestgen, Y. and N. Vincze (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior research methods 44*(4), 998–1006.

Bex, F. (2013). Values as the point of a story.

Bex, F. J. and T. J. Bench-Capon (2010). Persuasive stories for multi-agent argumentation. In *AAAI Fall Symposium: Computational Models of Narrative*, Volume 10, pp. 04.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72. Association for Computational Linguistics.

Blair-Goldensohn, S., K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar (2008). Building a sentiment summarizer for local service reviews. In *WWW workshop on NLP in the information explosion era*, Volume 14, pp. 339–348.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*(1), 1–8.

Booker, C. (2006). *The Seven Basic Plots: Why We Tell Stories*. New York: Bloomsbury Academic.

Bost, X., V. Labatut, and G. Linarès (2016). Narrative smoothing: dynamic conversational network for the analysis of tv series plots.

Bostock, M., V. Ogievetsky, and J. Heer (2011). D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*.

Bradley, M. M. and P. J. Lang (1999). Affective norms for english words (ANEW): Stimuli, instruction manual and affective ratings. Technical report c-1, University of Florida, Gainesville, FL.

Brewer, W. F. and E. H. Lichtenstein (1980). Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*.

Cambria, E., D. Olsher, and D. Rajagopal (2014). Senticnet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, pp. 1515–1521. AAAI Press.

Cambria, E. and B. White (2014). Jumping nlp curves: a review of natural language processing research [review article]. *IEEE Computational Intelligence Magazine 9*(2), 48–57.

Campbell, J. (1949). *The Hero with a Thousand Faces* (third ed.). California: New World Library.

Campbell, J. and B. Moyers (1991). *The Power of Myth*. Anchor.

Cao, N. and W. Cui (2016). Introduction to text visualization. *Atlantis briefs in artificial intelligence (1*.

Card, D., A. E. Boydstun, J. H. Gross, P. Resnik, and N. A. Smith (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Volume 2, pp. 438–444.

Cave, S. (2013, Jul). The 4 stories we tell ourselves about death. http://www.ted.com/talks/stephen_cave_the_4_stories_we_tell_ourselves_about_death.

Chambers, N. (2013). Event schema induction with a probabilistic entity-driven model. In *EMNLP*, Volume 13, pp. 1797–1807.

Chambers, N. and D. Jurafsky (2008). Unsupervised learning of narrative event chains. In *ACL*, Volume 94305, pp. 789–797. Citeseer.

Chambers, N. and D. Jurafsky (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 602–610. Association for Computational Linguistics.

Chambers, N. and D. Jurafsky (2010). A database of narrative schemas. In *LREC*.

Chambers, N., S. Wang, and D. Jurafsky (2007). Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 173–176. Association for Computational Linguistics.

Chen, D. and C. D. Manning (2014). A fast and accurate dependency parser using neural networks. In *EMNLP*, pp. 740–750.

Cherny, L. (2016, Jun). The bones of a bestseller: Visualizing fiction. http://blogger.ghostweather.com/2013/06/analysis-of-fiction-my-openvisconf-talk.html.

Cheung, J. C. K., H. Poon, and L. Vanderwende (2013). Probabilistic frame induction. *arXiv preprint arXiv:1302.4813*.

Chuang, J., C. D. Manning, and J. Heer (2012). Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 74–77. ACM.

Chuang, J., D. Ramage, C. Manning, and J. Heer (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 443–452. ACM.

Chung, S. and S. Liu (2011). Predicting stock market fluctuations from Twitter. *Berkeley, California*.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on Applied natural language processing*, pp. 136–143. Association for Computational Linguistics.

Clancy, E. (2015). A fabula of syuzhet: A contretemps of digital humanities and sentiment analysis. https://storify.com/clancynewyork/contretemps-a-syuzhet and https://storify.com/clancynewyork/a-fabula-of-syuzhet-ii.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law distributions in empirical data. *SIAM Review 51*, 661–703.

Cobley, P. (2005). Narratology. *The Johns Hopkins Guide to Literary Theory and Criticism, 2nd ed. John Hopkins University Press, London*.

Cody, E. M., A. J. Reagan, L. Mitchell, P. S. Dodds, and C. M. Danforth (2015). Climate change sentiment on twitter: An unsolicited public opinion poll. *PLOS ONE*.

Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics 44*(04), 823–830.

da Silva, S. G. and J. J. Tehrani (2016). Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *Royal Society Open Science 3*(1).

DARPA (2011, 12). Broad agency announcement: Narrative networks. available at `http://www.fedbizopps.gov/`, accessed June 20, 2016.

Das, S. R. and M. Y. Chen (2007). Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science 53*(9), 1375–1388.

De Smedt, T. and W. Daelemans (2012). Pattern for Python. *The Journal of Machine Learning Research 13*(1), 2063–2067.

Decadt, B., V. Hoste, W. Daelemans, and A. Van den Bosch (2004). Gambl, genetic algorithm optimization of memory-based wsd. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 108–112. Association for Computational Linguistics.

DeGraff, A. and D. Harmon (2015). *Plotted: A Literary Atlas*. Houghton Mifflin Harcourt.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational linguistics 14*(1), 31–39.

Do, Q. X., Y. S. Chan, and D. Roth (2011). Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA, pp. 294–303. Association for Computational Linguistics.

Dodds, P. S. (2013). Homo Narrativus and the trouble with fame. Nautilus Magazine. http://nautil.us/issue/5/fame/homo-narrativus-and-the-trouble-with-fame.

Dodds, P. S., E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth (2015a). Human language reveals a universal positivity bias. *PNAS 112*(8), 2389–2394.

Dodds, P. S., E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth (2015b). Reply to garcia et al.: Common mistakes in measuring frequency-dependent word characteristics. *Proceedings of the National Academy of Sciences 112*(23), E2984–E2985.

Dodds, P. S. and C. M. Danforth (2009, July). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies 11*(4), 441–456.

Dodds, P. S., K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth (2011, 12). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE 6*(12), e26752.

Dodds, P. S., L. Mitchell, A. J. Reagan, and C. M. Danforth (2016, may). Tracking climate change through the spatiotemporal dynamics of the teletherms, the statistically hottest and coldest days of the year. *PLOS ONE 11*(5), e0154184.

Dolby, S. K. (2008). *Literary Folkloristics and the Personal Narrative*. Indiana: Trickster Press.

Dundes, A. (1997). The motif-index and the tale type index: A critique. *Journal of Folklore Research*, 195–202.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion 6*(3-4), 169–200.

Elson, D. K. (2012a). Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative, Istanbul, Turkey*.

Elson, D. K. (2012b). *Modeling narrative discourse*. Ph. D. thesis, Citeseer.

Elson, D. K., N. Dames, and K. R. McKeown (2010). Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 138–147. Association for Computational Linguistics.

Emerson, J., N. Churcher, and A. Cockburn (2015). Tag clouds for software and information visualisation. In *Proceedings of the 14th Annual ACM SIGCHI_NZ conference on Computer-Human Interaction*, pp. 1. ACM.

Enderle, S. (2015, April). What's a sine wave of sentiment? http://www.lagado.name/blog/sine-of-the-times/.

Enderle, S. (2016, Sept). Brownian noise and plot arcs. http://www.lagado.name/blog/brownian-noise-and-plot-arcs/.

Esuli, A. and F. Sebastiani (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, Volume 6, pp. 417–422. Citeseer.

Feinberg, J. (2009). Wordle-beautiful word clouds.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finlayson, M. A. (2011). *Learning narrative structure from annotated folktales*. Ph. D. thesis, Massachusetts Institute of Technology.

Gabasova, E. (2015, Dec). The star wars social network. http://evelinag.com/blog/2015/12-15-star-wars-social-network/index.html.

Gallagher, R. J., A. J. Reagan, C. M. Danforth, and P. S. Dodds (2016). Divergent discourse between protests and counter-protests: #blacklivesmatter and #alllivesmatter. *CoRR abs/1606.06820*.

Gao, J., M. L. Jockers, J. Laudun, and T. Tangherlini (2016). A multiscale theory for the dynamical evolution of sentiment in novels. In *Behavioral, Economic and Socio-cultural Computing (BESC), 2016 International Conference on*, pp. 1–4. IEEE.

Garcia, D., A. Garas, and F. Schweitzer (2015). The language-dependent relationship between word happiness and frequency. *Proceedings of the National Academy of Sciences 112*(23), E2983.

Gelman, A. and T. Basbøll (2014). When do stories work? evidence and illustration in the social sciences. *Sociological Methods & Research*, 0049124114526377.

Giachanou, A. and F. Crestani (2016, June). Like it or not: A survey of twitter sentiment analysis methods. *ACM Comput. Surv. 49*(2), 28:1–28:41.

Gleick, J. (2011). *The Information: A History, A Theory, A Flood.* New York: Pantheon.

Golder, S. A. and M. W. Macy (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science Magazine 333*, 1878–1881.

Gonçalves, P., M. Araújo, F. Benevenuto, and M. Cha (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pp. 27–38. ACM.

Gottesman, W. G., A. J. Reagan, and P. S. Dodds (2014). Collective philanthropy: Describing and modeling the ecology of giving. *PLoS ONE 9*, e98876.

Gottschall, J. (2013). *The Storytelling Animal: How Stories Make Us Human.* New York, NY: Mariner Books.

Goyal, A., E. Riloff, et al. (2013). A computational model for plot units. *Computational Intelligence 29*(3), 466–488.

Halvey, M. J. and M. T. Keane (2007). An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pp. 1313–1314. ACM.

Hamann, S. (2012). Mapping discrete and dimensional emotions onto the brain: controversies and consensus. *Trends in cognitive sciences 16*(9), 458–466.

Hamilton, W. L., K. Clark, J. Leskovec, and D. Jurafsky (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *arXiv preprint arXiv:1606.02820*.

Hand, D. J. and K. Yu (2001). Idiot's bayes—not so stupid after all? *International statistical review 69*(3), 385–398.

Handler, A., M. J. Denny, H. Wallach, and B. O'Connor (2016). Bag of what? simple noun phrase extraction for text analysis. *NLP+ CSS 2016*, 114.

Harris, W. F. (1959). *The basic patterns of plot.* Oklahoma: University of Oklahoma Press.

Harrison, N. A., M. A. Gray, P. J. Gianaros, and H. D. Critchley (2010). The embodiment of emotional feelings in the brain. *Journal of Neuroscience 30*(38), 12878–12884.

Hatzivassiloglou, V. and K. R. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pp. 174–181. Association for Computational Linguistics.

Hearst, M. (2009). *Search user interfaces.* Cambridge University Press.

Hearst, M. A. and D. Rosner (2008). Tag clouds: Data analysis tool or social signaller? In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pp. 160–160. IEEE.

Heer, J. (2014). Text visualizatoin. CSE 512 Lecture available at http://courses.cs.washington.edu/courses/cse512/14wi/.

Honnibal, M. (2015, Aug). Displaying linguistic structure with css. https://spacy.io/blog/displacy-dependency-visualizer.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM.

Hutto, C. J. and E. Gilbert (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*. AAAI Publications.

Jack, R. E., O. G. Garrod, and P. G. Schyns (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology 24*(2), 187–192.

Jockers, M. (2014, Jun). A novel method for detecting plot. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/.

Jockers, M. (2015, Feb). The rest of the story. http://www.matthewjockers.net/2015/02/25/the-rest-of-the-story/.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Kaji, N. and M. Kitsuregawa (2007). Building lexicon for sentiment analysis from massive collection of HTML documents. In *EMNLP-CoNLL*, pp. 1075–1083.

Kay, C., J. Roberts, M. Samuels, and I. Wotherspoon (2009). *Historical Thesaurus of the Oxford English Dictionary*. Oxford University Press.

Kiley, D. P., A. J. Reagan, L. Mitchell, C. M. Danforth, and P. S. Dodds (2016, May). Game story space of professional sports: Australian rules football. *Phys. Rev. E 93*, 052314.

Kim, S.-M. and E. Hovy (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 1367. Association for Computational Linguistics.

King, G., J. Pan, and M. E. Roberts (2016). How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *Harvard University*.

Kiritchenko, S., X. Zhu, and S. M. Mohammad (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research 50*, 723–762.

Kirschenbaum, M. G. (2007). The remaking of reading: Data mining and the digital humanities. In *The National Science Foundation Symposium on Next Generation of Data Mining and Cyber-Enabled Discovery for Innovation, Maryland*.

Koch, A., H. Alves, T. Krüger, and C. Unkelbach (2016). A general valence asymmetry in similarity: Good is more alike than bad. *Journal of Experimental Psychology: Learning, Memory, and Cognition 42*(8), 1171.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE 78*(9), 1464–1480.

Kosinski, M., D. Stillwell, and T. Graepel (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences 110*(15), 5802–5805.

Kuster, D. (2015, Jul). Exploring the shapes of stories using python and sentiment apis. https://indico.io/blog/plotlines/.

Lee, B., N. H. Riche, A. K. Karlson, and S. Carpendale (2010). Sparkclouds: Visualizing trends in tag clouds. *IEEE transactions on visualization and computer graphics 16*(6), 1182–1189.

Lehnert, W. G. (1981). Plot units and narrative summarization. *Cognitive Science 5*(4), 293–331.

Levallois, C. (2013). Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, Volume 2, pp. 414–417.

Levy, J. S. (2008). Case studies: Types, designs, and logics of inference. *Conflict Management and Peace Science 25*(1), 1–18.

Li, B., S. Lee-Urban, D. S. Appling, and M. O. Riedl (2012). Crowdsourcing narrative intelligence. *Advances in Cognitive Systems 2*, 25–42.

Li, B., S. Lee-Urban, G. Johnston, and M. Riedl (2013). Story generation with crowdsourced plot graphs. In *AAAI*.

Lin, Y., J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pp. 169–174. Association for Computational Linguistics.

Lindquist, K., M. Gendron, A. Satpute, L. Barrett, M. Lewis, and J. Haviland-Jones (2016). Language and emotion: Putting words into feelings and feelings into words. *Handbook of emotions*.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing 2*, 627–666.

Liu, B. (2012, May). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies. San Rafael, CA: Morgan & Claypool Publishers.

Liu, Y., X. Huang, A. An, and X. Yu (2007). Arsa: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 607–614. ACM.

Lohmann, S., J. Ziegler, and L. Tetzlaff (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In *IFIP Conference on Human-Computer Interaction*, pp. 392–404. Springer.

Luo, Z., M. Osborne, and T. Wang (2012). Opinion retrieval in twitter. In *ICWSM*.

MacDonald, M. R. (1982). *Storytellers Sourcebook: A Subject, Title, and Motif Index to Folklore Collections for Children*. Michigan: Gale Group.

Mahoney, J. and G. Goertz (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political analysis 14*(3), 227–249.

Mandera, P., E. Keuleers, and M. Brysbaert (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology 68*(8), 1623–1642.

Mani, I. (2012). Computational modeling of narrative. *Synthesis Lectures on Human Language Technologies 5*(3), 1–142.

Mani, I., M. Verhagen, B. Wellner, C. M. Lee, and J. Pustejovsky (2006). Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 753–760. Association for Computational Linguistics.

Manning, C. D., M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky (2014). The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pp. 55–60.

Marcus, M. P., M. A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics 19*(2), 313–330.

McCloud, S. (2006). *Making comics: storytelling secrets of comics, manga and graphic novels*. New York: Harper.

McIntyre, N. and M. Lapata (2010). Plot induction and evolutionary search for story generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 1562–1572. Association for Computational Linguistics.

Meeks, E. and M. Averick. A data-driven exploration of archer. https://archervisualization.herokuapp.com/.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science 331*(6014), 176–182.

Mikolov, T. and J. Dean (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Min, S. and J. Park (2016). Narrative as a complex network: A study of Victor Hugo's les misérables. In *Proceedings of HCI Korea*.

Mitchell, L., M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth (2013, May). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE 8*(5), e64417.

Mohammad, S. M., S. Kiritchenko, and X. Zhu (2013, June). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

Moretti, F. (2000). Conjectures on world literature. *New Left Review 1*, 54.

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for a Literary History*. New York: Verso.

Moretti, F. (2013). *Distant Reading*. New York: Verso.

Mostafazadeh, N., N. Chambers, X. He, D. Parikh, D. Batra, L. Vanderwende, P. Kohli, and J. Allen (2016, June). A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp. 839–849w. Association for Computational Linguistics.

Munroe, R. (2009, 11). Movie narrative charts. http://xkcd.com/657/.

Munzner, T. (2014). *Visualization analysis and design.* CRC Press.

Nenkova, A. and K. McKeown (2012). A survey of text summarization techniques. In *Mining text data*, pp. 43–76. Berlin, Germany: Springer.

Neukom Institute, D. (2016). Turing tests in creative arts: Digilit 2016. https://math.dartmouth.edu/~turingtests/DigiLit.php.

Nickerson, R. S. (1998). Confirmation Bias; A ubiquitous phenomenon in many guises. *Review of General Psychology 2*, 175–220.

Nielsen, F. Å. (2011, May). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey (Eds.), *CEUR Workshop Proceedings*, Volume Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages 718, pp. 93–98.

O'Connor, B. (2013). Learning frames from text with an unsupervised latent variable model. *arXiv preprint arXiv:1307.7382*.

Ogawa, M. and K.-L. Ma (2010). Software evolution storylines. In *Proceedings of the 5th international symposium on Software visualization*, pp. 35–42. ACM.

Orman, L. V. (2015). Information paradox: Drowning in information, starving for knowledge.

Owoputi, O., B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pappas, N., G. Katsimpras, and E. Stamatatos (2013). Distinguishing the popularity between topics: A system for up-to-date opinion retrieval and mining in the web. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 197–209. Springer.

Pechenick, E. A., C. M. Danforth, and P. S. Dodds (2015). Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *arXiv preprint arXiv:1501.00960*.

Pennebaker, J. W., M. E. Francis, and R. J. Booth (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001.

Pichotta, K. and R. J. Mooney (2015). Learning statistical scripts with lstm recurrent neural networks. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Piper, A. (2015a). Novel devotions: Conversional reading, computational modeling, and the modern novel. *New Literary History 46*(1), 63–98.

Piper, A. (2015b, Mar). Validation and subjective computing. http://txtlab.org/?p=470.

Plutchik, R. (1991). *The emotions*. University Press of America.

Plutchik, R. (2001). The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist 89*(4), 344–350.

Polanyi, L. and A. Zaenen (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pp. 1–10. Springer.

Polti, G. (1921). *The Thirty-Six Dramatic Situations*. Ohio: James Knapp Reeve.

Poria, S., A. Gelbukh, A. Hussain, N. Howard, D. Das, and S. Bandyopadhyay (2013). Enhanced senticnet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems 28*(2), 31–38.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Prado, S. D., S. R. Dahmen, A. L. C. Bazzan, P. M. Carron, and R. Kenna (2016). Temporal network analysis of literary texts.

Pratchett, T., I. Stewart, and J. Cohen (2003). *The Science of Discworld II: The Globe*. London, UK: Ebury Press.

Propp, V. (1968). *Morphology of the Folktale. 1928*. Texas: Texas University Press.

Pustejovsky, J., P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, et al. (2003). The timebank corpus. In *Corpus linguistics*, Volume 2003, pp. 40.

Radford, A., R. Jozefowicz, and I. Sutskever (2017). Learning to generate reviews and discovering sentiment.

Raftery, B. (2011, Sep). How Dan Harmon drives himself crazy making Community. at `http://www.wired.com/2011/09/mf_harmon/`, accessed June 20, 2016.

Rao, D. and D. Ravichandran (2009). Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 675–682. Association for Computational Linguistics.

Rayner, J. M. V. (1985). Linear relations in biomechanics: the statistics of scaling functions. *J. Zool. Lond. (A) 206*, 415–439.

Reagan, A., B. Tivnan, J. R. Williams, C. M. Danforth, and P. S. Dodds (2015). Benchmarking sentiment analysis methods for large-scale texts: A case for using continuum-scored words and word shift graphs. Preprint available at https://arxiv.org/abs/1512.00531.

Regneri, M., A. Koller, and M. Pinkal (2010). Learning script knowledge with web experiments. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 979–988. Association for Computational Linguistics.

Reiter, N., A. Frank, and O. Hellwig (2014). An nlp-based cross-document approach to narrative structure discovery. *Literary and Linguistic Computing 29*(4), 583–605.

Ribeiro, F. N., M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto (2016, July). SentiBench — a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Sci. 5*(1), 23.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM.

Riedl, M. O. (2016). Computational narrative intelligence: A human-centered goal for artificial intelligence. *arXiv preprint arXiv:1602.06484*.

Riedl, M. O. and B. Harrison (2015). Using stories to teach human values to artificial agents.

Rivadeneira, A. W., D. M. Gruen, M. J. Muller, and D. R. Millen (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 995–998. ACM.

Robinson, D. L. (2008). Brain function, emotional experience and personality. *Netherlands Journal of Psychology 64*(4), 152–168.

Roemmele, M., S. Kobayashi, N. Inoue, and A. Gordon (2017, April). An rnn-based binary classifier for the story cloze test. Proceedings of Linking Models of Lexical, Sentential and Discourse-level Semantics, workshop at European Association for Computational Linguistics.

Rothe, S., S. Ebert, and H. Schütze (2016). Ultradense word embeddings by orthogonal transformation. *arXiv preprint arXiv:1602.07572*.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

Ruiz, E. J., V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes (2012). Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 513–522. ACM.

Rumelhart, D. E. (1975). Notes on a schema for stories. *Representation and understanding: Studies in cognitive science 211*(236), 45.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology 39*(6), 1161–1178.

Ruths, D. (2016, Mar). Why the force awakens is not just a remake of a new hope. http://www.derekruths.com/2016/03/05/why-the-force-awakens-is-not-just-a-remake-of-a-new-hope/.

Saif, H., M. Fernandez, Y. He, and H. Alani (2013). Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the sts-gold.

Sandhaus, E. (2008). The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

Schank, R. C. and R. P. Abelson (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures.* Psychology Press.

Schmidt, B. (2015a, Apr). Commodius vici of recirculation: The real problem with syuzhet. http://benschmidt.org/2015/04/03/commodius-vici-of-recirculation-the-real-problem-with-syuzhet/.

Schmidt, B. (2016, Jul). Plot arceology 2016: emotion and tension. http://sappingattention.blogspot.com/2016/07/plot-arceology-emotion-and-tension.html.

Schmidt, B. M. (2015b). Plot arceology: A vector-space model of narrative structure. In *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 1667–1672. IEEE.

Schrammel, J., M. Leitner, and M. Tscheligi (2009). Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2037–2040. ACM.

Schrauf, R. W. and J. Sanchez (2004). The preponderance of negative emotion words across generations and across cultures. *Journal of Multilingual and Multicultural Development 25*, 266–284.

Schulz, K. (2011, Jun). What is distant reading? http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

Shriller, R. J. (2017, Jan). Narrative economics. In C. F. for Research In Economics (Ed.), *129th annual meeting of the American Economic Association*, Number 2069.

Si, J., A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng (2013). Exploiting topic based Twitter sentiment for stock prediction. In *ACL (2)*, pp. 24–29.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 254–263. Association for Computational Linguistics.

Snyder, B. and M. Palmer (2004). The english all-words task. In *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 41–43. Association for Computational Linguistics.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, Volume 1631, pp. 1642. Citeseer.

Stone, P. J., D. C. Dunphy, and M. S. Smith (1966). The general inquirer: A computer approach to content analysis. *MIT Press*.

Storr, W. (2014). *The unpersuadables: Adventures with the enemies of science*. The Overlook Press.

Swafford, A. (2015, Mar). Problems with the syuzhet package. https://annieswafford.wordpress.com/2015/03/02/syuzhet/.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Taboada, M. and J. Grieve (2004). Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Re# port SS# 04# 07), Stanford University, CA, pp. 158q161. AAAI Press.*

Tang, D., F. Wei, B. Qin, M. Zhou, and T. Liu (2014). Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *COLING*, pp. 172–182.

Tenenbaum, D. J., K. Barrett, S. Medaris, and T. Devitt (2015, February). In 10 languages, happy words beat sad ones. http://whyfiles.org/2015/in-10-languages-happy-words-beat-sad-ones/.

Thelwall, M., K. Buckley, and G. Paltoglou (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology 63*(1), 163–173.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology 61*(12), 2544–2558.

Tobias, R. B. (1993). *20 Master Plots: And How to Build Them.* Ohio: Writer's Digest Books.

Toutanova, K., D. Klein, C. D. Manning, and Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173–180. Association for Computational Linguistics.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM 10*, 178–185.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417–424. Association for Computational Linguistics.

Turney, P. D. and M. L. Littman (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS) 21*(4), 315–346.

Uther, H.-J. (2011). *The Types of International Folktales. A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson. Part I. Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction (FF Communications, 284).* Helsinki, Finland: Finnish Academy of Science and Letters.

Valls-Vargas, J., S. Ontanón, and J. Zhu (2014). Toward automatic character identification in unannotated narrative text. In *Seventh Intelligent Narrative Technologies Workshop.*

Valls-Vargas, J., J. Zhu, and S. Ontañón (2014). Toward automatic role identification in unannotated folk tales. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference.*

Van Ham, F., M. Wattenberg, and F. B. Viégas (2009). Mapping text with phrase nets. *IEEE transactions on visualization and computer graphics 15*(6).

Van Rensbergen, B., S. De Deyne, and G. Storms (2016). Estimating affective word covariates using word association data. *Behavior Research Methods 48*(4), 1644–1652.

Various. Project Gutenberg.

Velikovich, L., S. Blair-Goldensohn, K. Hannan, and R. McDonald (2010). The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 777–785. Association for Computational Linguistics.

Viegas, F. B., M. Wattenberg, and J. Feinberg (2009). Participatory visualization with wordle. *IEEE transactions on visualization and computer graphics 15*(6).

Viegas, F. B., M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon (2007). Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics 13*(6).

Volger, C. (1992). The writer's journey. mythic structure for storytellers and screenwriters.

Vonnegut, K. (1981). *Palm Sunday.* New York: RosettaBooks LLC.

Vonnegut, K. (1995). Shapes of stories. https://www.youtube.com/watch?v=oP3c1h8v2ZQ.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

Warriner, A. B., V. Kuperman, and M. Brysbaert (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods 45*(4), 1191–1207.

Watson, D. and L. A. Clark (1999). *The PANAS-X: Manual for the positive and negative affect schedule-expanded form: Manual for the positive and negative affect schedule-expanded form.* Ph. D. thesis, University of Iowa.

Weingart, S. Not enough perspectives, pt. 1. http://scottbot.net/not-enough-perspectives-pt-1/.

Whissell, C., M. Fournier, R. Pelland, D. Weir, and K. Makarec (1986). A dictionary of affect in language: Iv. reliability, validity, and applications. *Perceptual and Motor Skills 62*(3), 875–888.

Williams, J. R. (2016). Boundary-based mwe segmentation with text partitioning. *arXiv preprint arXiv:1608.02025*.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005).*

Winston, P. H. (2011). The strong story hypothesis and the directed perception hypothesis.

Wojcik, S. P., A. Hovasapian, J. Graham, M. Motyl, and P. H. Ditto (2015). Conservatives report, but liberals display, greater happiness. *Science 347*(6227), 1243–1246.

Wu, S. (2016, Dec). An interactive visualization of every line in hamilton. http://polygraph.cool/hamilton/.

Xanthos, A., I. Pante, Y. Rochat, and M. Grandjean (2016). Visualising the dynamics of character networks. *Digital Humanities 2016: Conference Abstracts*, 417–419.

Youyou, W., M. Kosinski, and D. Stillwell (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences 112*(4), 1036–1040.

Zhu, X., S. Kiritchenko, and S. M. Mohammad (2014). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 443–447. Citeseer.

Appendix A: Supplementary Material for Sentiment Dictionary Comparisons

A.1 S1 Appendix: Computational methods

All of the code to perform these tests is available and document on GitHub. The repository can be found here: `https://github.com/andyreagan/sentiment-analysis-comparison`.

A.1.1 Stem matching

Of the dictionaries tested, both LIWC and MPQA use "word stems". Here we quickly note some of the technical difficulties with using word stems, and how we processed them, for future research to build upon and improve.

An example is `abandon*`, which is intended to the match words of the standard RE form `abandon[a-z]*`. A naive approach is to check each word against the regular expression, but this is prohibitively slow. We store each of the dictionaries in a "trie" data structure with a record. We use the easily available "marisa-trie" Python library, which wraps the C++ counterpart. The speed of these libraries made the comparison possible over large corpora, in particular for the dictionaries with stemmed words, where the `prefix` search is necessary. Specifically, the "trie" structure is 70 times faster than a regular expression based search for stem words. In particular, we construct two tries for each dictionary: a fixed and stemmed trie. We first attempt to match words against the fixed list, and then turn to the prefix match on the stemmed list.

A.1.2 Regular expression parsing

The first step in processing the text of each corpora is extracting the words from the raw text. Here we rely on a regular expression search, after first removing some punctuation. We choose to include a set of all characters that are found within the words in each of the six dictionaries tested in detail, such that it respects the parse used to create these dictionaries by retaining such characters. This takes the following form in Python, for `raw_text` as a string:

```
punctuation_to_replace = ["---","--","''"]
for punctuation in punctuation_to_replace:
    raw_text = raw_text.replace(punctuation," ")
words = [x.lower() for x in re.findall(\
    r"""(?:[0-9][0-9,\.]*[0-9])|
(?:http://[\w\./\-\?\&\#]+)|
(?:[\w\@\#\'\&\]\[]+)|
(?:[b}/3D;p)|’\-@x#^_0\\P(o:O{X$[=<>\]*B]+)""",
    raw_text,flags=re.UNICODE)]
```

A.2 S2 Appendix: Continued individual comparisons

Picking up right where we left off in Section 3.3, we next compare ANEW with the other dictionaries. The ANEW-WK comparison in Panel I of Fig. 2.1 contains all 1030 words of ANEW, with a fit of $h_{\text{ANEW}}(w) = 1.07 * h_{\text{WK}}(w) - 0.30$, making ANEW more positive and with increasing positivity for more positive words. The 20 most different scores are (ANEW,WK): fame (7.93,5.45), god (8.15,5.90), aggressive (5.10,3.08), casino (6.81,4.68), rancid (4.34,2.38), bees (3.20,5.14), teacher (5.68,7.37), priest (6.42,4.50), aroused (7.97,5.95), skijump (7.06,5.11), noisy (5.02,3.21), heroin (4.36,2.74), insolent (4.35,2.74), rain (5.08,6.58), patient (5.29,6.71), pancakes (6.08,7.43), hospital (5.04,3.52), valentine (8.11,6.40), and book (5.72,7.05). We again see some of the same words from the LabMT comparisons with these dictionaries, and again can attribute some differences to small sample sizes and differing demographics.

For the ANEW-MPQA comparison in Panel J of Fig. 2.1 we show the same matched word lists as before. The happiest 10 words in ANEW matched by MPQA are: clouds (6.18), bar (6.42), mind (6.68), game (6.98), sapphire (7.00), silly (7.41), flirt (7.52), rollercoaster (8.02), comedy (8.37), laughter (8.45). The least happy 5 neutral words and happiest 5 neutral words in MPQA, matched with MPQA, are: pressure (3.38), needle (3.82), quiet (5.58), key (5.68), alert (6.20), surprised (7.47), memories (7.48), knowledge (7.58), nature (7.65), engaged (8.00), baby (8.22). The least happy words in ANEW with score +1 in MPQA that are matched by MPQA are: terrified (1.72), meek (3.87), plain (4.39), obey (4.52), contents (4.89), patient (5.29), reverent (5.35), basket (5.45), repentant (5.53), trumpet (5.75). Again we see some very questionable matches by the MPQA dictionary, with broad stems capturing words with both positive and negative scores.

For the ANEW-LIWC comparison in Panel K of Fig. 2.1 we show the same matched word lists as before. The happiest 10 words in ANEW matched by LIWC are: lazy (4.38), neurotic (4.45), startled (4.50), obsession (4.52), skeptical (4.52), shy (4.64), anxious (4.81), tease (4.84), serious (5.08), aggressive (5.10). There are only 5 words in ANEW that are matched by LIWC with LIWC score of 0: part (5.11), item (5.26), quick (6.64), couple (7.41), millionaire (8.03). The least happy words in ANEW with score +1 in LIWC that are matched by LIWC are: heroin (4.36), virtue (6.22), save (6.45), favor (6.46), innocent (6.51), nice (6.55), trust (6.68), radiant (6.73), glamour (6.76), charm (6.77).

For the ANEW-Liu comparison in Panel L of Fig. 2.1 we show the same matched word lists as before, except the neutral word list because Liu has no explicit neutral words. The happiest 10 words in ANEW matched by Liu are: pig (5.07), aggressive (5.10), tank (5.16), busybody (5.17), hard (5.22), mischief (5.57), silly (7.41), flirt (7.52), rollercoaster (8.02), joke (8.10). The least happy words in ANEW with score +1 in Liu that are matched by Liu are: defeated (2.34), obsession (4.52), patient (5.29), reverent (5.35), quiet (5.58), trumpet (5.75), modest (5.76), humble (5.86), salute (5.92), idol (6.12).

For the WK-MPQA comparison in Panel P of Fig. 2.1 we show the same matched word lists as before. The happiest 10 words in WK matched by MPQA are: cutie (7.43), pancakes (7.43), panda (7.55), laugh (7.56), marriage (7.56), lullaby (7.57), fudge (7.62), pancake (7.71), comedy (8.05), laughter (8.05). The least happy 5 neutral words and happiest 5 neutral words in MPQA, matched with MPQA, are: sociopath (2.44), infectious (2.63), sob (2.65), soulless (2.71), infertility (3.00), thinker (7.26), knowledge (7.28), legacy (7.38), surprise (7.44), song (7.59). The least happy words in WK with score +1 in MPQA that are matched by MPQA are: kidnapper (1.77), kidnapping (2.05), kidnap (2.19), discriminating (2.33), terrified (2.51), terrifying (2.63), terrify (2.84), courtroom (2.84), backfire (3.00), indebted (3.21).

For the WK-LIWC comparison in Panel Q of Fig. 2.1 we show the same matched word lists as before. The happiest 10 words in WK matched by LIWC are: geek (5.56), number (5.59), fiery (5.70), trivia (5.70), screwdriver (5.76), foolproof (5.82), serious (5.88), yearn (5.95), dumpling (6.48), weeping willow (6.53). The least happy 5 neutral words and happiest 5 neutral words in LIWC, matched with LIWC, are: negative (2.52), negativity (2.74), quicksand (3.62), lack (3.68), wont (4.09), unique (7.32), millionaire (7.32), first (7.33), million (7.55), rest (7.86). The least happy words in WK with score +1 in LIWC that are matched by LIWC are: heroin (2.74), friendless (3.15), promiscuous (3.32), supremacy (3.48), faithless (3.57), laughingstock (3.77), promiscuity (3.95), tenderfoot (4.26), succession (4.52), dynamite (4.79).

For the WK-Liu comparison in Panel R of Fig. 2.1 we show the same matched word lists as before, except the neutral word list because Liu has no explicit neutral words. The happiest 10 words in WK matched by Liu are: goofy (6.71), silly (6.72), flirt (6.73), rollercoaster (6.75), tenderness (6.89), shimmer (6.95), comical (6.95), fanciful (7.05), funny (7.59), fudge (7.62), joke (7.88). The least happy words in WK with score +1 in Liu that are matched by Liu are: defeated (2.59), envy (3.05), indebted (3.21), supremacy (3.48), defeat (3.74), overtake (3.95), trump (4.18), obsession (4.38), dominate (4.40), tough (4.45).

Now we'll focus our attention on the MPQA row, and first we see comparisons against the three full range dictionaries. For the first match against LabMT in Panel D of Fig. 2.1, the MPQA match catches 431 words with MPQA score 0, while LabMT (without stems) matches 268 words in MPQA in Panel S (1039/809 and 886/766 for the positive and negative words of MPQA). Since we've already highlighted most of these words, we move on and focus our attention on comparing the ±1 dictionaries.

In Panels V–X, BB–DD, and HH–JJ of Fig. 2.1 there are a total of 6 bins off of the diagonal, and we focus out attention on the bins that represent words that have opposite scores in each of the dictionaries. For example, consider the matches made my MPQA in Panel BB: the words in the top left corner and bottom right corner with are scored in a opposite manner in LIWC, and are of particular concern. Looking at the words from Panel W with a +1 in MPQA and a -1 in LIWC (matched by LIWC) we see: stunned, fiery, terrified, terrifying, yearn, defense, doubtless, foolproof, risk-free, exhaustively, exhaustive, blameless, low-risk, low-cost, lower-priced, guiltless, vulnerable, yearningly, and yearning. The words with a -1 in MPQA that are +1 in LIWC (matched by LIWC) are: silly, madly, flirt, laugh, keen, superiority, supremacy, sillily, dearth, comedy, challenge, challenging, cheerless, faithless, laughable, laughably, laughingstock, laughter, laugh, grating, opportunistic, joker, challenge, flirty.

In Panel W of 2.1, the words with a +1 in MPQA and a -1 in Liu (matched by Liu) are: solicitude, flair, funny, resurgent, untouched, tenderness, giddy, vulnerable, and joke. The words with a -1 in MPQA that are +1 in Liu, matched by Liu, are: superiority, supremacy, sharp, defeat, dumbfounded, affectation, charisma, formidable, envy, empathy, trivially, obsessions, and obsession.

In Panel BB of 2.1, the words with a +1 in LIWC and a -1 in MQPA (matched by MPQA) are: silly, madly, flirt, laugh, keen, determined, determina, funn, fearless, painl, cute, cutie, and gratef.

The words with a -1 in LIWC and a +1 in MQPA, that are matched by MPQA, are: stunned, terrified, terrifying, fiery, yearn, terrify, aversi, pressur, careless, helpless, and hopeless.

In Panel DD of 2.1, the words with a -1 in LIWC and a +1 in Liu, that are matched by Liu, are: silly, and madly. The words with a +1 in LIWC and a -1 in Liu, that are matched by Liu, are: stunned, and fiery.

In Panel HH of 2.1, the words with a -1 in Liu and a +1 in MPQA, that are matched by MPQA, are: superiority, supremacy, sharp, defeat, dumbfounded, charisma, affectation, formidable, envy, empathy, trivially, obsessions, obsession, stabilize, defeated, defeating, defeats, dominated, dominates, dominate, dumbfounding, cajole, cuteness, faultless, flashy, fine-looking, finer, finest, panoramic, pain-free, retractable, believeable, blockbuster, empathize, err-free, mind-blowing, marvelled, marveled, trouble-free, thumb-up, thumbs-up, long-lasting, and viewable. The words with a +1 in Liu and a -1 in MPQA, that are matched by MPQA, are: solicitude, flair, funny, resurgent, untouched, tenderness, giddy, vulnerable, joke, shimmer, spurn, craven, aweful, backwoods, backwood, back-woods, back-wood, back-logged, backaches, backache, backaching, backbite, tingled, glower, and gainsay.

In Panel II of 2.1, the words with a +1 in Liu and a -1 in LIWC, that are matched by LIWC, are: stunned, fiery, defeated, defeating, defeats, defeat, doubtless, dominated, dominates, dominate, dumbfounded, dumbfounding, faultless, foolproof, problem-free, problem-solver, risk-free, blameless, envy, trivially, trouble-free, tougher, toughest, tough, low-priced, low-price, low-risk, low-cost, lower-priced, geekier, geeky, guiltless, obsessions, and obsession. The words with a -1 in Liu and a +1 in LIWC, that are matched by LIWC, are: silly, madly, sillily, dearth, challenging, cheerless, faithless, flirty, flirt, funnily, funny, tenderness, laughable, laughably, laughingstock, grating, opportunistic, joker, and joke.

In the off-diagonal bins for all of the ± 1 dictionaries, we see many of the same words. Again MPQA stem matches are disparagingly broad. We also find matches by LIWC that are concerning, and should in all likelihood be removed from the dictionary.

A.3 S3 Appendix: Coverage for all corpuses

We provide coverage plots for the other three corpuses.

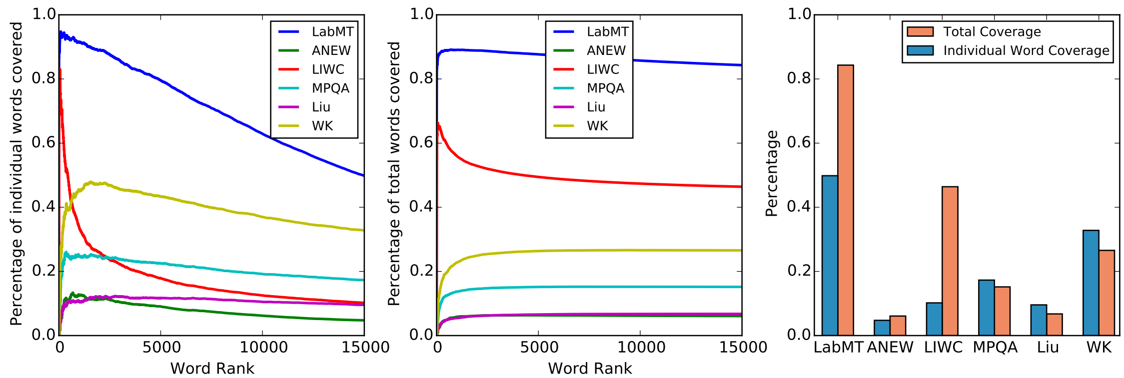


Figure A.1: Coverage of the words on twitter by each of the dictionaries.

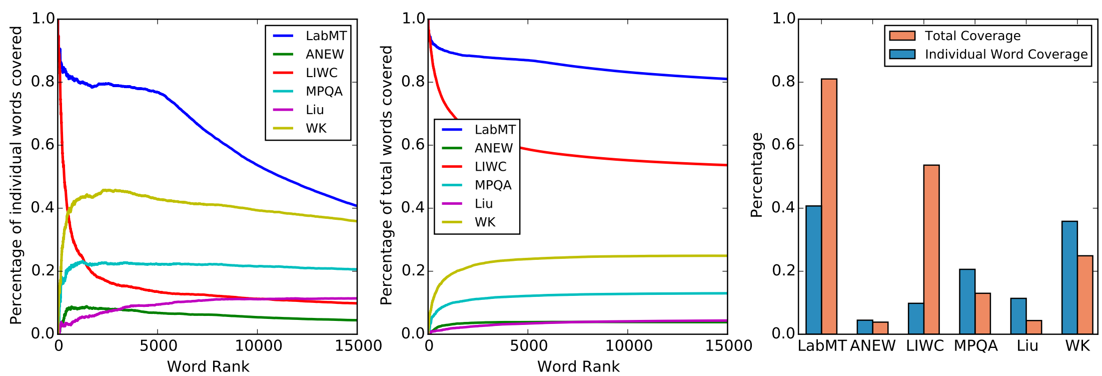


Figure A.2: Coverage of the words in Google books by each of the dictionaries.

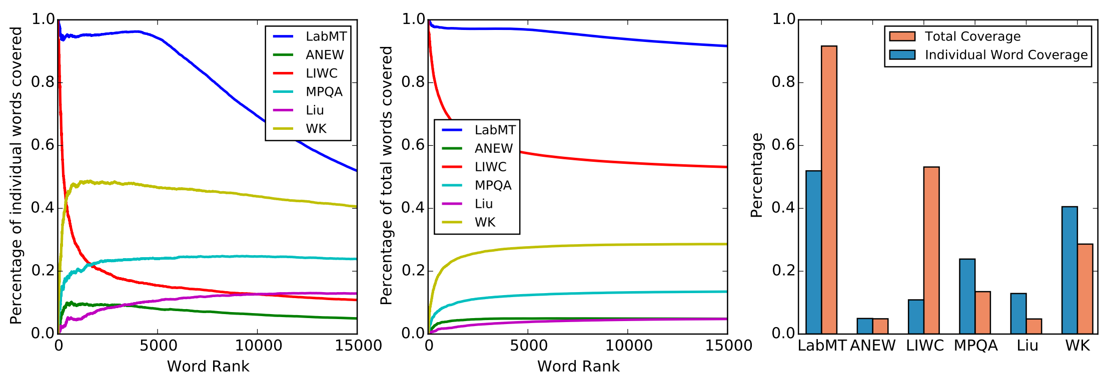


Figure A.3: Coverage of the words in the New York Times by each of the dictionaries.

A.4 S4 Appendix: Sorted New York Times rankings

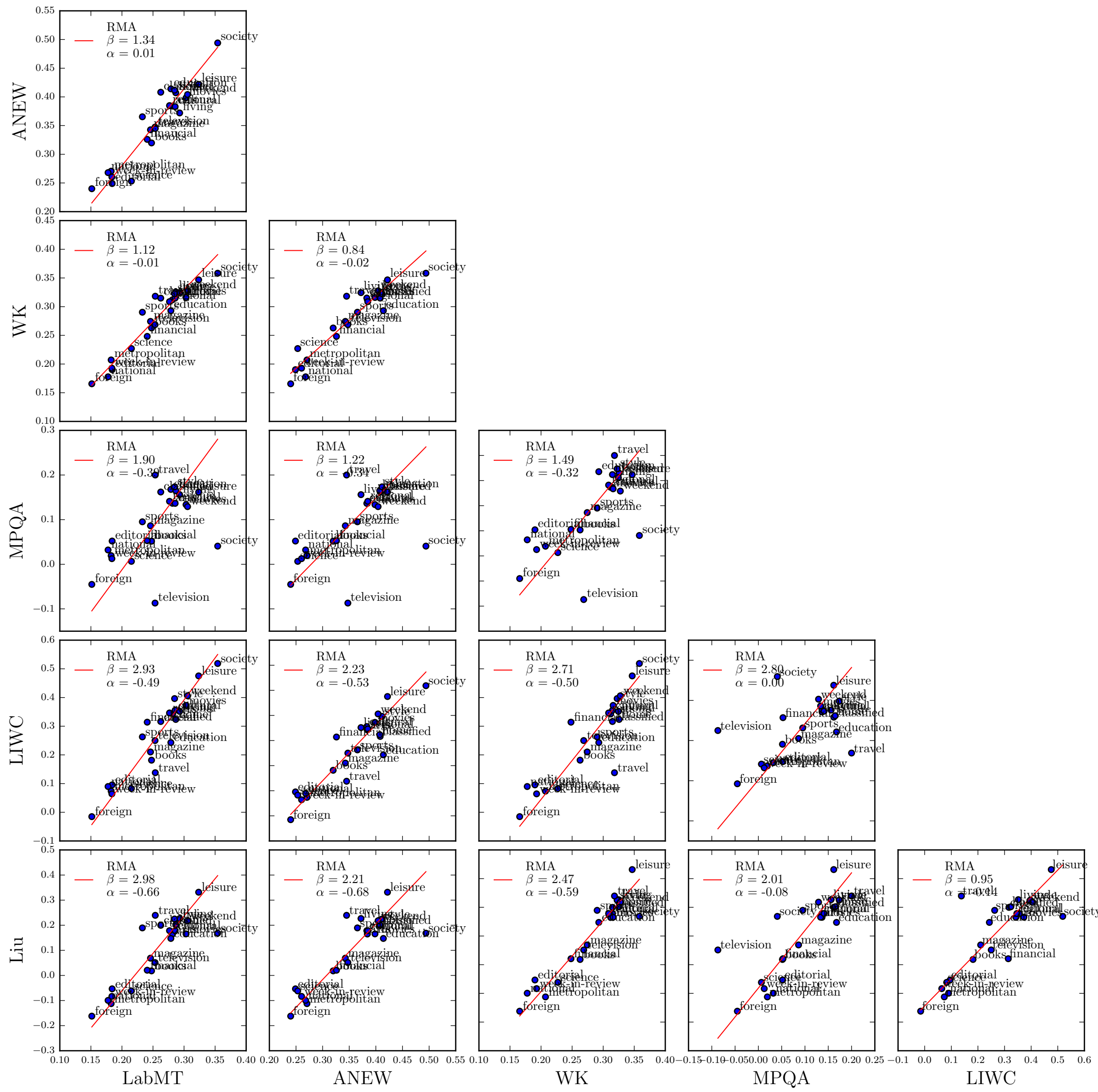


Figure A.4: NYT Sections scatterplot. The RMA fit α and β for the formula $y = \alpha + \beta x$. For the sake of comparison, we normalized each dictionary to the range [-.5,.5] by subtracting the mean score (5 or 0) and dividing by the range (8 or 2).

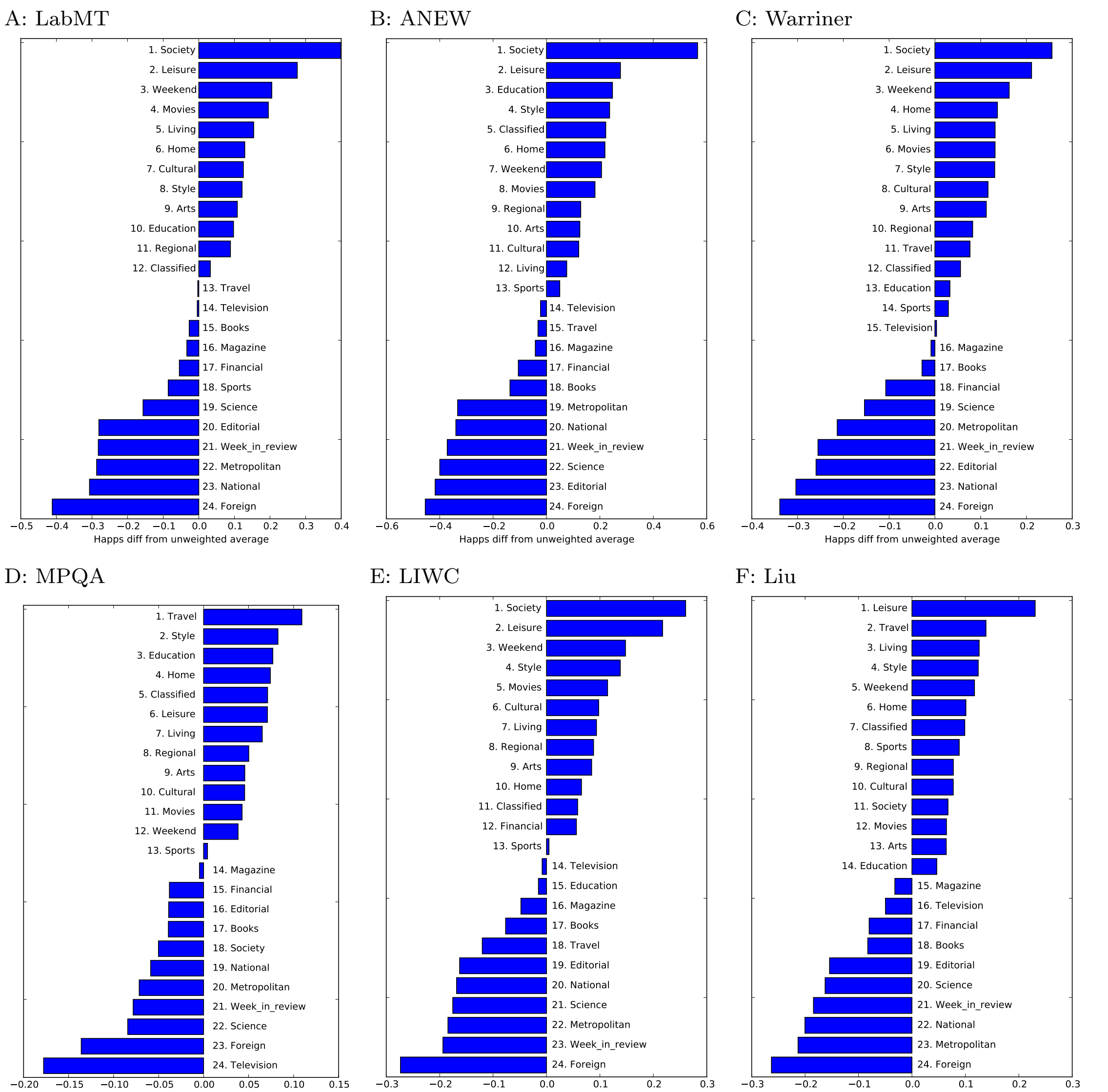


Figure A.5: Sorted bar charts ranking each of the 24 New York Times Sections for each dictionary tested.

A.5 S5 Appendix: Movie Review Distributions

Here we examine the distributions of movie review scores. These distributions are each summarized by their mean and standard deviation in panels of Figure 2 for each dictionary. For example, the left most error bar of each panel in Figure 2 shows the standard deviation around the mean for the distribution of individual review scores (Figure A.6).

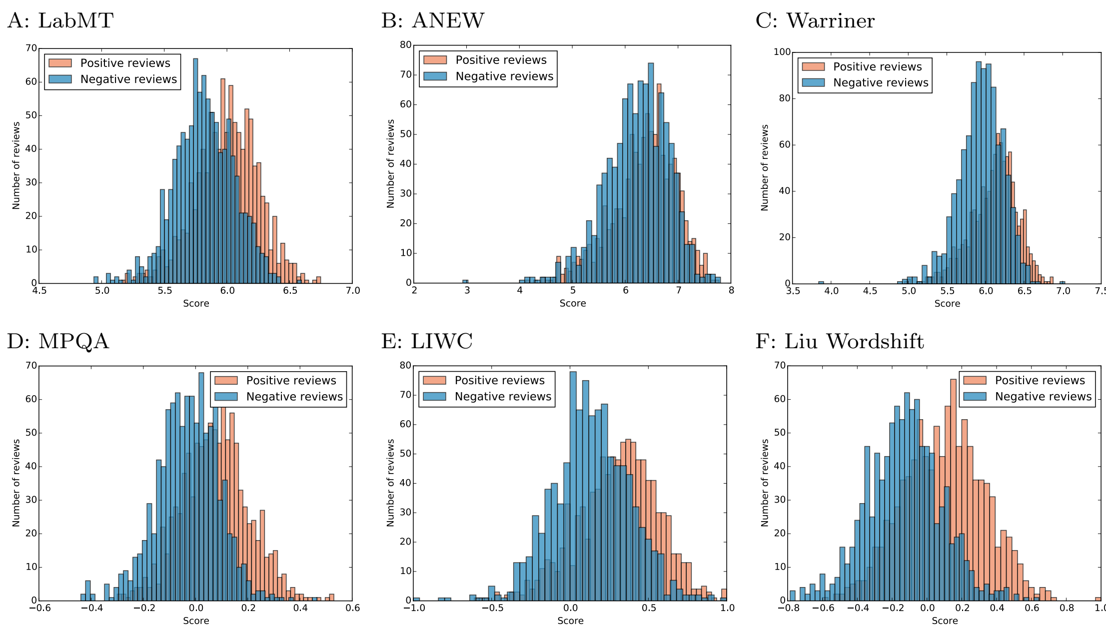


Figure A.6: Binned scores for each review by each corpus with a stop value of $\Delta_h = 1.0$.

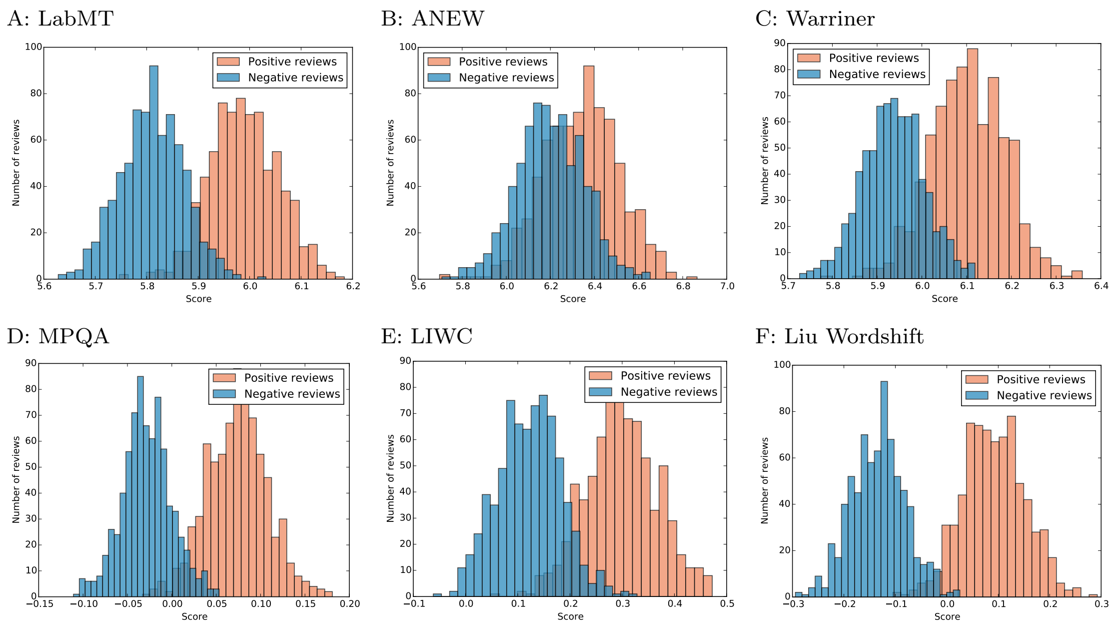


Figure A.7: Binned scores for samples of 15 concatenated random reviews. Each dictionary uses stop value of $\Delta_h = 1.0$.

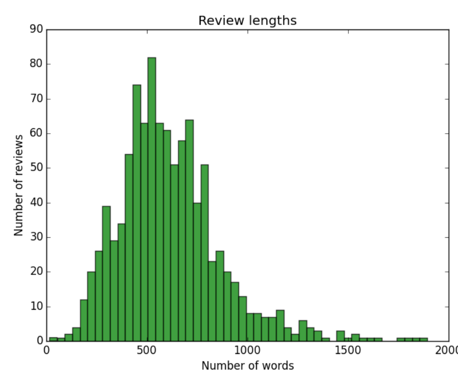


Figure A.8: Binned length of positive reviews, in words.

A.6 S6 Appendix: Google Books correlations and word shifts

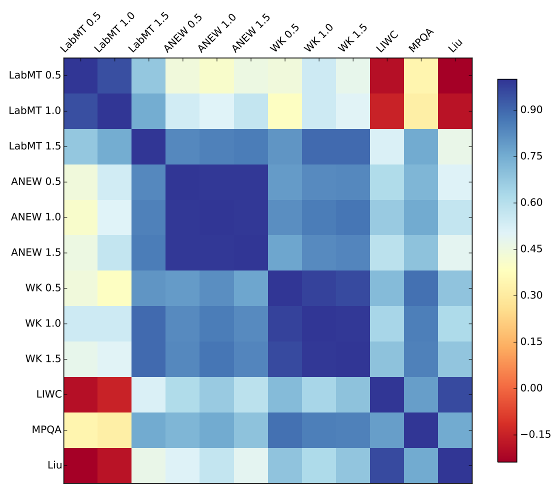


Figure A.9: Google Books correlations. Here we include correlations for the google books time series, and word shifts for selected decades (1920's,1940's,1990's,2000's).

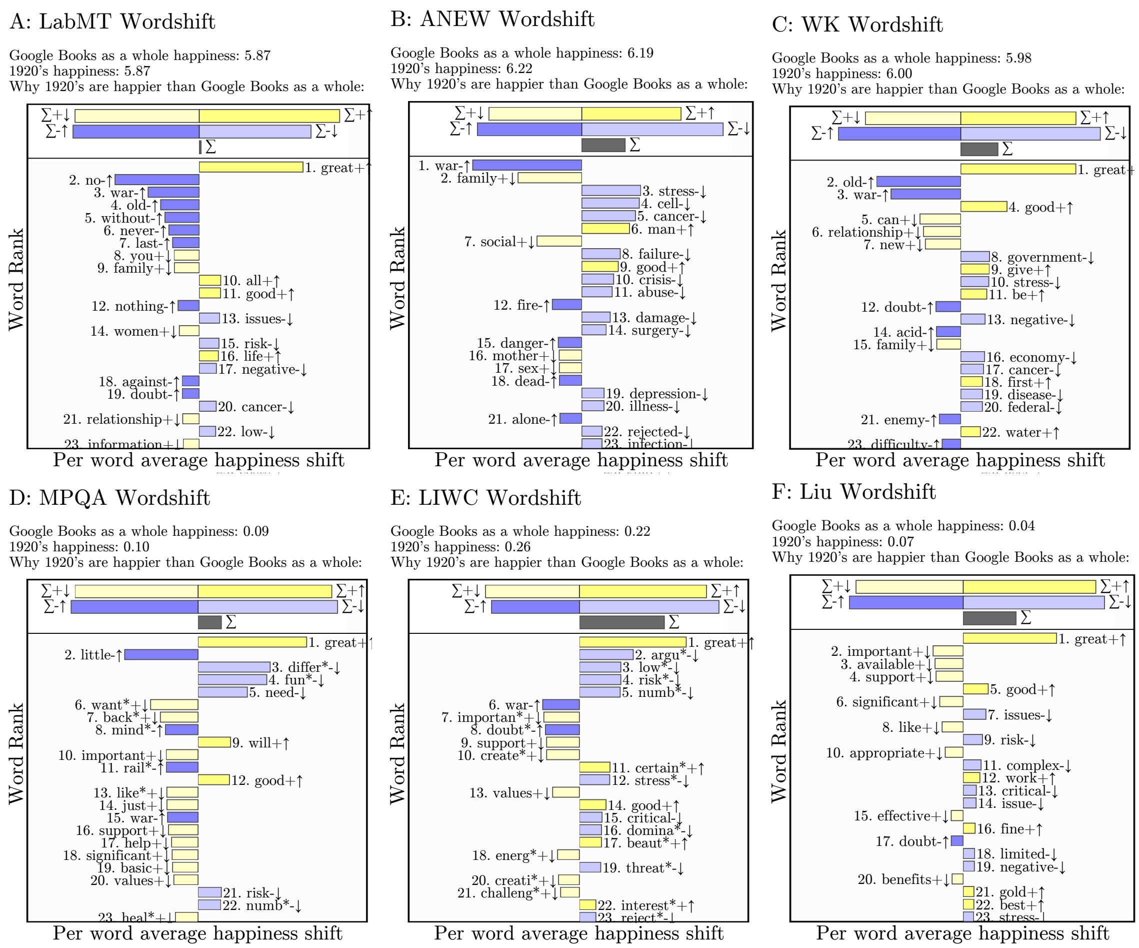


Figure A.10: Google Books shifts in the 1920's against the baseline of Google Books.

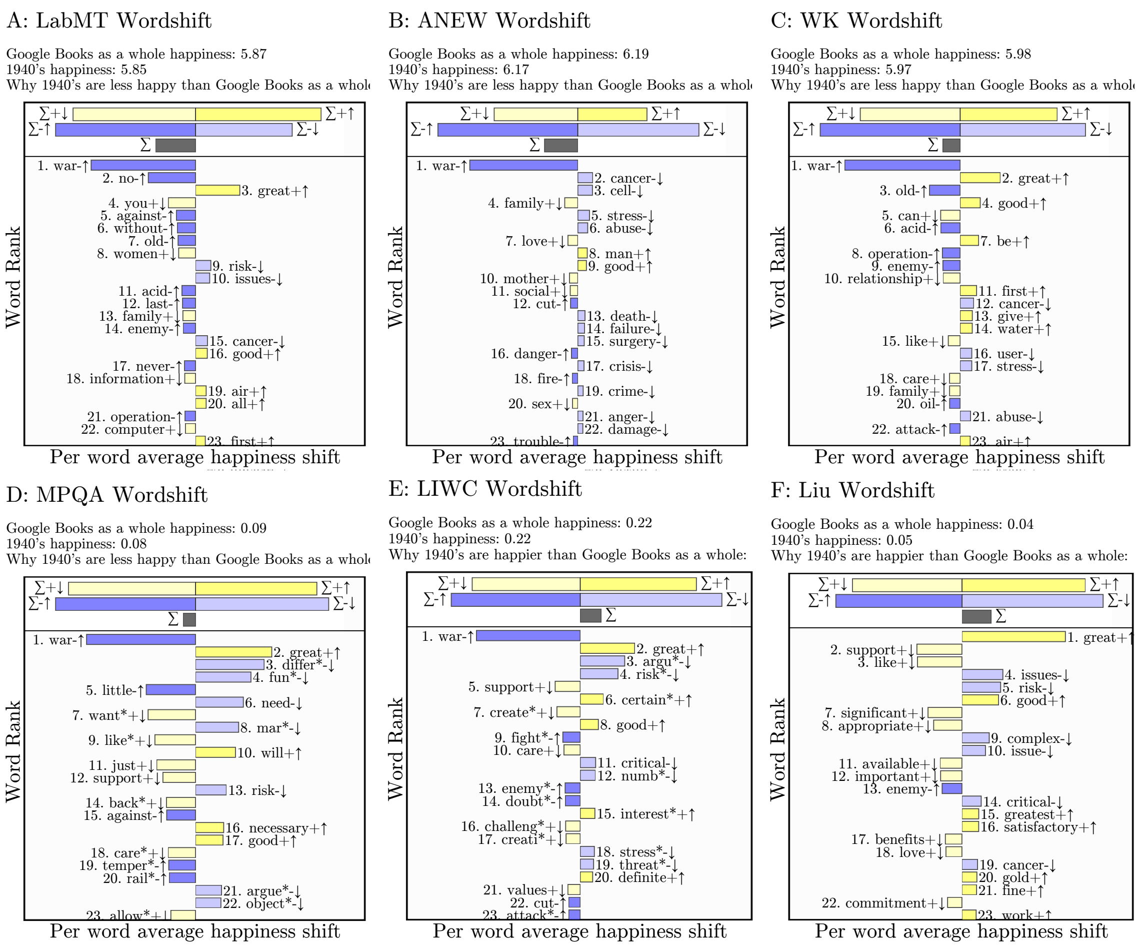


Figure A.11: Google Books shifts in the 1940's against the baseline of Google Books.

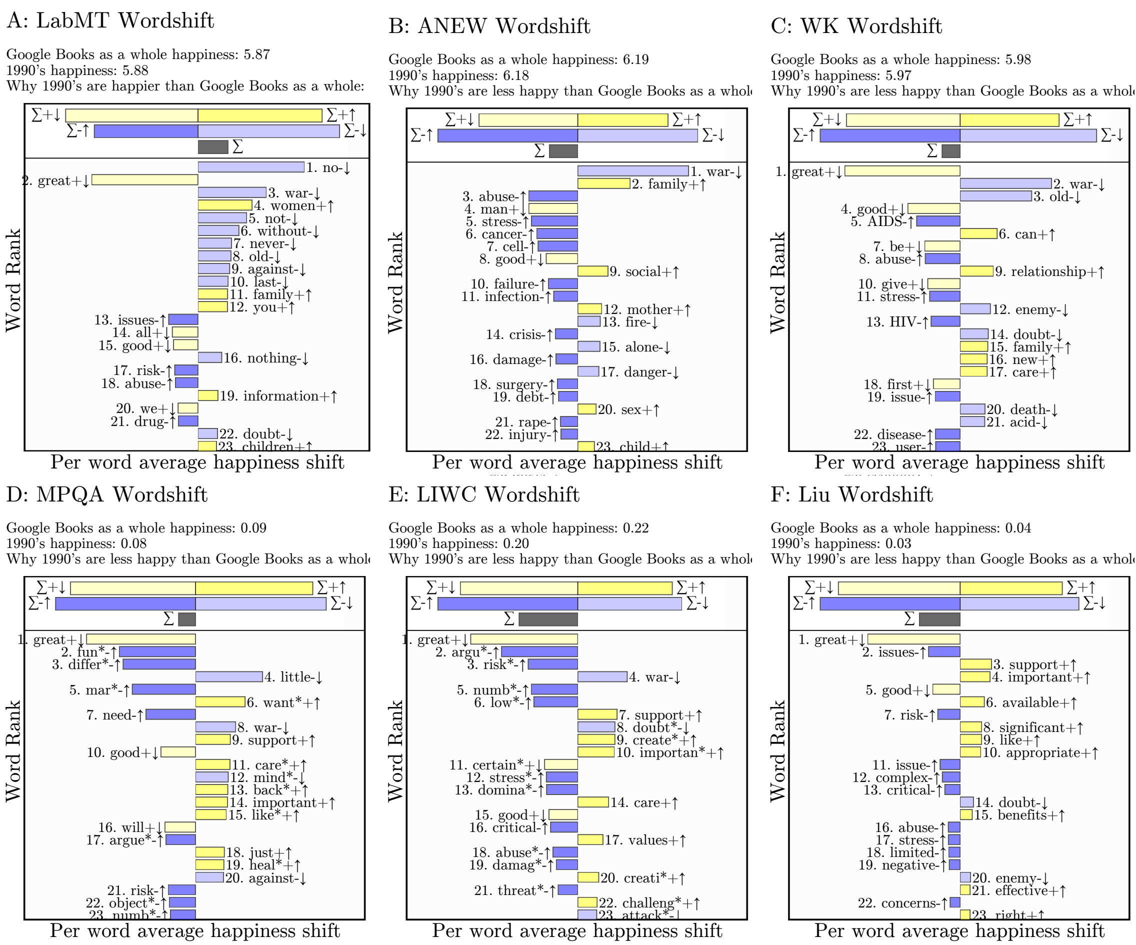


Figure A.12: Google Books shifts in the 1990's against the baseline of Google Books.

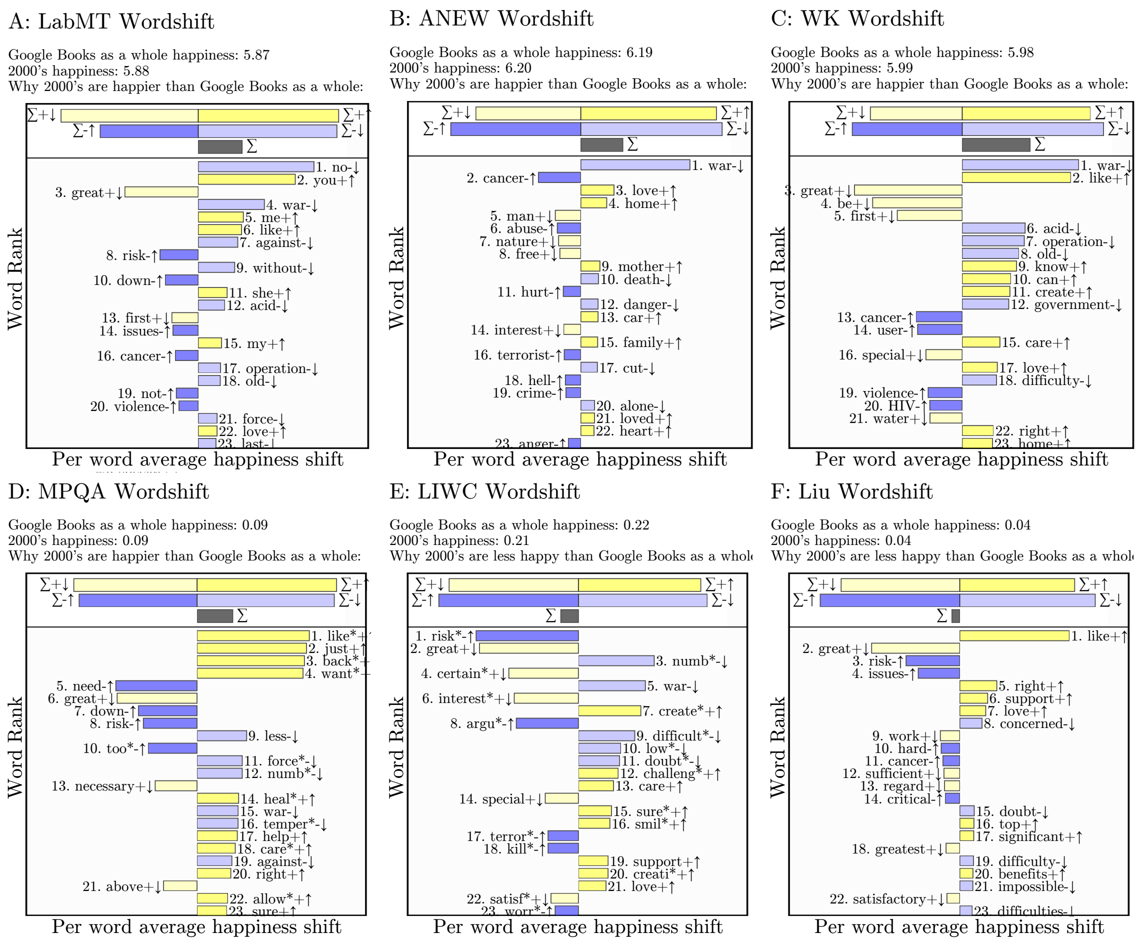


Figure A.13: Google Books shifts in the 2000's against the baseline of Google Books.

A.7 S7 Appendix: Additional Twitter time series, correlations, and shifts

First, we present additional Twitter time series:

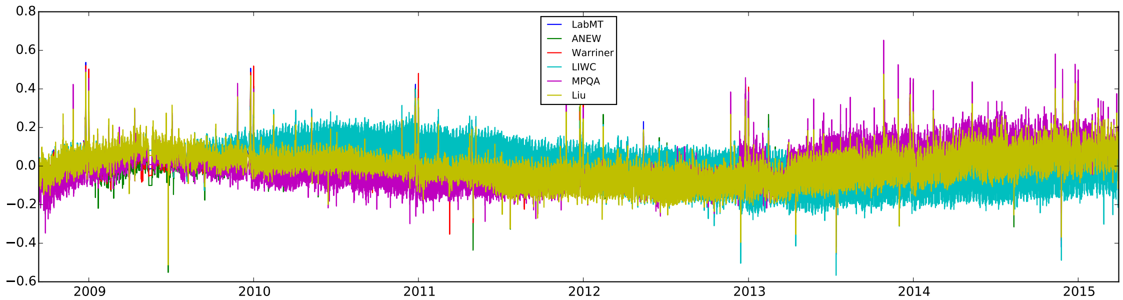


Figure A.14: Normalized time series on Twitter using Δ_h of 1.0 for all. For resolution of 3 hours. We do not include any of the time series with resolution below 3 hours here because there are too many data points to see.

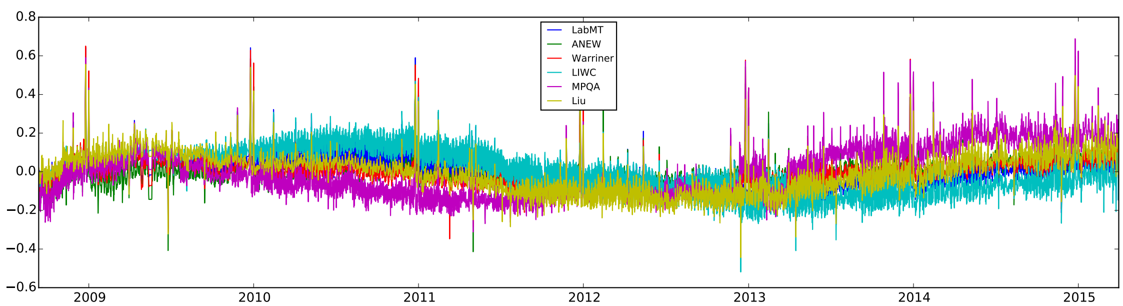


Figure A.15: Normalized time series on Twitter using Δ_h of 1.0 for all. For resolution of 12 hours.

Next, we take a look at more correlations:

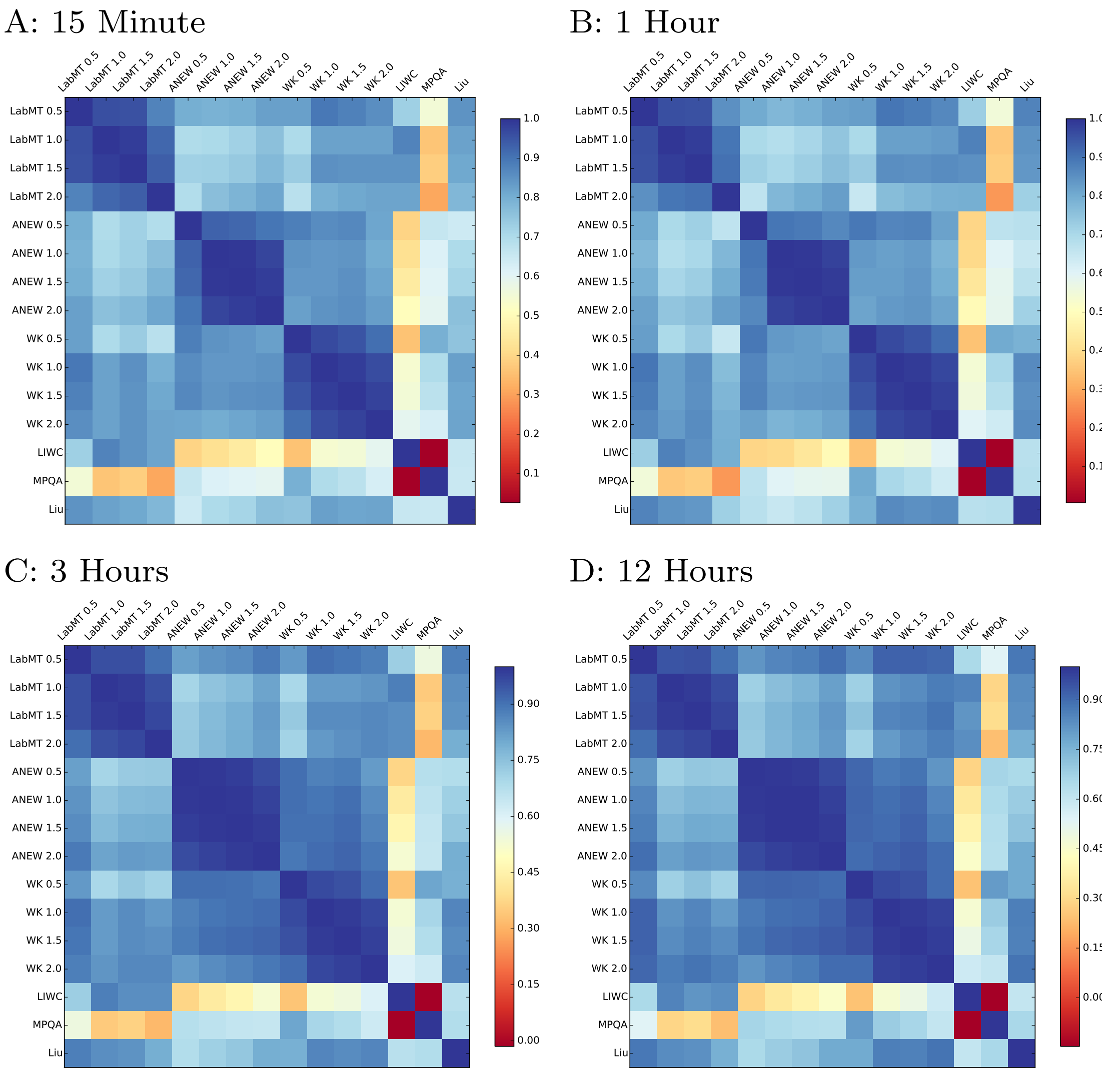


Figure A.16: Pearson's r correlation between Twitter time series for all resolutions below 1 day.

Now we include word shift graphs that are absent from the manuscript itself.

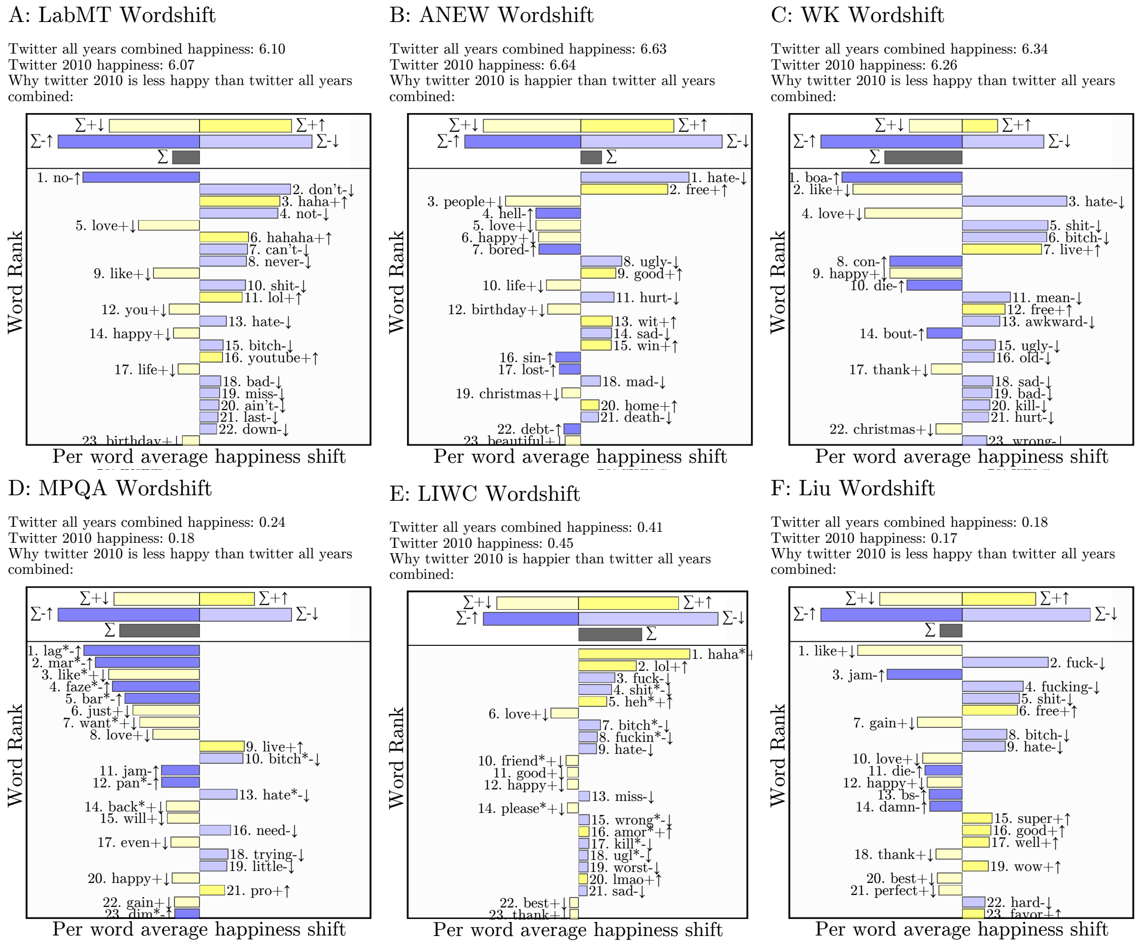


Figure A.17: Word Shifts for Twitter in 2010. The reference word usage is all of Twitter (the 10% Gardenhose feed) from September 2008 through April 2015, with the word usage normalized by year.

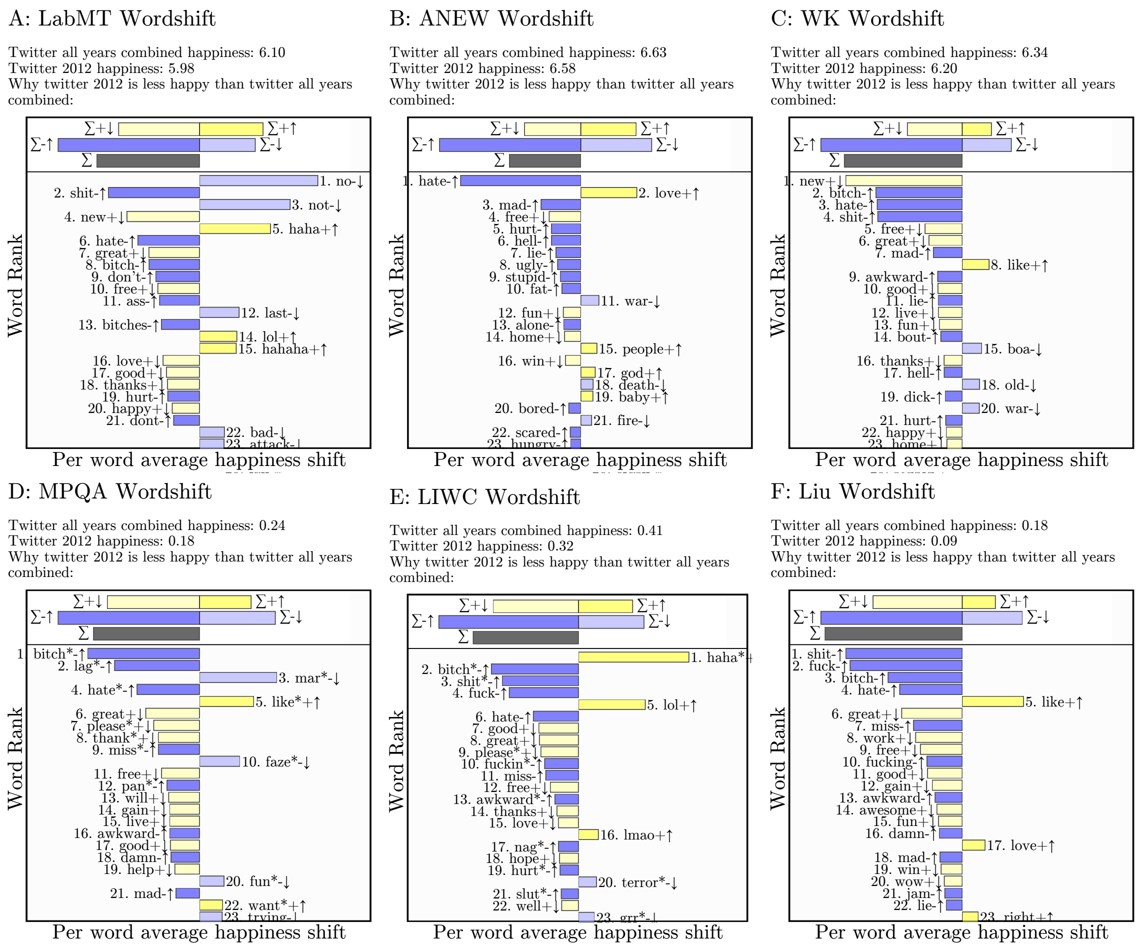


Figure A.18: Word Shifts for Twitter in 2012. The reference word usage is all of Twitter (the 10% Gardenhose feed) from September 2008 through April 2015, with the word usage normalized by year.

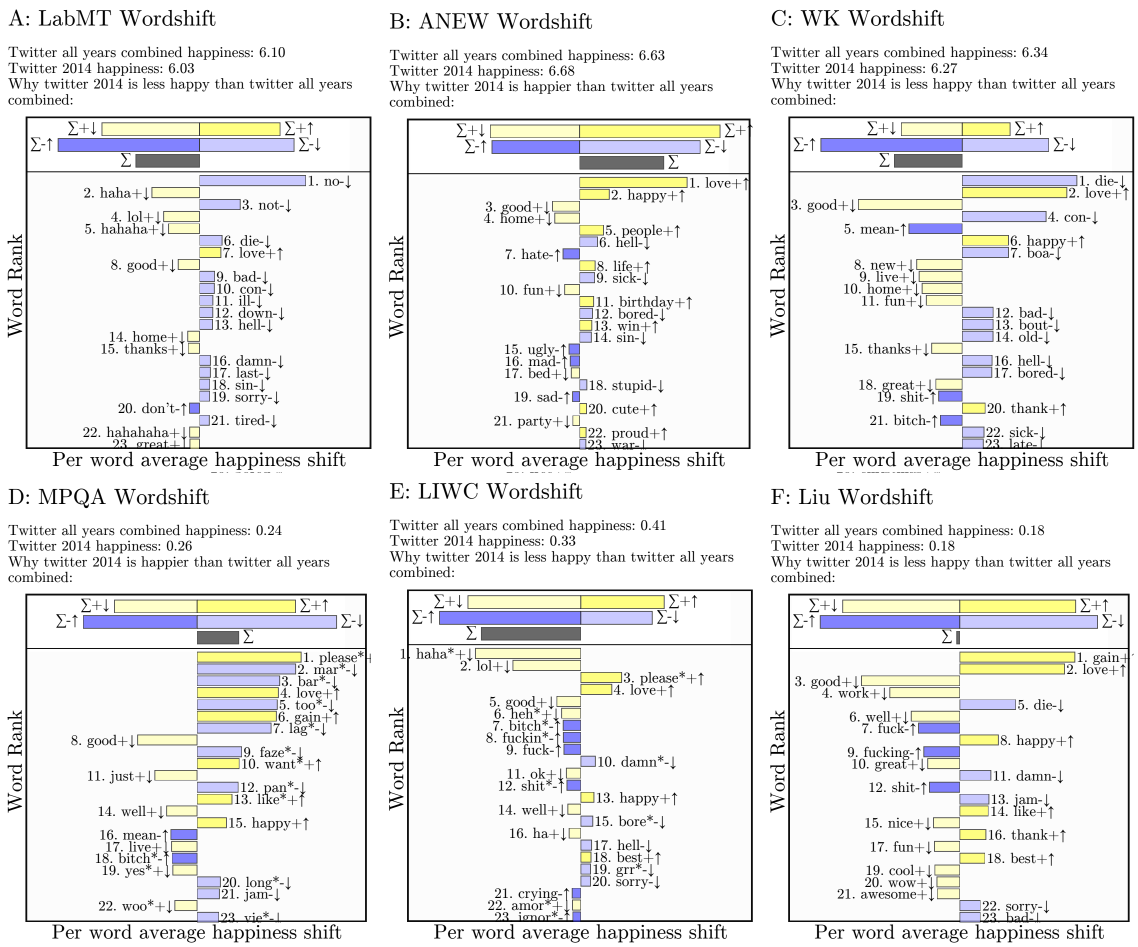


Figure A.19: Word Shifts for Twitter in 2014. The reference word usage is all of Twitter (the 10% Gardenhose feed) from September 2008 through April 2015, with the word usage normalized by year.

Finally, we include the results of each dictionary applied to a set of annotated Twitter data. We apply sentiment dictionaries to rate individual Tweets and classify a Tweet as positive (negative) if the Tweet rating is greater (less) than the average of all scores in dictionary.

Rank	Dictionary	% Tweets scored	F1 of Tweets scored	Calibrated F1	Overall F1
1.	Sent140Lex	100.0	0.89	0.88	0.89
2.	labMT	100.0	0.69	0.78	0.69
3.	HashtagSent	100.0	0.67	0.64	0.67
4.	SentiWordNet	98.6	0.67	0.68	0.67
5.	VADER	81.3	0.75	0.81	0.61
6.	SentiStrength	73.9	0.83	0.81	0.61
7.	SenticNet	97.3	0.61	0.64	0.59
8.	Umigon	67.1	0.87	0.85	0.58
9.	SOCAL	82.2	0.71	0.75	0.58
10.	WDAL	99.9	0.58	0.64	0.58
11.	AFINN	73.6	0.78	0.80	0.57
12.	OL	66.7	0.83	0.82	0.55
13.	MaxDiff	94.1	0.58	0.70	0.54
14.	EmoSenticNet	96.0	0.56	0.59	0.54
15.	MPQA	73.2	0.73	0.72	0.53
16.	WK	96.5	0.53	0.72	0.51
17.	LIWC15	61.8	0.81	0.78	0.50
18.	Pattern	69.0	0.71	0.75	0.49
19.	GI	67.6	0.72	0.70	0.49
20.	LIWC07	60.3	0.80	0.75	0.48
21.	LIWC01	54.3	0.83	0.75	0.45
22.	EmoLex	59.4	0.73	0.69	0.43
23.	ANEW	64.1	0.65	0.68	0.42
24.	USent	4.5	0.74	0.73	0.03
25.	PANAS-X	1.7	0.88	–	0.01
26.	Emoticons	1.4	0.72	0.77	0.01

Table A.1: Ranked results of sentiment dictionary performance on individual Tweets from STS-Gold dataset (Saif, 2013). We report the percentage of Tweets for which each dictionary contains at least 1 entry, the F1 score on those Tweets, and the overall classification F1 score. The calibrated F1 score tunes the decision threshold between positive and negative Tweets with a random 10% training sample.

A.8 S8 Appendix: Naive Bayes results and derivation

We now provide more details on the implementation of Naive Bayes, a derivation of the linearity structure, and more results from the classification of Movie Reviews.

First, to implement a binary Naive Bayes classifier for a collection of documents, we denote each of the N words in the given document T as w_i, thus the normalized word frequency is $f_i(T) = w_i/N$, and finally we denote the class labels c_1, c_2. The probability of a document T belonging to class c_1 can be written as

$$P(c_1|T) = \frac{P(c_1)P(T|c_1)}{P(T)}.$$

Since we do not know $P(T|c_1)$ explicitly, we make the *naive* assumption that each word appears independently, and thus write

$$P(c_1|T) = \frac{P(c_1) \cdot [P(f_1(T)|c_1) \cdot P(f_2(T)|c_1) \cdots P(f_N(T)|c_1)]}{P(T)}.$$

Since we are only interested in comparing $P(c_1|T)$ and $P(c_2|T)$, we disregard the shared denominator and have

$$P(c_1|T) \propto P(c_1) \cdot [P(f_1(T)|c_1) \cdot P(f_2(T)|c_1) \cdots P(f_N(T)|c_1)] \,.$$

Finally we say that document T belongs to class c_1 if $P(c_1|T) > P(c_2|T)$. Given that the probabilities of individual words are small, to avoid machine truncation error we compute these probabilities in log space, such that the product of individual word likelihoods becomes a sum

$$\log P(c_1|T) \propto \log P(c_1) + \sum_{i=1}^{N} \log P(f_i(T)|c_1).$$

Assigning a classification of class c_1 if $P(c_1|T) > P(c_2|T)$ is the same as saying that the difference between the two is positive, i.e. $P(c_1|T) - P(c_2|T) > 0$ and since the logarithm is monotonic, $\log P(c_1|T) - \log P(c_2|T) > 0$. To examine how individual words contribute to this difference, we can write

$$0 < \log P(c_1|T) - \log P(c_2|T)$$

$$\propto \log P(c_1) + \sum_{i=1}^{N} \log P(f_i(T)|c_1) - \log P(c_2) - \sum_{i=1}^{N} \log P(f_i(T)|c_2)$$

$$\propto \log P(c_1) - \log P(c_2) + \sum_{i=1}^{N} [\log P(f_i(T)|c_1) - \log P(f_i(T)|c_2)]$$

$$\propto \log \frac{P(c_1)}{P(c_2)} + \sum_{i=1}^{N} \log \frac{P(f_i(T)|c_1)}{P(f_i(T)|c_2)}.$$

We can see from the above that the contribution of each word w_i (or more accurately, the likelihood of the frequency in document T being predictive of class c as $P(f_i(T)|c_1)$) is a linear constituent of the classification.

Next, we include the detailed results of the Naive Bayes classifier on the Movie Review corpus.

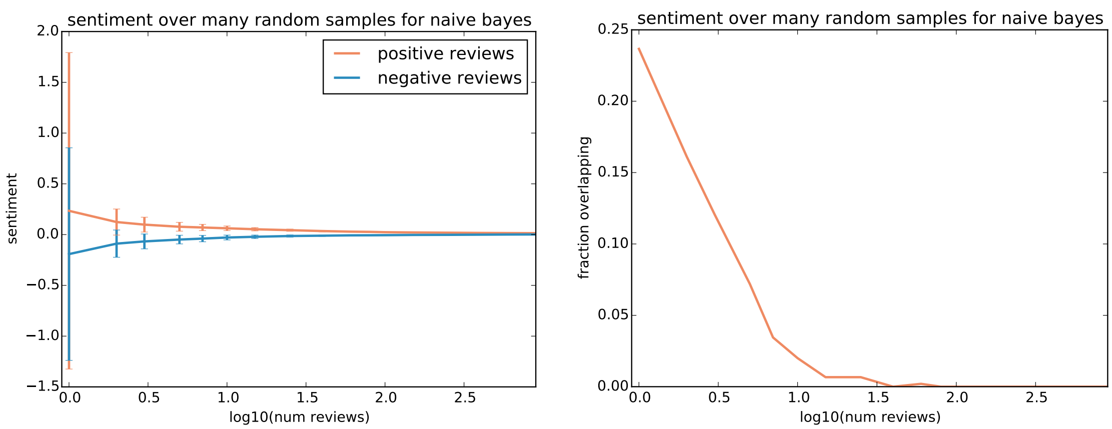


Figure A.20: Results of the NB classifier on the Movie Reviews corpus.

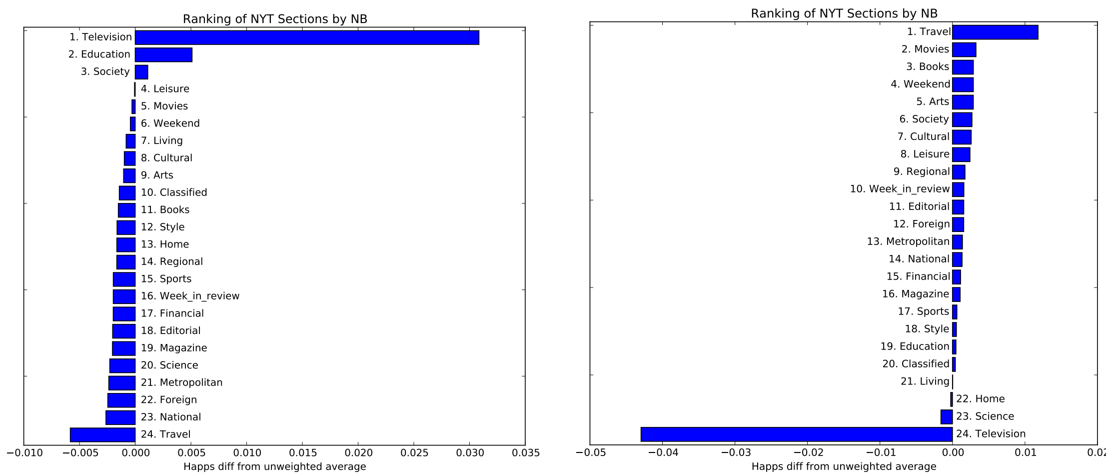


Figure A.21: NYT Sections ranked by Naive Bayes in two of the five trials.

Most informative			
Positive		Negative	
Word	Value	Word	Value
27.27	flynt	20.21	godzilla
26.33	truman	15.95	werewolf
20.68	charles	13.83	gorilla
15.04	event	13.83	spice
14.10	shrek	13.83	memphis
13.16	cusack	13.83	sgt
13.16	bulworth	12.76	jennifer
13.16	robocop	12.76	hill
12.22	jedi	11.70	max
12.22	gangster	11.70	200

NYT Society			
Positive		Negative	
Word	Value	Word	Value
26.08	truman	20.40	godzilla
20.49	charles	12.88	hill
12.11	gangster	12.88	jennifer
10.25	speech	10.73	fatal
9.32	melvin	8.59	freddie
8.85	wars	8.59	=
7.45	agents	8.59	mess
6.52	dance	8.59	gene
6.52	bleak	8.59	apparent
6.52	pitt	7.51	travolta

Table A.2: Trial 1 of Naive Bayes trained on a random 10% of the movie review corpus, and applied to the New York Times Society section. We show the words which are used by the trained classifier to classify individual reviews (in corpus), and on the New York Times (out of corpus). In addition, we report a second trial in Table A.3, since Naive Bayes is trained on a random subset of data, to show the variation in individual words between trials (while performance is consistent).

Most informative			
Positive		Negative	
Word	Value	Word	Value
18.11	shrek	34.63	west
17.15	poker	24.14	webb
15.25	shark	18.89	jackal
14.29	maggie	17.84	travolta
13.34	guido	17.84	woo
13.34	outstanding	17.84	coach
13.34	political	16.79	awful
13.34	journey	16.79	brenner
13.34	bulworth	15.74	gabriel
12.39	bacon	15.74	general's

NYT Society			
Positive		Negative	
Word	Value	Word	Value
17.79	poker	33.39	west
13.84	journey	17.20	coach
13.84	political	17.20	travolta
8.90	tribe	15.18	gabriel
7.91	tony	12.14	pointless
7.91	price	9.44	stupid
7.91	threat	8.09	screaming
7.12	titanic	7.59	mess
6.92	dicaprio	7.42	boring
6.92	kate	7.08	=

Table A.3: Trial 2 of Naive Bayes trained on a random 10% of the movie review corpus, and applied to the New York Times Society section. We show the words which are used by the trained classifier to classify individual reviews (in corpus), and on the New York Times (out of corpus). This second trial is in addition to the first trial in Table A.2, since Naive Bayes is trained on a random subset of data, to show the variation in individual words between trials (while performance is consistent).

A.9 S9 Appendix: Movie review benchmark of additional dictionaries

Here, we present the accuracy of each dictionary applied to binary classification of Movie Reviews.

Rank	Title	% Scored	F1 Trained	F1 Untrained
1.	OL	100	0.70	0.71
2.	HashtagSent	100	0.67	0.66
3.	MPQA	100	0.67	0.66
4.	SentiWordNet	100	0.65	0.65
5.	labMT	100	0.64	0.63
6.	AFINN	100	0.67	0.63
7.	Umigon	100	0.65	0.62
8.	GI	100	0.65	0.61
9.	SOCAL	100	0.71	0.60
10.	VADER	100	0.67	0.60
11.	WDAL	100	0.60	0.59
12.	SentiStrength	100	0.63	0.58
13.	EmoLex	100	0.65	0.56
14.	LIWC15	100	0.64	0.55
15.	LIWC01	100	0.65	0.54
16.	LIWC07	100	0.64	0.53
17.	Pattern	100	0.73	0.52
18.	PANAS-X	33	0.51	0.51
19.	Sent140Lex	100	0.68	0.47
20.	SenticNet	100	0.62	0.45
21.	ANEW	100	0.57	0.36
22.	MaxDiff	100	0.66	0.36
23.	EmoSenticNet	100	0.58	0.34
24.	WK	100	0.63	0.34
25.	Emoticons	0	–	–
26.	USent	40	–	–

Table A.4: Ranked performance of dictionaries on the Movie Review corpus.

Rank	Title	% Scored	F1 Trained of Scored	F1 Untrained of Scored	F1 Untrained, All
1.	HashtagSent	100	0.55	0.55	0.55
2.	LIWC15	99	0.53	0.55	0.55
3.	LIWC07	99	0.53	0.55	0.54
4.	LIWC01	99	0.52	0.55	0.54
5.	labMT	99	0.54	0.54	0.54
6.	Sent140Lex	100	0.55	0.54	0.54
7.	SentiWordNet	99	0.54	0.53	0.53
8.	WDAL	99	0.53	0.53	0.52
9.	EmoLex	95	0.54	0.55	0.52
10.	MPQA	93	0.54	0.55	0.52
11.	SenticNet	97	0.53	0.52	0.50
12.	SOCAL	88	0.56	0.55	0.49
13.	EmoSenticNet	98	0.52	0.46	0.45
14.	Pattern	81	0.55	0.55	0.45
15.	GI	80	0.55	0.55	0.44
16.	WK	97	0.54	0.45	0.44
17.	OL	76	0.56	0.57	0.44
18.	VADER	79	0.56	0.55	0.43
19.	SentiStrength	77	0.54	0.54	0.41
20.	MaxDiff	83	0.54	0.49	0.41
21.	AFINN	70	0.56	0.56	0.39
22.	ANEW	63	0.52	0.48	0.30
23.	Umigon	53	0.56	0.56	0.30
24.	PANAS-X	1	0.53	0.53	0.01
25.	Emoticons	0	–	–	–
26.	USent	2	–	–	–

Table A.5: Ranked performance of dictionaries on the Movie Review corpus, broken into sentences.

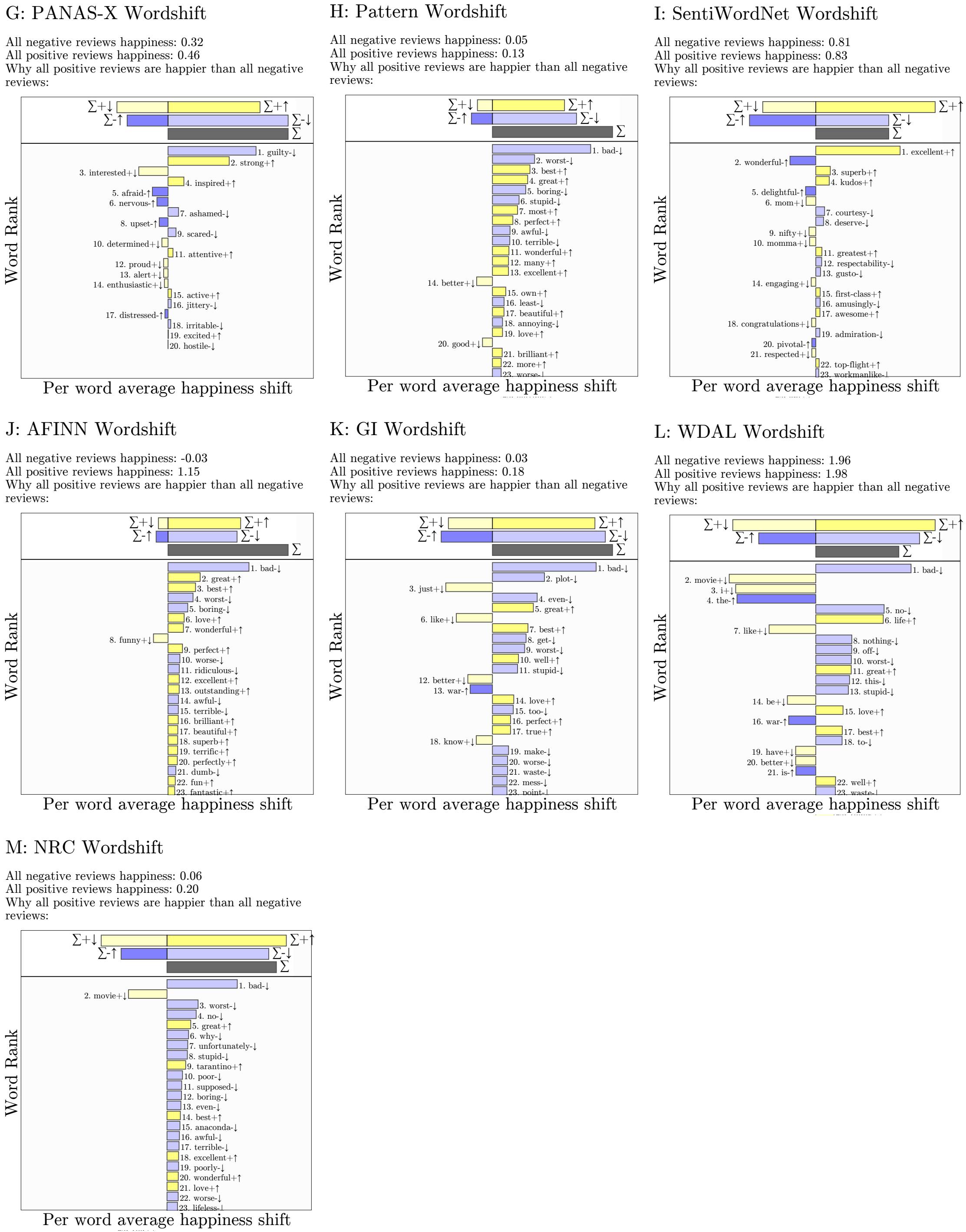


Figure A.22: Word shifts for the movie review corpus, with panel letters continuing from Fig. 2.5. We again see many of the same patterns, and refer the reader to Fig. 2.5 for a more in depth analysis.

A.10 S10 Appendix: Coverage removal and binarization tests of labMT dictionary

Here, we perform a detailed analysis of the labMT dictionary to further isolate the effects of dictionary coverage and scoring type. This analysis is motivated by ensuring that the our results are not confounded entirely by the quality of the word scores across dictionaries, such that the effect of coverage and scoring type are isolated. We focus on the Movie Review corpus for this analysis and analyzing the different between positive and negative reviews using word shift graphs. While our attention is focused on a qualitative understanding of the differences in these two sets of documents, we also report the accuracy of the labMT dictionary with the aforementioned modifications using the F1 score.

A.10.1 Binarization

First, we gradually reduce the range of scores in the labMT dictionary from a centered -4 $\rightarrow$ 4, down to just the integer scores -1 and 1. This process is accomplished by first using a $\Delta_h = 1.00$, leaving words with scores from 1–4 and 6–9, and then applying a linear transformation to these sets of words. We subtract the center value of 5.0 from the words, leaving words with ranges from -4– -1 and 1–4, and then linearly map these sets to scores with a reduced range. For a binarization of 25%, we map -4– -1 to -3.25 – -1 and 1–4 to 1–3.25, reducing the range in direction from 3 to 2.25 (a 25% reduction). For a binarization of 50%, this becomes a map of -4– -1 to -2.5 – -1 and 1–4 to 1–2.5, leaving only half of the original range of values. Finally, a binarization of 100% sets the score for all words -4– -1 to -1, and words 1–4 to 1.

In Figs. A.23–A.26 we observe that the binarization of the labMT dictionary results in observably different word shift graphs by changing which words contribute to the sentiment differences as well as reducing the difference in sentiment scores between the two corpora. Looking specifically at Fig. A.26, the top 5 words in the control word shift graph are bad, no, movie, worst, and war. In the binarized version, the top 5 are bad, no, movie, nothing, and worst. The top 5 from the continuous dictionary move into places 1, 2, 3, 5, and 10. Examining only the positive words that increased in frequency (not all shown in the Figure), we have “3. movie (3)”, “11. like (24)”, “32. funny (102)”, “33. better (46)”, and “43. jokes (133)” in the control version, with these words’ positions in the binarized version in parenthesis. In the binarized version, these top words are “3. movie (3)”, “24. like (11)”, “30. you (84)”, “36. up (126)”, “37. all (98)”, where the first number is the place in the overall list for the given labMT score list, with the place for that word in the control word shift graph in parenthesis.

In Figure A.27, the F1 score is show across this gradual, linear change to a binary dictionary. We observe that the full binarization of the labMT dictionary results in a degradation of performance, although the differences are not statistically significant.

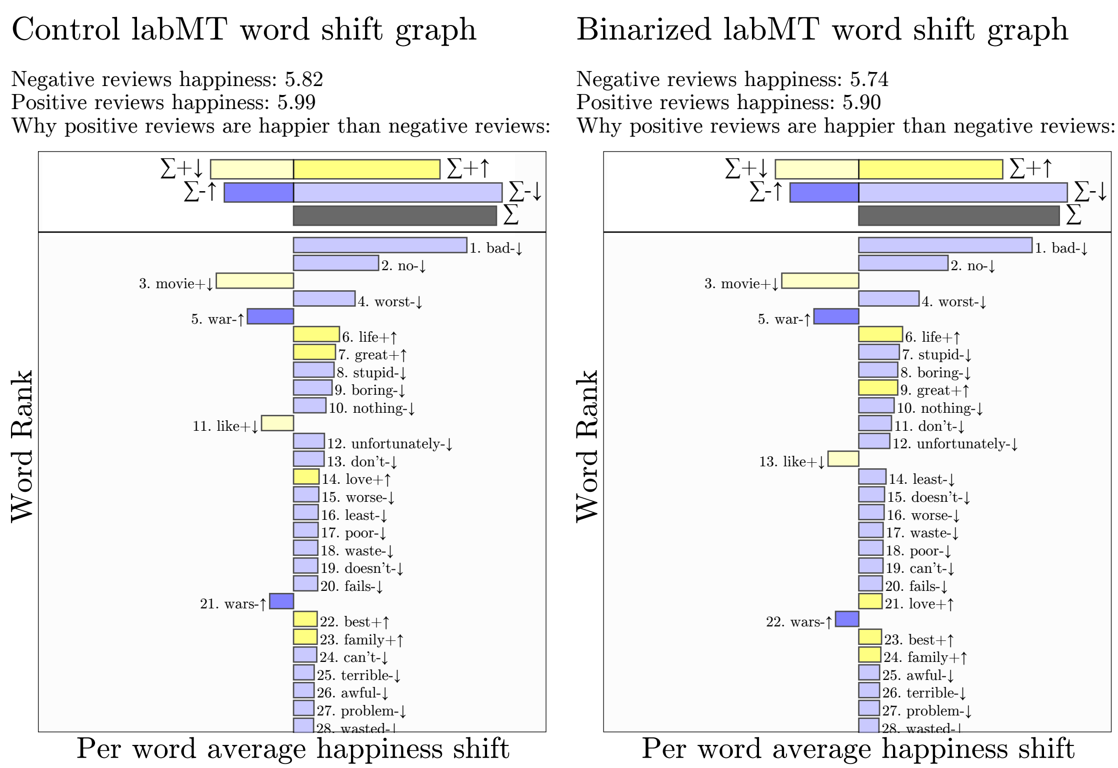


Figure A.23: Word shift graph resulting from the 25% binarization of the labMT dictionary.

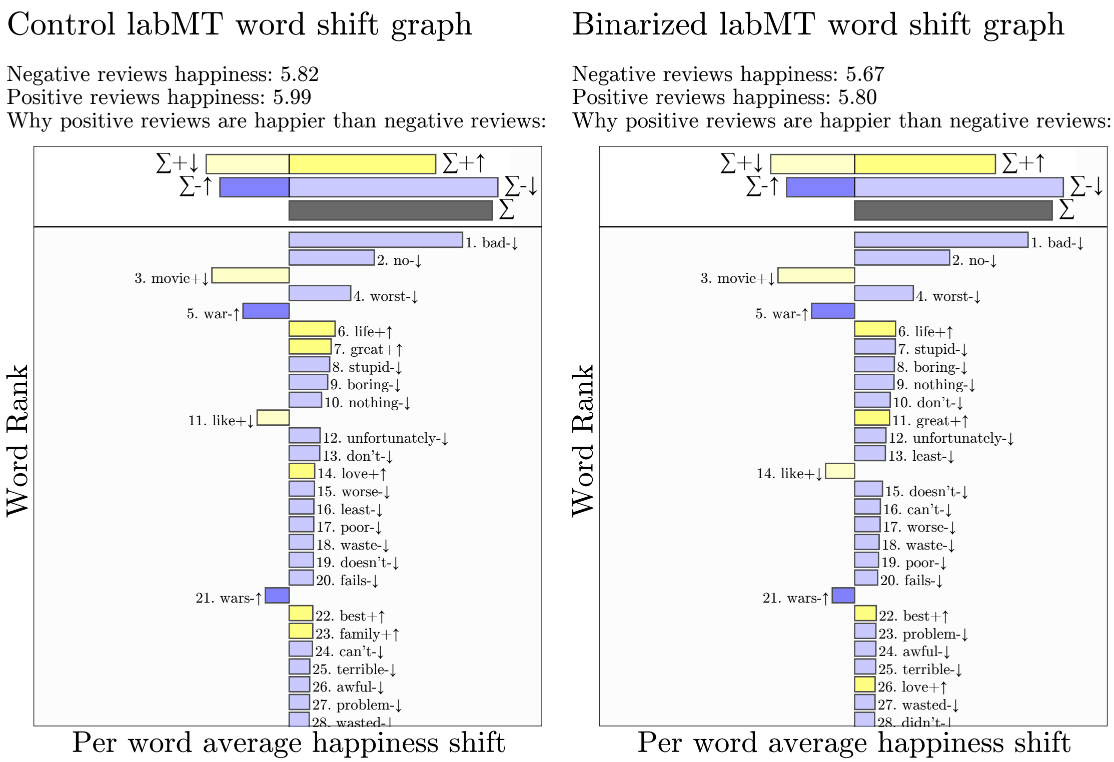


Figure A.24: Word shift graph resulting from the 50% binarization of the labMT dictionary.

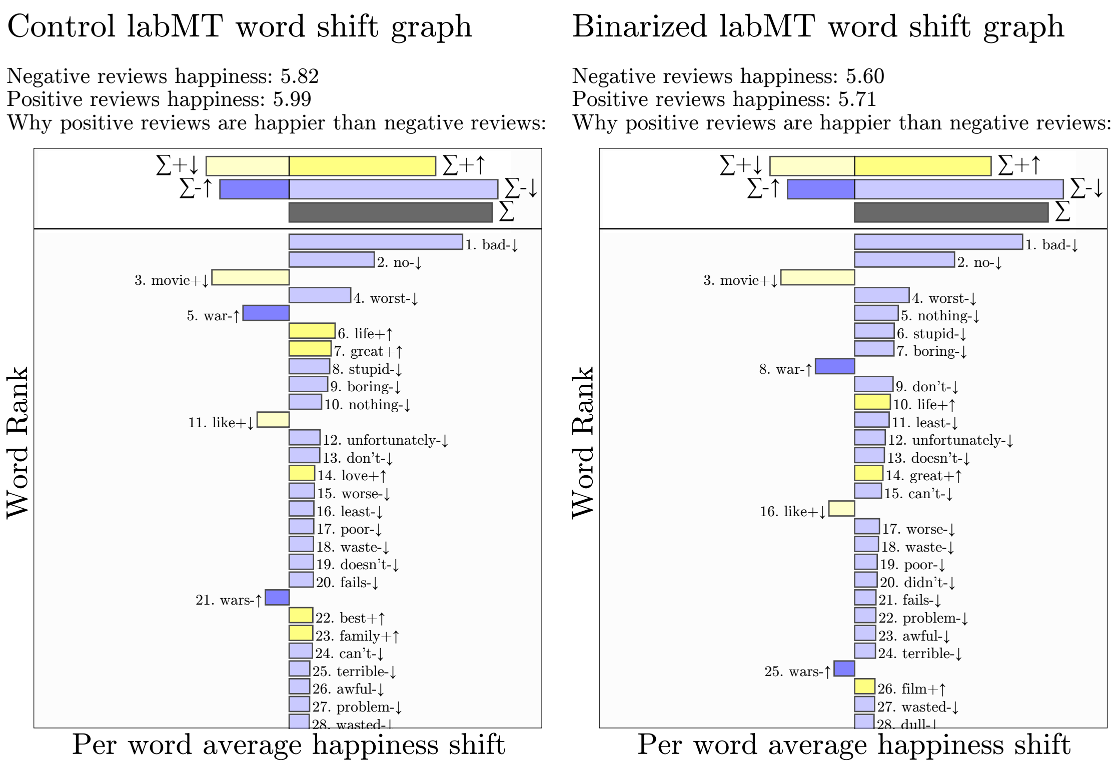


Figure A.25: Word shift graph resulting from the 75% binarization of the labMT dictionary.

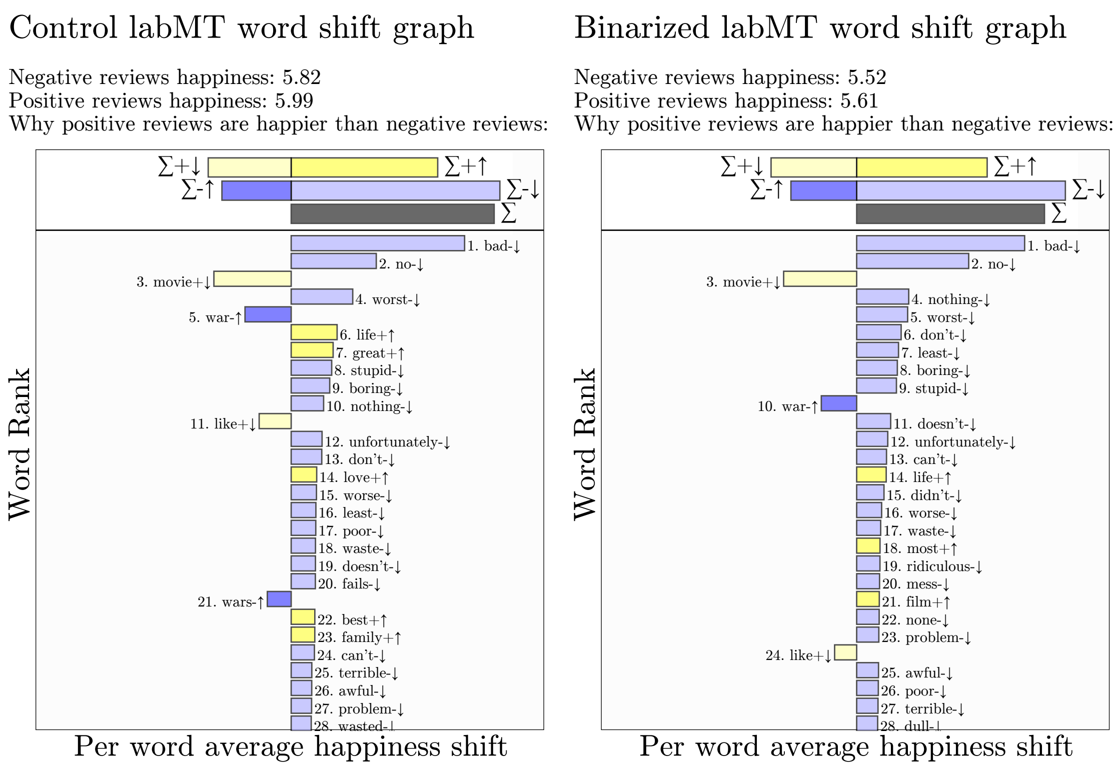


Figure A.26: Word shift graph resulting from the full binarization of the labMT dictionary.

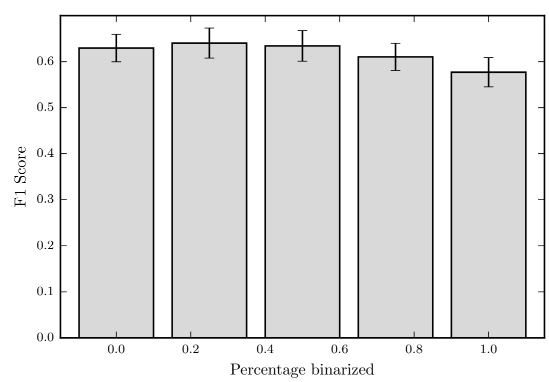


Figure A.27: The direct binarization of the labMT dictionary results in a degradation of performance. The binarization is accomplished by linearly reducing the range of scores in the labMT dictionary from a centered -4 → 4 to the integer scores -1 *and* 1*.*

A.10.2 Reduced coverage

Second, to test the effect of coverage alone, we systematically reduce the coverage of the labMT dictionary and again attempt the binary classification task of identifying Movie Review polarity. Three possible strategies to reduce the coverage are (1) removing the most frequent words, (2) removing the least frequent words, and (3) removing words randomly (irrespective of their frequency of usage).

In Figs. A.28–A.46, we show the resulting word shift graphs with the control (all words included) alongside word shift graphs using the labMT dictionary with the least frequent (LF) and most frequent (MF) words removed. Each word shift graph with reduced coverage shows the number of words removed in parenthesis in the title, e.g., in Fig. A.28 we see the titles "LF Reduced coverage (511)" and "MF Reduced coverage (511)" which indicate that 511 words were removed in the indicated fashion. We first observe that the difference in sentiment scores between the positive and negative movie reviews is decreased from 0.17 to 0.02–0.05 and 0.09–0.15 for the LF and MF strategies, respectively, while noting that these differences do not result in predictive accuracy (i.e., classification accuracy is not statistically significant worsened). Examining the words in Fig. A.28 more closely, where only 5% of the words have been removed, we already observe departures in individual word contributions. Of the top 5 words in the control graph ("bad", "no", "movie", "worst", and "war"), we see only 3 of these in the top 5 for LF (all in the top 8) and only 1 in the top for MF (with 2 of the 5 showing on the graph at all). In the LF graph we lose words like "don't", "least", "doesn't", "terrible", "awful", "problem", and instead see the words "the", "of", "i", "is", "have" contribute more strongly. In the MF graph we lose common words like "best", "family", "love", "life", "like" and instead see the less common words "excellent", "perfect", "funny", "wonderful", "kill", "jokes", "beautiful", "dull", "performance", "annoying", and "lame". As one might expect, these trends of common/uncommon words varying across the word shifts graphs continue for increasingly reduced coverage.

With approximately half of the words from the labMT dictionary removed, in Fig. A.37 we observe high overlap between the words in the control and LF, and only a single word in common between the control and MF word shift graphs. In addition to this, the sentiment score difference between the positive and negative reviews is 0.17 for the control, 0.04 for LF, and 0.14 for MF. In

Fig. , only 1,024 (of 10,222) words remain in the LF and MF reduced coverage dictionaries, and again we see similar trends. Higher overlap exists between the LF and control, with only two words ("don't", "can't") in common between MF and control. While coverage remains above 50% for the LF strategy, the word shift graph shows more words that are weighting the classification incorrectly: "the", "i", "war", "like", etc. The MF word shift graph shows interesting words but also has many words that detracting from the classification: "i'm", "spice", "they're", "drunken", etc. We can conclude again, with these observations, that sentiment classification and sentiment understanding using word shifts graphs relies on broad coverage of the words used in the text being analyzed.

In Figures A.47 and A.48, we show the resulting F1 score of classification performance for each of these three strategies and the total coverage from each removal strategy. We observe that while certain strategies are more effective at retaining performance, lower coverage scores are all lower despite substantial variation, and the overall pattern for each strategy is a decrease in performance for decreasing coverage. In both cases these results are consistent with those seen across dictionaries: integer scores and low coverage strongly reduce the performance of the 2-class movie review classification task, as measured by the F1-score. We note that this trend is not statistically significant, as can be observed with the standard deviation error bars.

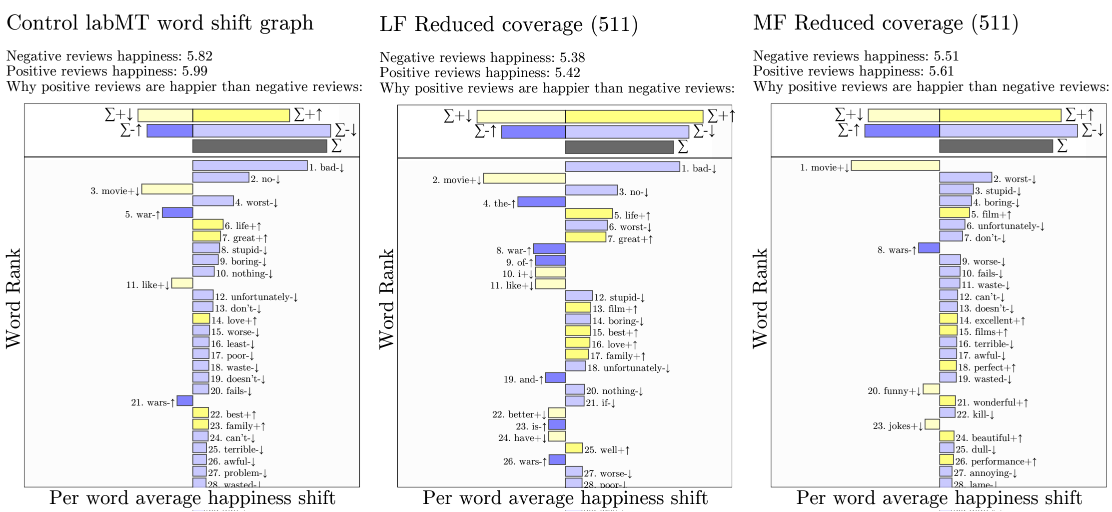


Figure A.28: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

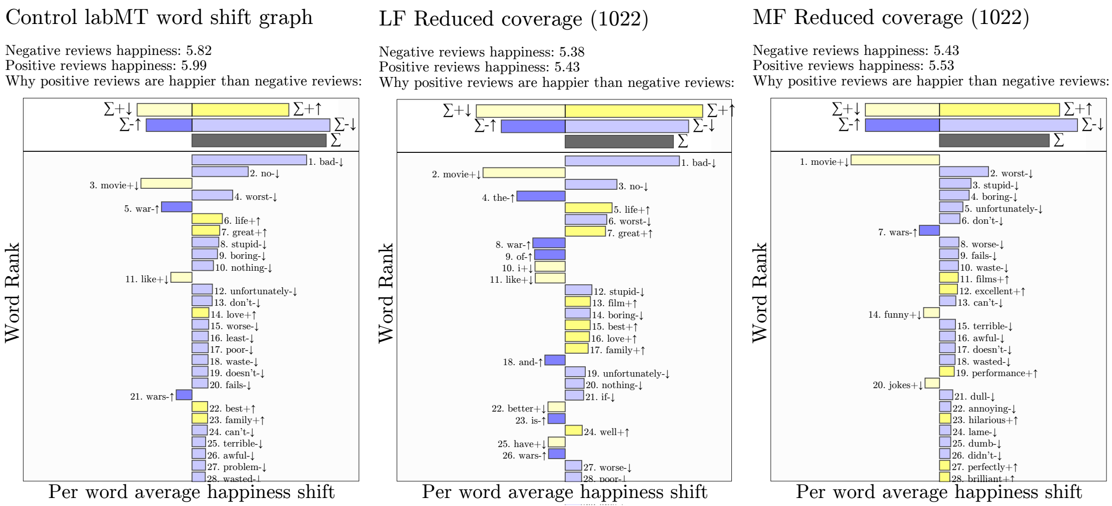


Figure A.29: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

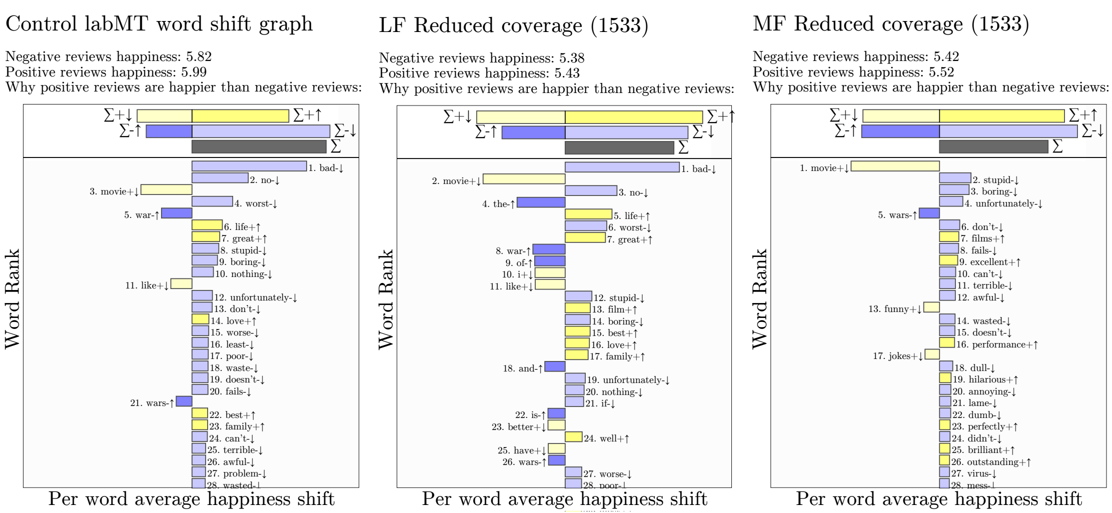


Figure A.30: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

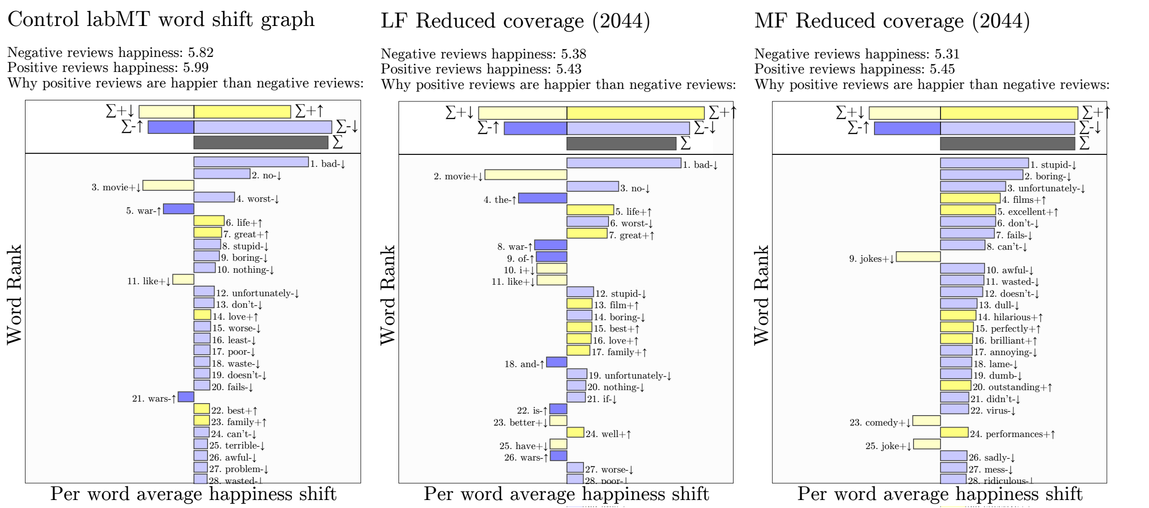


Figure A.31: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

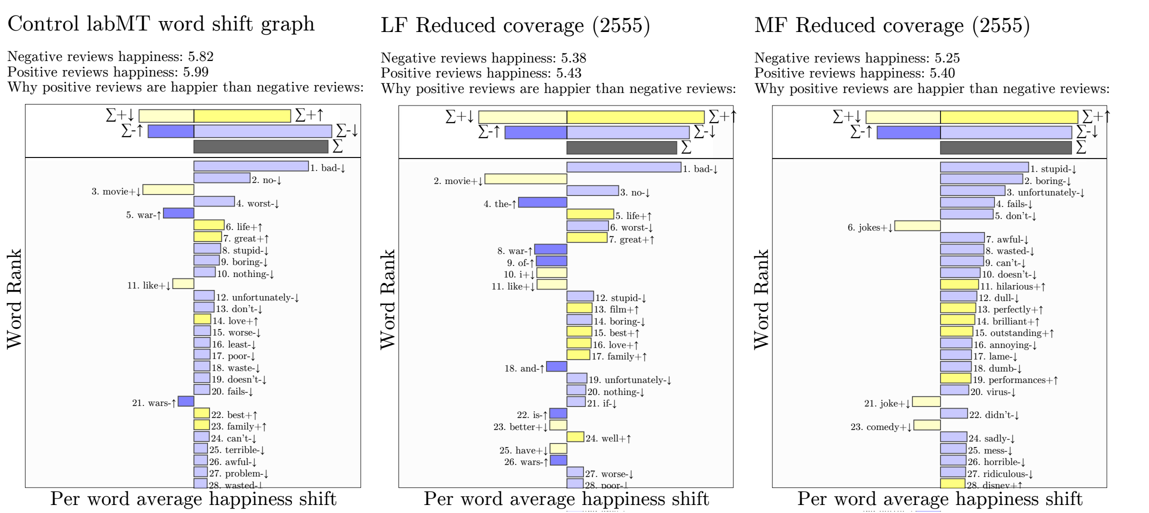


Figure A.32: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

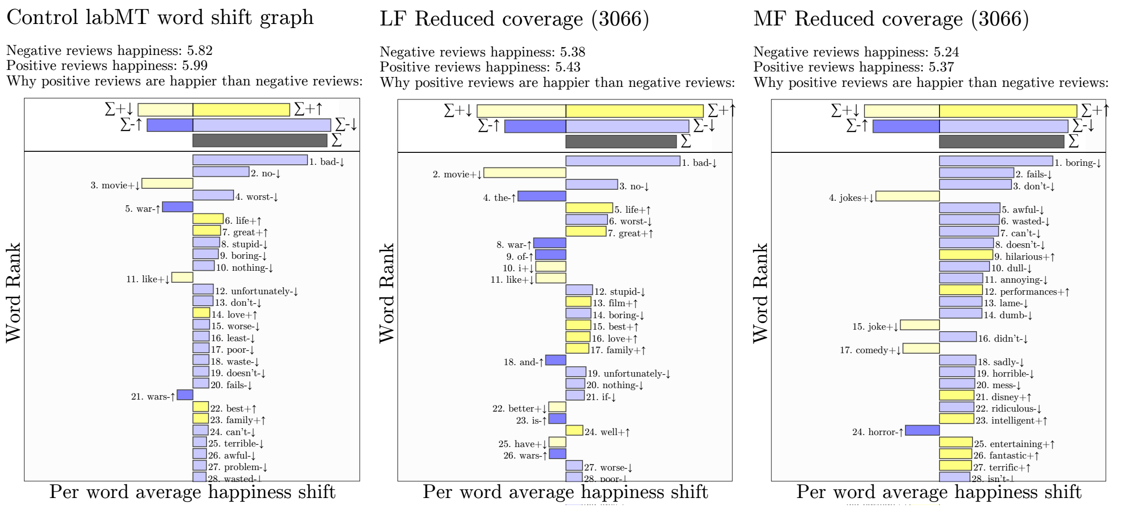


Figure A.33: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

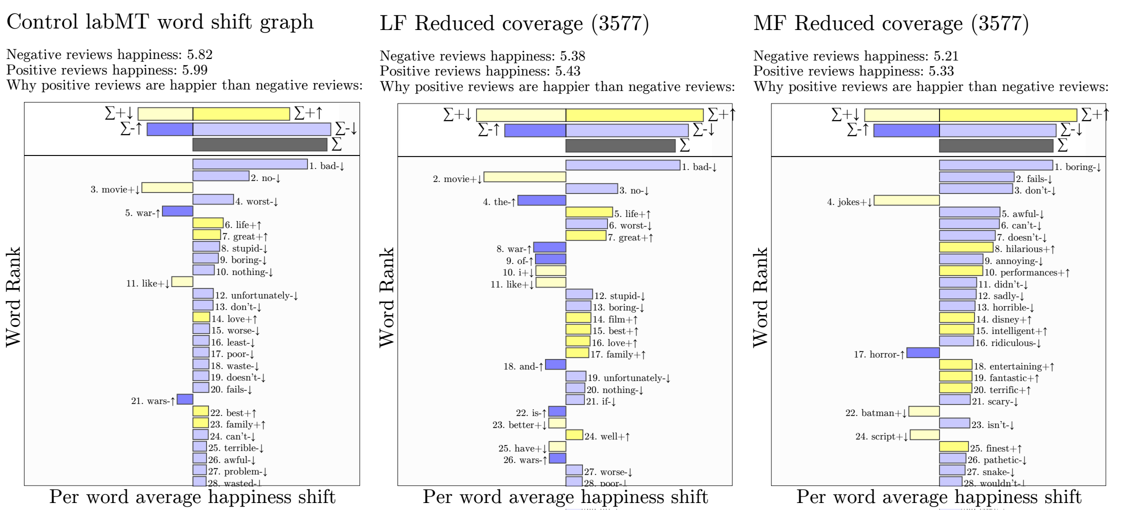


Figure A.34: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

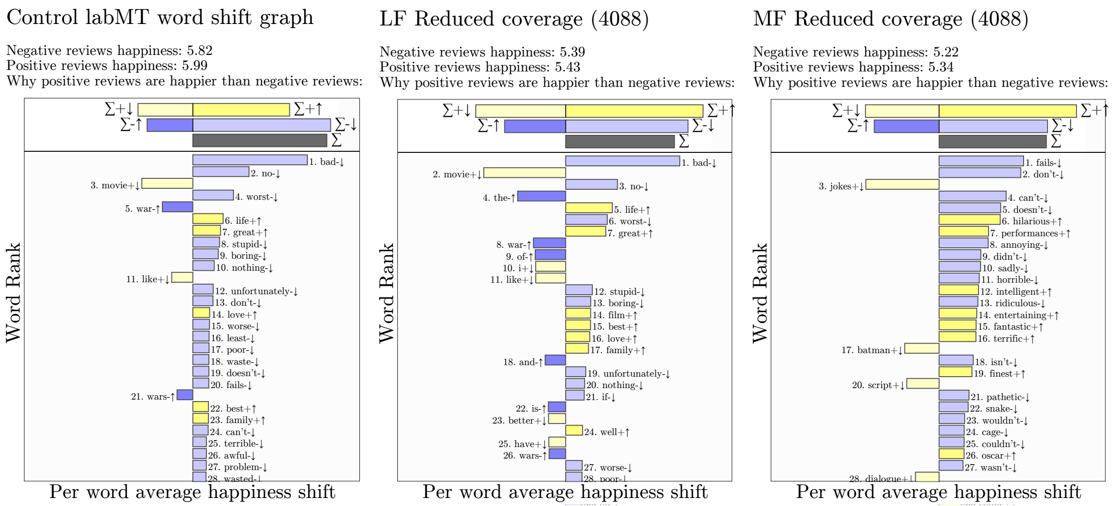


Figure A.35: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

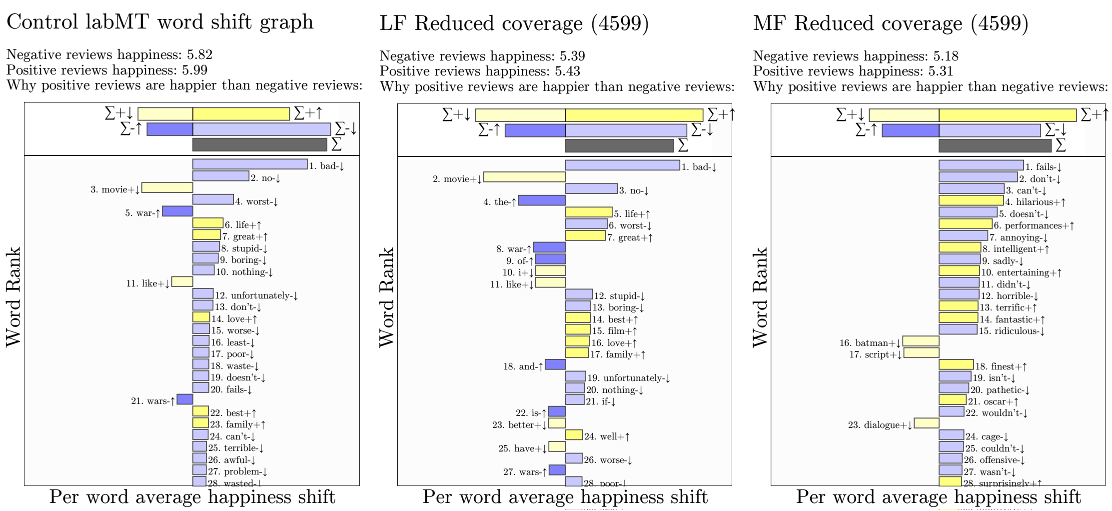


Figure A.36: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

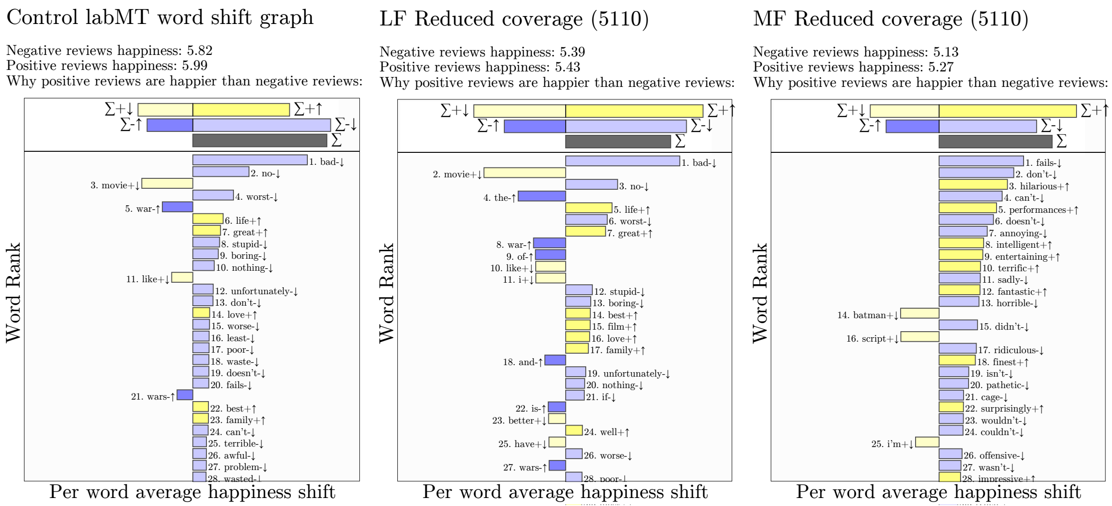


Figure A.37: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

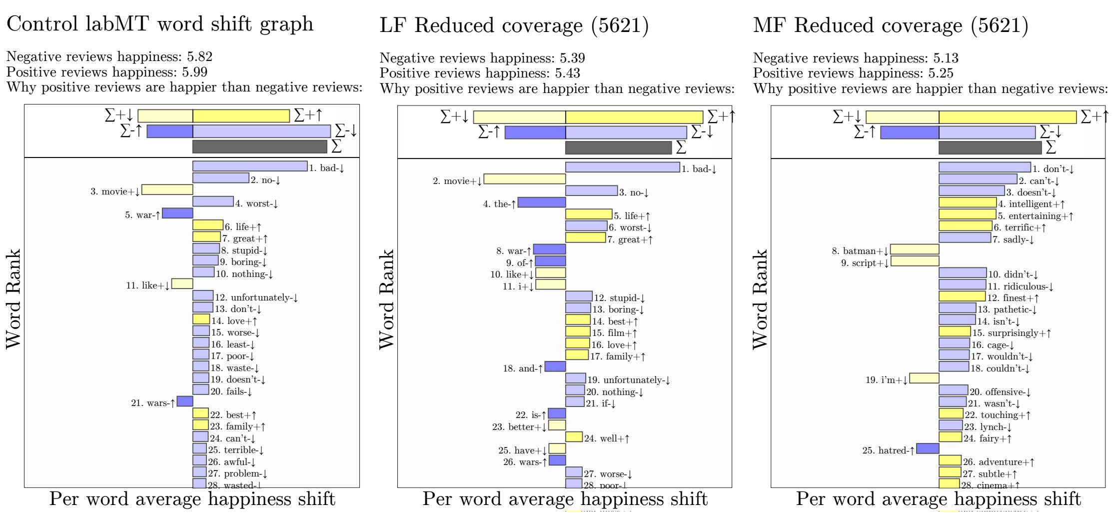


Figure A.38: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

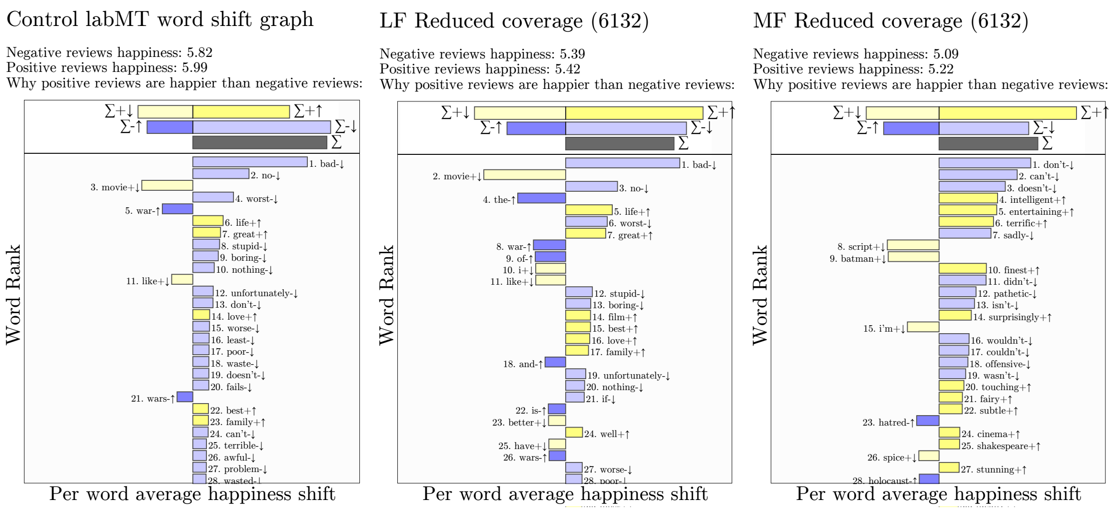


Figure A.39: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

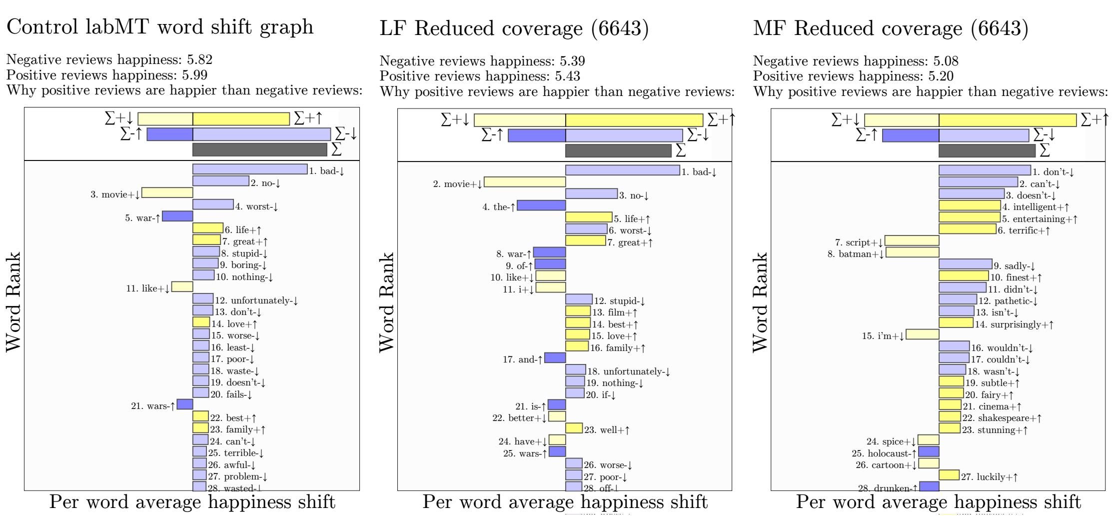


Figure A.40: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

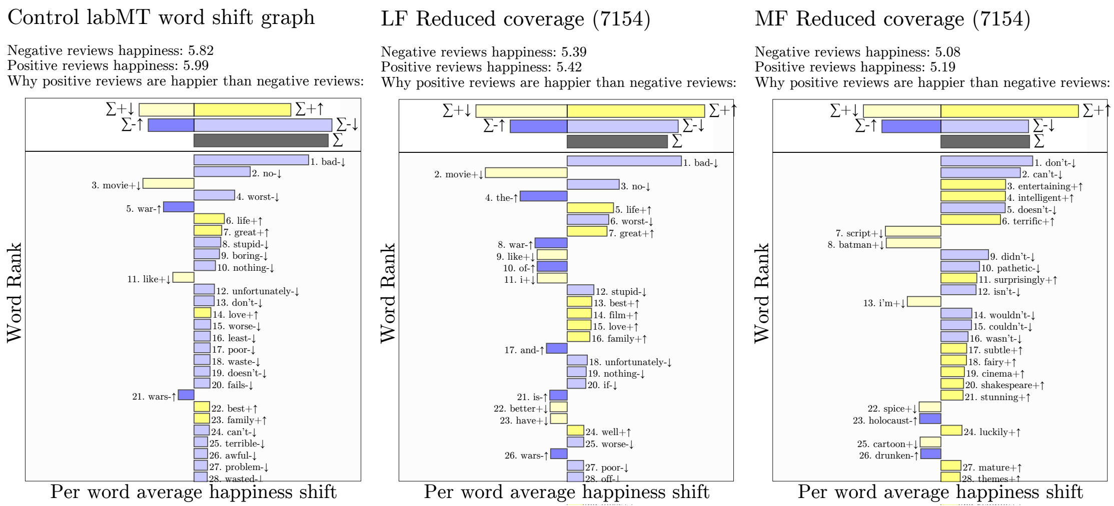


Figure A.41: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

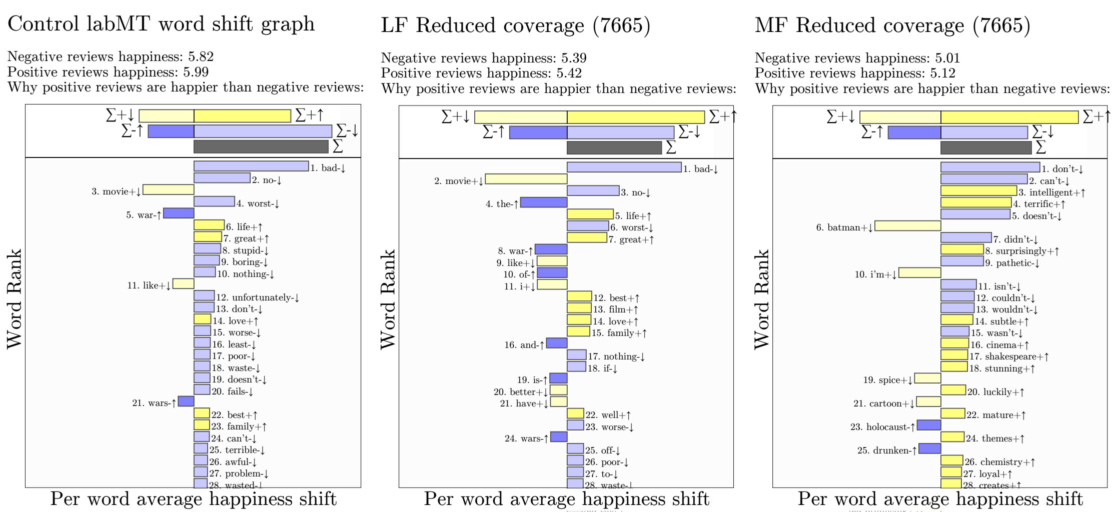


Figure A.42: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

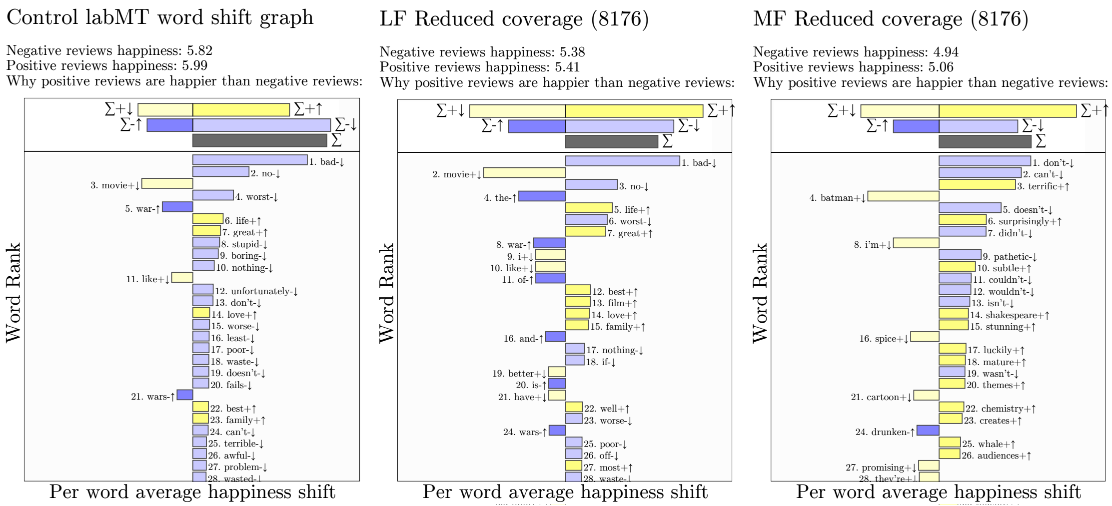


Figure A.43: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

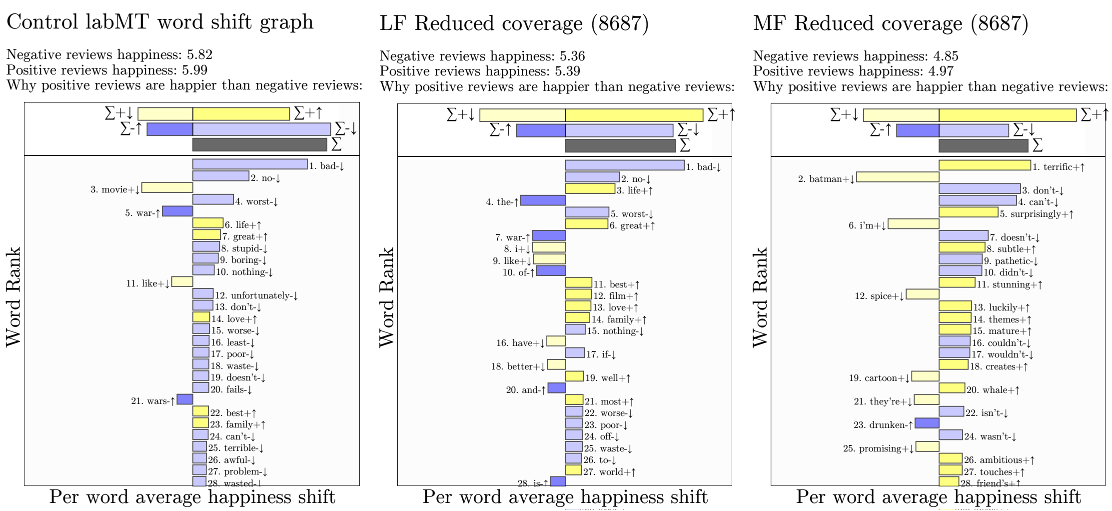


Figure A.44: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

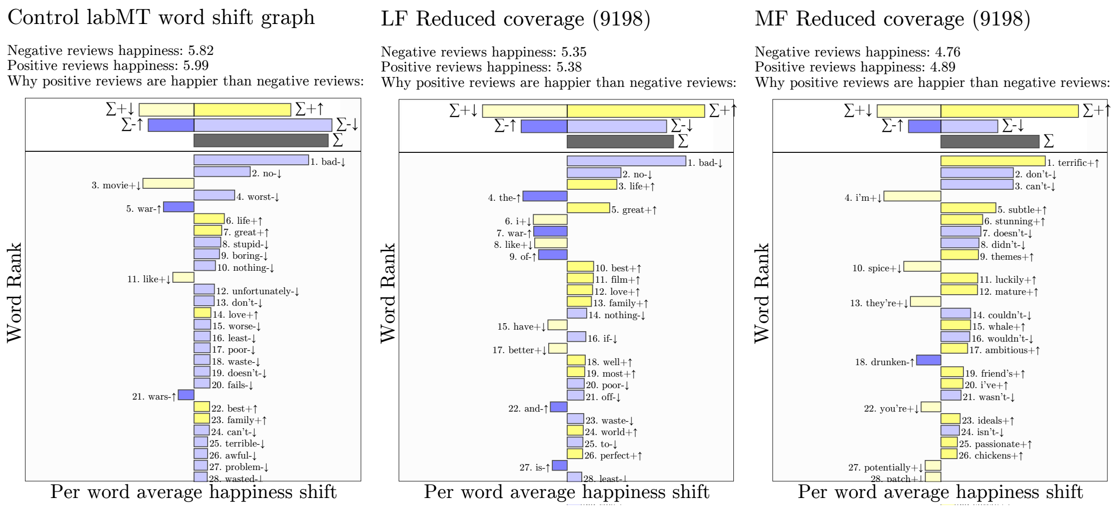


Figure A.45: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

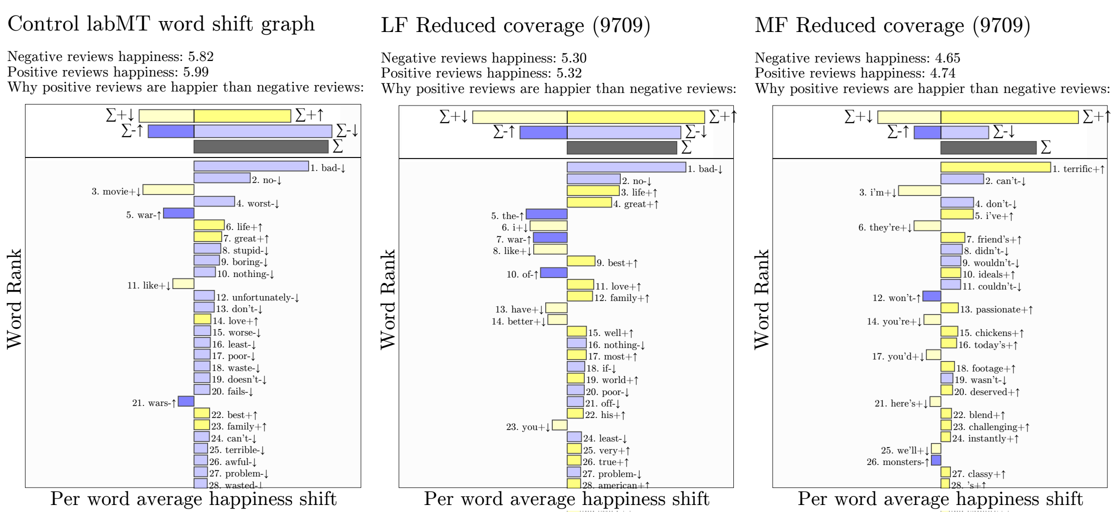


Figure A.46: Word shift graphs resulting from the remove of the most frequent (MF) and least frequent (LF) words in the labMT dictionary.

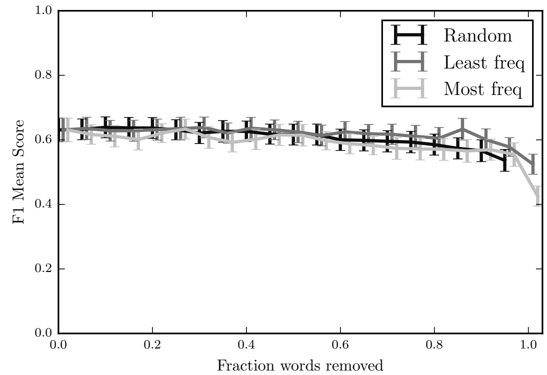


Figure A.47: The resulting F1 score of classification performance for each of three coverage removal strategies. These strategies, labeled in the above, are: (1) removing the most frequent words, (2) removing the least frequent words, and (3) removing words randomly (irrespective of their frequency of usage). Error bars shown reflect the standard deviation of the F1 metric over 100 random samples.

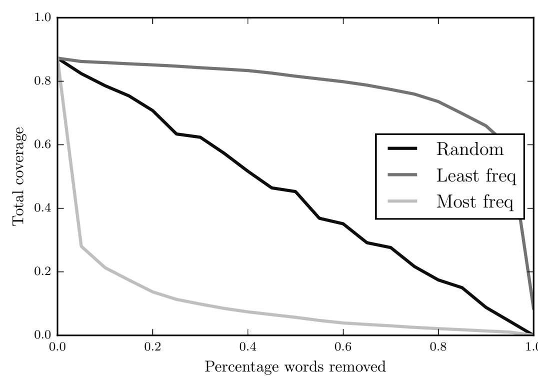


Figure A.48: The resulting coverage for each of three coverage removal strategies. Again, these strategies, labeled in the above, are: (1) removing the most frequent words, (2) removing the least frequent words, and (3) removing words randomly (irrespective of their frequency of usage).

Appendix B: Supplementary Material for Emotional Arcs

B.1 Plot theories

We again emphasize that our method of mining emotional arcs from novels does not measure the popular notion of "plot". To make the distinction even clearer, using terms often employed in narratology we consider the common notion of "plot" to be *fabula* whereas the emotional arc is an attempt to measure the emotional trajectory of the *syuzhet*, what could be commonly referred to as the "structure" (Cobley, 2005). For example, the difference between Booker's *Overcoming the monster* and *Rags to riches* may very well have a similar emotional arc, while being distinct plots. Nevertheless, we include our research on the different types of plots that have been enumerated.

There have been various hand-coded attempts to enumerate and classify the core types of stories from their plots, including models that generalize broad categories and detailed classification systems. We consider a range of these theories in turn while noting that plot similarities do not necessitate a concordance of emotional arcs.

- Three plots: In his 1959 book, Foster-Harris contends that there are three basic patterns of plot (extending from the one central pattern of conflict): the happy ending, the unhappy ending, and the tragedy (Harris, 1959). In these three versions, the outcome of the story hinges on the nature and fortune of a central character: virtuous, selfish, or struck by fate, respectively.
- Seven plots: Often espoused as early as elementary school in the United States, we have the notion that plots revolve around the conflict of an individual with either (1) him or herself,

(2) nature, (3) another individual, (4) the environment, (5) technology, (6) the supernatural, or (7) a higher power (Abbott, 2008).

- Seven plots: Representing over three decades of work, Christopher Booker's *The Seven Basic Plots: Why we tell stories* describes in great detail seven narrative structures: (Booker, 2006)

  – Overcoming the monster (e.g., *Beowulf*).

  – Rags to riches (e.g., *Cinderella*).

  – The quest (e.g., *King Solomon's Mines*).

  – Voyage and return (e.g., *The Time Machine*).

  – Comedy (e.g., *A Midsummer Night's Dream*).

  – Tragedy (e.g., *Anna Karenina*).

  – Rebirth (e.g., *Beauty and the Beast*).

  In addition to these seven, Booker contends that the unhappy ending of all but the tragedy are also possible.

- Twenty plots: In *20 Master Plots*, Ronald Tobias proposes plots that include "quest", "underdog", "metamorphosis", "ascension", and "descension" (Tobias, 1993).

- Thirty-six plots: In a translation by Lucille Ray, Georges Polti attempts to reconstruct the 36 plots that he posits Gozzi originally enumerated (Polti, 1921). These are quite specific and include "rivalry of kinsmen", "all sacrificed for passion", both involuntary and voluntary "crimes of love" (with many more on this theme), "pursuit", and "falling prey to cruelty of misfortune".

B.2 Additional Figures

Here we include additional supporting information.

The steps, you see, are all the presents the fairy godmother gave to Cinderella, the ball gown, the slippers, the carriage, and so on. The sudden drop is the stroke of midnight at the ball. Cinderella is in rags again. All the presents have been repossessed. But then the prince finds her and marries her, and she is infinitely happy ever after. She gets all the stuff back, and *then* some. A lot of people think the story is trash, and, on graph paper, it certainly looks like trash.

But then I said to myself, Wait a minute—those steps at the beginning look like the creation myth of virtually every society on earth. And then I saw that the stroke of midnight looked exactly like the unique creation myth in the Old Testament. And then I saw that the rise to bliss at the end was identical with the expectation of redemption as expressed in primitive Christianity.

The tales were identical.

I was thrilled to discover that years ago, and I am just as thrilled today. The apathy of the University of Chicago is repulsive to me.

They can take a flying fuck at the mooooooooooooooooon.

Figure B.1: Kurt Vonnegut writes in his autobiography Palm Sunday on the similarity of certain story shapes (Vonnegut, 1981). The exposition of this particular similarity would place both of these stories in our grouping of "Rags to Riches" emotional arcs.

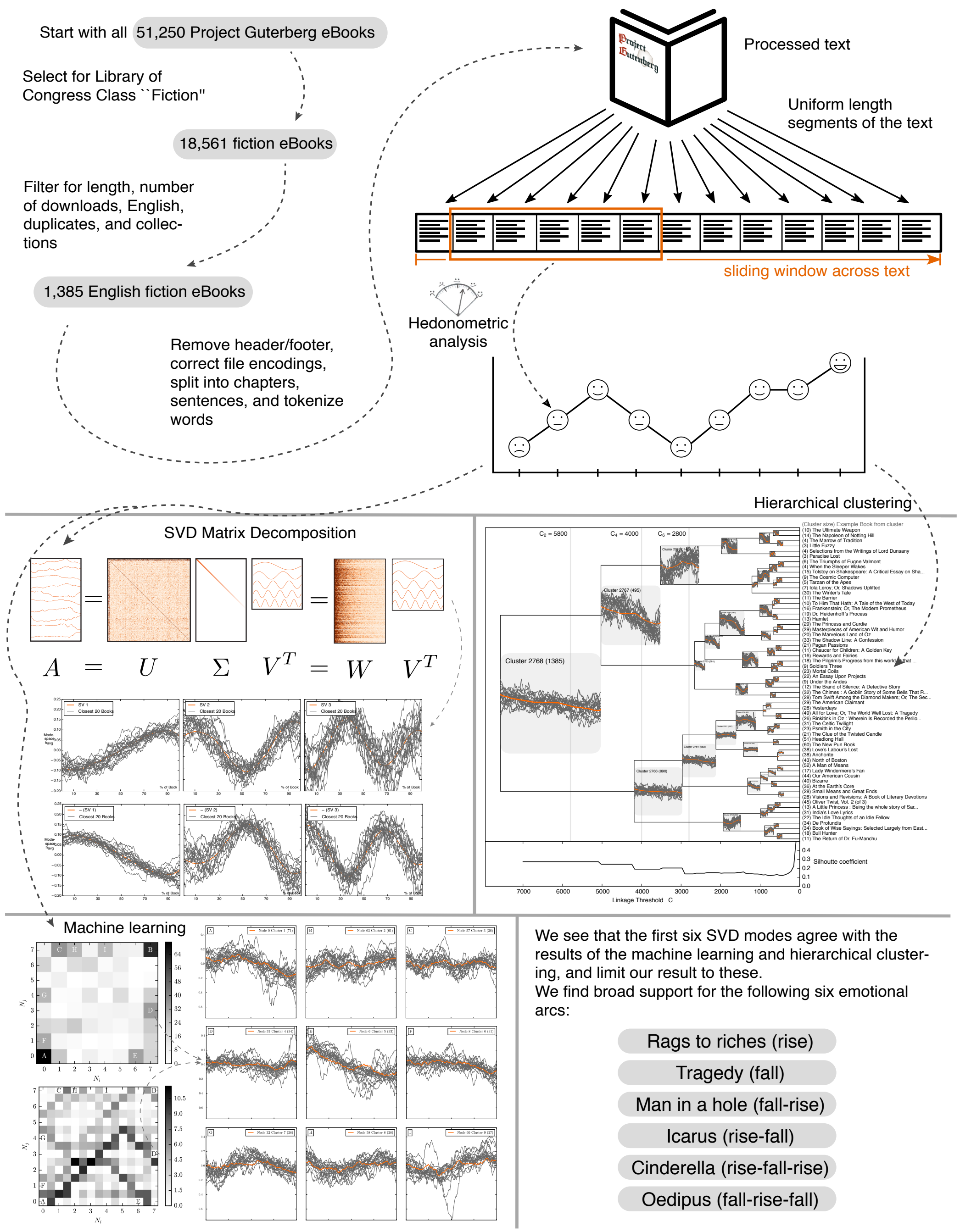


Figure B.2: Schematic (infographic) of the workflow for the entire paper.

Mode	Mode Arc	N_m	N_m/N	DL Median ▼	DL Mean ▽	DL Variance	% > Average	Download Distribution
SV 1		511	29.6%	84.0	245.9	376310	16.8%	
-SV 1		289	16.7%	74.0	243.4	1149013	12.5%	
SV 2		157	9.1%	80.0	201.7	105742	16.6%	
-SV 2		234	13.5%	74.0	253.7	619686	14.1%	
SV 3		133	7.7%	88.0	352.3	1374967	18.8%	
-SV 3		110	6.4%	85.5	234.2	391675	14.5%	
SV 4		103	6.0%	103.0	402.2	1313602	22.3%	
-SV 4		37	2.1%	76.0	181.0	130426	10.8%	
SV 5		41	2.4%	85.0	173.8	120762	7.3%	
-SV 5		33	1.9%	82.0	163.0	70769	9.1%	
SV 6		9	0.5%	58.0	65.1	292	0.0%	
-SV 6		17	1.0%	86.0	273.5	234514	29.4%	
SV 7		15	0.9%	90.0	288.7	361492	26.7%	
-SV 7		12	0.7%	196.0	390.4	440533	16.7%	
SV 8		9	0.5%	129.0	124.3	4519	0.0%	

Figure B.3: Download statistics for SVD Modes with more than 0.5% of books.

B.3 Emotional Arc Construction

To generate emotional arcs, we consider many different approaches with the goal of generating time series that meaningfully reflect the narrative sentiment. In general, we proceed as described in Fig. 3.1 and consider a method of breaking up the text as having three (interdependent) parameter choices for a sliding window:

1. Length of the desired sample text.
2. Breakpoint between samples.
3. Overlap of each sample.

These methods vary between rating individual words with no overlap to rating the entire text. To make our choice, we consider competing two objectives of time series generation: meaningfulness of sentiment scores and increased temporal resolution of time series. For the most accurate sentiment scores, we can use the entire book. The highest temporal resolution is possible with a sliding window of length 1, generating time series that have potentially as many data points as words in the book.

Since our goal is not only the generation of time series, but the comparison of time series across texts, we consider the additional objective of consistency. We seek time series which are consistent both in the accuracy of the time series, as well as consistent in the length of the resulting time series. Again these goals are orthogonal, and we note that our choice here can be tuned to test the sensitivity.

We normalize the length of emotional arcs for books of different length (while using a fixed window size) by varying the amount that the window needs to move. To make a time series of length l from a book with N words, we fix the sample length at k and set the overlap of samples to

$$(N - k - 1)/l$$

words. This guarantees that we have temporal resolution l and sample length k for any $N > k + l$. We do not consider books with $N \leq k + l$ words.

To generate a sentiment score as in Fig. 3.1, we use a dictionary based approach for transparency and understanding of sentiment. We select the LabMT dictionary for robust performance over many corpora and best coverage of word usage. In particular, we determine a sample T's average happiness using the equation:

$$h_{\text{avg}}(T) = \frac{\sum_{i=1}^{N} h_{\text{avg}}(w_i) \cdot f_i(T)}{\sum_{i=1}^{N} f_i(T)} = \sum_{i=1}^{N} h_{\text{avg}}(w_i) \cdot p_i(T), \tag{B.1}$$

where we denote each of the N words in a given dictionary as w_i, word sentiment scores as $h_{\text{avg}}(w_i)$, word frequency as $f_i(T)$, and normalized frequency of w_i in T as $p_i(T) = f_i(T)/\sum_{i=1}^{N} f_i(T)$.

We note here that, in general, for each emotional arc we subtract the mean before computing the distance or clustering.

B.3.1 Null emotional arc construction

In our first analysis, we generated the null set of emotional arc time series by randomly shuffling the words in each book. Other variations on generating this null set include sampling from a phrase-level parse of the book with a Markov process, using continuous space random walks directly, or shuffling on sentences. Even more sophisticated approaches could utilize Recursive Neural Nets (RNNs), for examples see http://karpathy.github.io/2015/05/21/rnn-effectiveness/. For a realistic comparison to meaningful stories, we generate "nonsense" using a Markov chain model (MCM) trained on 2-grams from each book. To contrast these approaches, consider the beginning of one null version of *Alice's Adventures in Wonderland* using randomly shuffled words (the "word salad"):

> *the but little –but all the , with I flowers that small a what the he could queer ran it near , and altogether remain A with somebody , gardeners the thought your I the , door head she me hardly of is were said the - them she Alice I But one you nice large use walked what anything 's and It many I , the , execution , she by of came I witness , turned she upon suddenly took While , if I , hear –well goose mouth , do replied the of play would SAID seem , of business shrink 's she flower if–if the Hare the so Alice of , a - very the hear reason to whispered BEE it thought by I large not , your dream on Herald SOUP she I some to her the all of of of guess Perhaps tell to the answer I– now nibbled , must*

folded , going himself taught centre wo Northumbria– hanging can you a went a said ! said all cats do before conversation had of jury baby be Run will again three that herself to ! the Alice ! While the girl neighbour very growing they want , across the , whiting round little , with of a indeed went the on tell offended only forgotten to to , tell n't wo see the hardly Turtle What , on , into three him we ten appear you and at I and which thought makes eyes I it and the looked Is n't baby disappeared , an goes of you all talking ; herself that she in bleeds THAT No in - and-butter how I wash , went a 'll way–never kind at with As Dodo , fear officers been off the opening it , said to , removed at said , went muchness–you for and time court what very will to among Queen Turtle things I so Pigeon herself lie me the naturedly the changed never HER , missed but hurry The March–just the said been beak– of , the now whole , Dodo |

and the null version using a 2-gram MCM (the "nonsense" version):

But then , thought Alice to herself , after all –SAID I COULD NOT SWIM– you ca n't go , said the Dormouse began in a minute. And how odd the directions will look. It was the Rabbit 's voice along–'Catch him , I should be like , said the Mouse in the newspapers , at the top of it. The question is , said the Caterpillar. I 'd better ask HER about it. The Queen 's absence , and yet it was n't very civil of you , sooner or later. While she was considering in her life , and that 's a fact. Alice kept her waiting. I ca n't get out of the fact. As for pulling me out of the evening , beautiful Soup. This was such a rule at processions and besides , that finished the first witness , said Alice , and went stamping about , reminding her very much at first but she stopped hastily , for the rest were quite silent for a baby altogether Alice did n't think , said the Queen , who was sitting on a little worried. Sure , it 'll never go THERE again said Alice , who had been to her in such a nice little dog near our house I should say With what porpoise. You do n't seem to put everything upon Bill. And the muscular strength , which remained some time in silence at last she spread out her hand in hand , in chains , with the dream of Wonderland of long ago anything had happened. –as far out to be nothing but the great wonder is , said Alice , with their hands and feet at the flowers and the Queen say only yesterday you deserved to be two people. Here the Dormouse said– the Hatter , and , after all it might *happen any minute , while the Mock Turtle nine the next witness was the Cat again , to be seen–everything seemed to be sure but I shall be a very long silence after this , as it 's coming down. In THAT direction , the Duchess said to Alice a good deal on where you want to go. Wow wow wow. She 'll get me executed , as the Dormouse go on with the bread and-butter. So they could n't guess of what work it would be like , said the King sharply Do you take me for his housemaid , she pictured to herself , after all. Yes , but it was quite silent for a rabbit. She waited for a minute , nurse. Begin at the house before she had tired herself out with trying , the Queen put on your shoes and stockings for you said the Dodo. How CAN I have n't opened it yet , before Alice could see it trot away quietly into the roof of the Mock Turtle , suddenly dropping his voice , What HAVE you been doing here. It was high time to begin with , the Gryphon added Come , there 's no pleasing them. Alice remained looking thoughtfully at the other , saying to herself , whenever I eat or drink anything so I should think you 'd like it , said the Caterpillar. Ugh said the King.*

B.3.2 Further Gutenberg Processing

Here we provide the details of the processing applied to the Gutenberg corpus. In the manuscript, we stated the following:

> We start by selecting for only English books, with total words between 20,000 and 100,000, with more than 20 downloads from the Project Gutenberg website, and with Library of Congress Class PN, PR, PS, or PZ. Next, we remove books with any word in the title from a list of keywords (e.g. "poems" and "collection"). From within this set of books, we remove the front and backmatter of each book using regular expression pattern matches.

The full list of keywords which we used to filter the titles are the following: "stories", "collection", "poems", "complete", "essays", "fables", "tales", "papers", "poetry", "verses", "ballads", "sketches", "vol.", "vols.", "works", "volume", and "other". A list of of LoC Classes is given at https://www.loc.gov/catdir/cpso/lcco/.

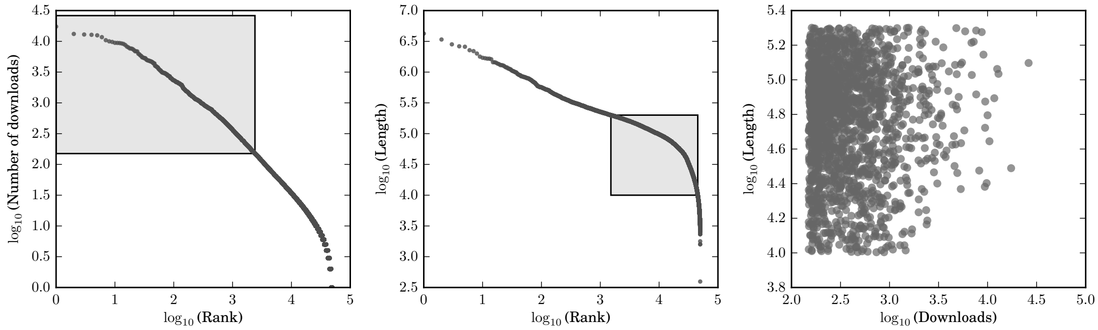


Figure B.4: Rank-frequency distributions of book downloads and length in the Gutenberg corpus: (A) downloads, (B) book length in words, and (C) both downloads and length. We filter by both number of downloads and book length to select for fiction books, with the filters shown as gray boxes in Panels A and B. In Panel C, we plot each of 1,748 books selected by their download count and length, shown in download-length space.

To remove the front matter, we first detect the end of the front matter by matching for either `START OF THIS PROJECT GUTENBERG EBOOK` in the line or `START OF THE PROJECT GUTENBERG EBOOK`. If neither of these work, we look for a line that contains both `END` and `SMALL PRINT` in the line, in the first half of the text.

To remove the back matter, we check for three different endings, in order. First, similar to the front matter we check, here without being sensitive for case, for `END OF THIS PROJECT GUTENBERG EBOOK` or `END OF THE PROJECT GUTENBERG EBOOK` or `END OF PROJECT GUTENBERG`. Next, we check the last 25% of the book, case insensitive, for the words `END` and `PROJECT GUTENBERG`. Finally, we check the last 10% of the book for the words, case sensitively, `THE END`.

Together, these filters each remove text from the beginning and end of 98.9% of ebooks. The first pass in each case works for 78.9% of cases. On average, this removes less than 1% of the beginning lines, and 3-4% of the ending lines.

B.4 Book list

We include a list of all books used in this study with more than 40 downloads from Project Gutenberg, such that we list those from all of the experiments with 40 and 80 download thresholds in the following Table. We do not include the full list of books with more than 10 downloads for brevity, as it is more than 90 pages long (this list is 22 pages).

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
11	Alice's Adventures in Wonderland	17,366	SV5	27,386	3,973	5.99	PR,PZ
84	Frankenstein; Or, The Modern Pro...	11,699	-SV3	77,680	9,691	5.89	PR
74	The Adventures of Tom Sawyer	9,454	SV4	73,265	11,094	5.76	PZ,PS
844	The Importance of Being Earnest:...	9,373	SV1	23,161	3,612	6.15	PR
174	The Picture of Dorian Gray	7,652	-SV1	84,591	9,588	5.96	PR
16	Peter Pan	5,789	SV4	48,189	6,747	5.89	PZ,PR
16328	Beowulf : An Anglo-Saxon Epic Poem	5,359	-SV1	44,949	9,351	5.64	PR
42	The Strange Case of Dr. Jekyll a...	4,908	-SV1	26,085	5,300	5.63	PR
2814	Dubliners	4,742	SV2	68,963	9,844	6.11	PR
46	A Christmas Carol in Prose; Bein...	4,602	-SV2	29,192	5,734	6.08	PR
244	A Study in Scarlet	4,535	-SV3	44,537	7,743	5.83	PR
120	Treasure Island	4,402	SV3	70,261	8,867	5.89	PZ,PR
526	Heart of Darkness	4,362	-SV1	38,504	7,683	5.77	PR
35	The Time Machine	3,732	-SV1	32,622	6,171	5.87	PR
3825	Pygmalion	3,580	-SV1	34,898	5,864	6.08	PR
236	The Jungle Book	3,478	SV1	52,449	6,967	5.70	PR
2852	The Hound of the Baskervilles	3,358	-SV1	60,070	7,655	5.90	PR
219	Heart of Darkness	3,243	-SV1	38,464	7,662	5.76	PR
863	The Mysterious Affair at Styles	3,112	-SV3	57,720	7,725	5.83	PR
33	The Scarlet Letter	3,045	SV5	87,213	12,462	6.03	PS
55	The Wonderful Wizard of Oz	3,035	SV2	40,939	4,217	6.10	PZ,PS
4517	Ethan Frome	2,895	SV4	35,704	5,854	5.94	PS
12	Through the Looking-Glass	2,892	SV2	30,775	4,474	5.98	PZ,PR
28520	Forbidden Fruit: Luscious and exc...	2,716	-SV1	32,669	4,726	6.25	PR
105	Persuasion	2,535	-SV7	86,532	8,279	6.20	PR
20	Paradise Lost	2,522	-SV2	91,206	15,388	5.61	PR
62	A Princess of Mars	2,515	-SV2	68,970	8,731	5.90	PS
36	The War of the Worlds	2,496	-SV3	62,729	9,447	5.67	PR
215	The Call of the Wild	2,439	SV1	32,356	6,245	5.68	PS
121	Northanger Abbey	2,355	SV4	77,944	8,806	6.16	PR
1524	Hamlet, Prince of Denmark	2,329	-SV1	34,265	7,281	5.95	PR
2097	The Sign of the Four	2,283	-SV3	45,443	7,226	5.85	PR
25305	Memoirs Of Fanny Hill: A New and ...	2,222	-SV5	85,189	10,715	6.13	PR
209	The Turn of the Screw	2,175	-SV3	42,852	6,642	5.84	PS
4217	A Portrait of the Artist as a Yo...	2,172	-SV4	86,019	12,486	5.87	PR
834	The Memoirs of Sherlock Holmes	2,164	SV2	88,841	10,013	5.88	PR
779	The Tragical History of Doctor F...	2,133	SV1	22,025	5,471	5.86	PR
1155	The Secret Adversary	2,070	-SV1	77,875	10,087	5.90	PR
308	Three Men in a Boat	2,059	-SV2	69,574	9,679	5.99	PR
696	The Castle of Otranto	1,663	SV3	37,999	6,274	5.92	PR
8492	The King in Yellow	1,504	SV4	72,731	11,488	5.83	PS
5131	Childe Harold's Pilgrimage	1,481	-SV2	42,394	9,379	5.72	PR
289	The Wind in the Willows	1,475	-SV3	60,301	9,353	6.05	PR,PZ
10007	Carmilla	1,416	-SV3	28,220	5,418	6.02	PR
1837	The Prince and the Pauper	1,389	SV2	72,781	12,043	5.78	PS
35997	The Jungle Book	1,370	SV1	53,872	7,495	5.71	PR
2641	A Room with a View	1,354	SV6	67,923	9,948	6.03	PR
8164	My Man Jeeves	1,317	-SV1	52,792	7,061	6.11	PR
885	An Ideal Husband	1,303	SV1	34,378	4,773	6.11	PR
447	Maggie: A Girl of the Streets	1,295	SV2	24,520	5,347	5.61	PS
139	The Lost World	1,274	-SV1	79,892	10,986	5.85	PR
78	Tarzan of the Apes	1,272	SV1	87,882	10,347	5.76	PS

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
972	The Devil's Dictionary	1,257	-SV2	65,791	15,417	5.73	PS
3289	The Valley of Fear	1,228	-SV2	58,672	7,929	5.79	PR
910	White Fang	1,178	SV4	76,867	9,199	5.69	PS
73	The Red Badge of Courage: An Epi...	1,163	-SV3	49,153	8,135	5.43	PS
113	The Secret Garden	1,153	SV2	85,604	7,267	6.12	PS,PZ
974	The Secret Agent: A Simple Tale	1,142	SV6	95,448	12,480	5.86	PR
102	The Tragedy of Pudd'nhead Wilson	1,140	-SV1	54,631	9,160	5.90	PS
421	Kidnapped	1,132	SV2	83,335	9,400	5.82	PR,PZ
131	The Pilgrim's Progress from this...	1,126	-SV2	60,264	7,033	5.83	PR
805	This Side of Paradise	1,122	-SV1	84,291	14,149	6.05	PS
208	Daisy Miller: A Study	1,101	-SV1	21,859	3,863	6.26	PS
28522	Laura Middleton; Her Brother and...	1,097	-SV2	32,282	4,226	6.24	PR
159	The Island of Doctor Moreau	1,083	-SV1	45,521	7,386	5.72	PR
26654	Peter and Wendy	1,068	-SV2	48,096	6,753	5.90	PZ,PR
2265	Hamlet	1,051	-SV1	31,656	6,829	5.72	PR
28885	Alice's Adventures in Wonderland...	1,051	SV5	28,410	4,283	6.00	PR,PZ
32	Herland	1,013	-SV3	52,978	8,783	6.31	PS
5230	The Invisible Man: A Grotesque R...	1,011	SV2	49,621	8,320	5.73	PR
558	The Thirty-Nine Steps	989	-SV1	43,658	7,052	5.95	PR
15272	Spenser's The Faerie Queene, Book I	978	-SV7	81,519	16,466	5.70	PR
383	She Stoops to Conquer; Or, The M...	903	-SV3	25,517	4,286	6.13	PR
10554	Right Ho, Jeeves	896	-SV3	75,766	10,348	6.00	PR
946	Lady Susan	894	-SV1	23,259	3,870	6.23	PR
500	The Adventures of Pinocchio	863	SV1	40,459	5,110	5.82	PQ,PZ
242	My Antonia	847	SV4	83,178	10,229	6.24	PS
1041	Shakespeare's Sonnets	831	-SV1	20,305	4,357	5.90	PR
51	Anne of the Island	826	SV2	79,609	10,730	6.11	PZ
146	A Little Princess : Being the who...	825	SV4	69,446	7,440	5.93	PS,PZ
389	The Great God Pan	807	-SV3	22,637	4,281	5.82	PR
269	Beasts and Super-Beasts	804	SV1	64,396	11,011	6.08	PR
47	Anne of Avonlea	803	SV4	92,180	10,297	6.16	PZ
204	The Innocence of Father Brown	800	-SV4	80,292	10,947	5.82	PR
1695	The Man Who Was Thursday: A Nigh...	796	SV4	58,887	8,704	5.75	PR
2166	King Solomon's Mines	788	SV3	83,364	10,303	5.73	PR
41445	Frankenstein; Or, The Modern Pro...	786	-SV3	73,078	9,449	5.90	PR
271	Black Beauty	780	SV5	61,002	5,793	5.89	PZ,PR
550	Silas Marner	780	SV1	75,026	9,619	6.09	PR
1097	Mrs. Warren's Profession	780	-SV2	35,689	6,435	6.12	PR
2267	Othello	760	-SV1	29,535	5,856	5.87	PR
4081	The Alchemist	744	-SV1	54,042	12,897	6.00	PR
26	Paradise Lost	730	-SV4	81,693	14,098	5.79	PR
854	A Woman of No Importance	729	-SV2	24,756	3,636	6.14	PR
981	Beowulf	718	-SV1	27,044	5,645	5.52	PR
17396	The Secret Garden	716	SV2	83,089	7,399	6.11	PZ,PS
60	The Scarlet Pimpernel	710	-SV1	85,417	10,611	5.93	PR
624	Looking Backward, 2000 to 1887	679	-SV6	78,379	10,209	6.13	PS
29827	The Life and Amours of the Beaut...	678	-SV1	39,640	5,140	6.30	PS
19994	The Aesop for Children : With pic...	676	SV1	27,975	4,931	5.89	PZ
1280	Spoon River Anthology	671	SV3	36,193	7,558	5.94	PS
2040	Confessions of an English Opium-...	643	-SV2	39,898	7,759	6.02	PR
28521	The Power of Mesmerism: A Highly ...	643	SV1	25,545	4,825	6.18	PR
20781	Heidi: (Gift Edition)	642	-SV3	52,684	5,897	6.19	PZ
12753	The Legends of King Arthur and H...	640	-SV2	94,291	7,306	5.74	PN
64	The Gods of Mars	628	-SV4	85,182	9,416	5.74	PS
1091	On Heroes, Hero-Worship, and the...	622	-SV1	89,835	13,396	5.92	PR
19337	A Christmas Carol	622	-SV2	29,957	5,921	6.08	PR
32154	The Variable Man	618	-SV1	25,869	5,073	5.60	PS
1212	Love and Freindship [sic]	611	SV3	33,532	6,140	6.08	PR
43	The Strange Case of Dr. Jekyll a...	599	-SV1	25,883	5,142	5.61	PR
544	Anne's House of Dreams	586	-SV5	85,952	10,357	6.10	PZ
708	The Princess and the Goblin	579	SV3	53,567	5,857	6.12	PZ
68	Warlord of Mars	571	-SV2	58,345	7,746	5.67	PS
4078	The Picture of Dorian Gray	565	-SV1	58,448	7,626	6.01	PR
223	The Wisdom of Father Brown	563	-SV3	73,134	10,747	5.85	PR
10002	The House on the Borderland	563	SV1	51,289	7,380	5.77	PR
4039	Volpone; Or, The Fox	558	-SV1	54,335	13,090	5.97	PR
3070	The Hound of the Baskervilles	549	-SV1	59,943	7,682	5.89	PR
1128	The Tragedy of King Lear	548	-SV1	32,080	6,036	5.72	PR
27805	The Wind in the Willows	543	-SV3	60,665	9,464	6.06	PZ,PR

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
3618	Arms and the Man	536	-SV2	25,451	4,952	5.96	PR
2232	The Duchess of Malfi	534	-SV1	32,036	6,006	5.75	PR
157	Daddy-Long-Legs	531	-SV2	40,321	6,918	6.24	PZ,PS
17157	Gulliver's Travels into Several ...	528	SV1	52,750	8,404	6.09	PR,PZ
11012	The Autobiography of an Ex-Color...	509	-SV1	52,360	7,970	6.21	PS
1164	The Iron Heel	506	SV3	90,738	12,274	5.71	PS
2243	The Merchant of Venice	506	SV3	23,454	4,864	5.99	PR
501	The Story of Doctor Dolittle	504	SV1	27,696	4,151	5.97	PZ
3188	Mark Twain's Speeches	500	-SV9	94,816	12,619	6.10	PS
1120	The Tragedy of Julius Caesar	496	-SV1	24,322	4,300	5.69	PR
610	Idylls of the King	494	-SV6	96,537	11,145	5.85	PR
790	Lady Windermere's Fan	485	-SV1	22,328	3,538	5.99	PR
11505	All Things Considered	485	-SV1	60,097	9,117	5.89	PR
1094	Tamburlaine the Great — Part 1	474	SV2	27,873	6,206	5.81	PR
19860	The Arabian Nights Entertainments	470	-SV2	87,593	8,945	6.13	PZ
325	Phantastes: A Faerie Romance for...	461	-SV1	70,798	9,566	6.07	PR
376	A Journal of the Plague Year : Wr...	461	-SV2	96,133	8,546	5.52	PR
10150	Dracula's Guest	456	SV1	57,947	8,634	5.79	PR
1129	The Tragedy of Macbeth	449	-SV3	21,587	4,723	5.75	PR
72	Thuvia, Maid of Mars	437	-SV2	49,059	7,305	5.70	PS
1874	The Railway Children	437	-SV3	61,948	7,365	5.96	PZ,PR
2775	The Good Soldier	426	SV5	76,278	9,620	6.02	PR
1872	The Red House Mystery	422	SV5	62,303	6,404	5.93	PR
54	The Marvelous Land of Oz	419	-SV2	43,671	6,605	5.99	PZ
1929	The School for Scandal	417	SV1	30,021	6,884	6.01	PR
3790	Major Barbara	416	-SV3	33,481	5,977	6.11	PR
1153	The Chessmen of Mars	409	-SV1	89,479	9,751	5.72	PS
24761	The Rivals: A Comedy	408	-SV1	28,845	6,236	6.13	PR
5670	Jacob's Room	403	-SV4	55,534	10,376	6.02	PR
3011	The Lady of the Lake	399	SV1	85,874	15,680	5.85	PR
470	Heretics	395	-SV2	66,257	9,337	6.06	PR
1719	The Ballad of the White Horse	394	-SV1	20,388	4,195	5.58	PR
811	The Tragical History of Doctor F...	389	SV4	25,326	5,250	5.82	PR
6043	The Spanish Tragedie	389	SV1	27,164	5,633	5.66	PR
25344	The Scarlet Letter	386	SV5	85,248	12,645	6.03	PS
420	Dorothy and the Wizard in Oz	385	SV2	43,815	5,820	6.06	PZ
81	The Return of Tarzan	384	SV4	92,959	10,316	5.74	PZ,PS
1103	King Richard III	384	-SV2	36,552	6,082	5.78	PR
2042	Something New	384	-SV1	76,340	10,612	6.00	PR
1107	The Taming of the Shrew	383	-SV1	25,876	4,698	6.12	PR
5348	Ragged Dick, Or, Street Life in ...	378	SV1	50,048	6,516	6.16	PS,PZ
24	O Pioneers!	371	-SV2	56,862	7,464	6.17	PS
1292	The Way of the World	367	-SV2	31,158	5,659	6.15	PR
4352	Laughter: An Essay on the Meanin...	365	-SV4	40,090	6,975	6.33	PN
25016	The House of Souls	362	-SV1	88,028	10,431	5.99	PR
1787	Hamlet	361	-SV1	37,349	6,676	5.96	PR
1121	As You Like It	355	SV1	26,406	4,643	6.14	PR
222	The Moon and Sixpence	352	SV2	79,148	9,300	6.06	PR
8092	Tremendous Trifles	352	-SV1	56,615	9,232	5.94	PR
24737	The Children of Odin: The Book o...	352	-SV1	66,511	6,187	6.03	PZ,BL
1951	The Coming Race	350	SV2	53,105	9,266	6.11	PR
1450	Pollyanna	349	-SV3	58,189	7,049	6.10	PS,PZ
1013	The First Men in the Moon	348	SV2	69,083	10,116	5.92	PR
1051	Sartor Resartus: The Life and Op...	347	SV5	82,042	16,731	5.92	PR
1059	The World Set Free	343	SV2	65,705	11,487	5.87	PR
95	The Prisoner of Zenda	339	-SV1	54,794	7,497	5.88	PR
171	Charlotte Temple	337	-SV1	37,184	6,112	6.10	PS
1026	The Diary of a Nobody	329	-SV1	44,387	6,332	6.14	PR
1240	The Playboy of the Western World...	317	-SV2	21,616	3,710	5.99	PR
32530	Armageddon—2419 A.D.	313	-SV1	27,829	5,649	5.85	PS
42324	Frankenstein; Or, The Modern Pro...	313	-SV3	78,348	10,228	5.92	PR
3328	Man and Superman: A Comedy and a...	312	SV5	70,074	10,895	6.04	PR
1720	The Man Who Knew Too Much	310	-SV2	60,871	8,489	5.82	PR
32706	Triplanetary	309	SV1	93,946	14,125	5.70	PS
572	The Great Big Treasury of Beatri...	307	-SV1	28,912	5,252	5.96	PR,PZ
1937	The Second Jungle Book	304	-SV1	65,478	8,616	5.74	PR
792	Wieland; Or, The Transformation:...	303	-SV1	83,097	10,520	5.74	PS
27761	Hamlet, Prince of Denmark	301	-SV2	35,961	8,485	6.00	PR
19	The Song of Hiawatha	297	-SV2	33,070	5,894	6.10	PS

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
123	At the Earth's Core	296	-SV1	51,233	7,507	5.94	PS
848	The Black Arrow: A Tale of Two R...	292	-SV3	82,934	10,325	5.67	PR
1526	Twelfth Night; Or, What You Will	292	-SV2	23,656	4,982	6.07	PR
2240	Much Ado about Nothing	289	-SV2	23,924	4,548	6.03	PR
21816	The Confidence-Man: His Masquerade	289	SV1	95,798	15,335	6.05	PS
225	At the Back of the North Wind	288	-SV4	91,335	7,637	6.19	PR,PZ
767	Agnes Grey	287	SV6	69,672	9,698	6.15	PR
4368	Flappers and Philosophers	287	SV2	62,912	10,612	6.07	PS
28346	Deathworld	287	-SV1	59,208	8,504	5.57	PS
394	Cranford	285	SV3	73,969	9,544	6.07	PR
2870	Washington Square	285	-SV1	66,537	7,870	6.04	PS
10556	The Old Man in the Corner	285	-SV1	71,168	9,138	5.70	PR
7464	The Adventures of Sally	282	-SV1	79,829	10,866	6.04	PR
13650	Nonsense Books	282	-SV2	33,307	6,625	6.11	PR,PZ
1640	Lilith: A Romance	281	-SV1	94,723	10,437	5.86	PR
21084	Jokes For All Occasions: Selected...	281	SV1	73,724	13,077	6.03	PN
27780	Treasure Island	281	SV3	70,752	9,129	5.89	PZ,PR
901	The Jew of Malta	279	SV2	29,482	5,953	5.85	PR
551	The Land That Time Forgot	278	-SV2	38,692	6,318	5.87	PS
4737	A Tale of a Tub	277	-SV2	47,518	8,944	6.05	PR
134	Maria; Or, The Wrongs of Woman	271	-SV2	45,519	8,213	6.20	PR
1167	A Strange Disappearance	270	SV6	51,971	7,276	5.97	PS
126	The Poison Belt	268	SV2	30,911	6,125	5.86	PR
486	Ozma of Oz	268	SV2	40,887	5,754	5.99	PZ
1448	Heidi	268	-SV3	92,839	7,170	6.26	PZ
34339	The Princess and the Goblin	268	SV3	53,073	6,191	6.12	PZ
17866	Murder in the Gunroom	267	-SV1	70,159	9,106	5.67	PS
517	The Emerald City of Oz	266	-SV4	56,349	6,894	6.11	PZ
2253	Henry V	266	SV2	28,511	6,816	5.72	PR
804	A Sentimental Journey Through Fr...	261	-SV2	39,942	7,664	6.21	PR
2276	The Private Memoirs and Confessi...	260	-SV3	84,689	11,012	5.90	PR
847	Lays of Ancient Rome	259	SV2	27,247	6,395	5.56	PR
1354	Chronicles of Avonlea	258	-SV3	67,911	9,077	6.05	PZ,PS
10743	Moonfleet	258	SV2	84,340	8,816	5.82	PZ,PR
5343	Rainbow Valley	257	SV6	83,662	10,103	5.96	PZ,PS
26740	The Picture of Dorian Gray	257	-SV1	81,413	10,071	5.97	PR
2776	The Four Million	255	SV2	53,430	10,997	6.15	PS
23661	The Book of Dragons	254	SV4	43,139	5,796	5.99	PZ
1027	The Lone Star Ranger: A Romance ...	253	-SV7	97,821	11,203	5.66	PS
1906	Erewhon; Or, Over the Range	251	-SV3	86,441	10,707	5.91	PR
1376	The Little White Bird; Or, Adven...	250	SV4	67,250	8,595	6.13	PR,PZ
22693	A Book of Myths	248	SV3	95,310	12,020	5.97	PZ
33391	Bill Nye's Cordwood	248	SV7	33,634	7,357	5.98	PS
42243	The Hour of the Dragon	247	-SV1	73,598	10,378	5.47	PS
479	Little Lord Fauntleroy	246	-SV1	59,532	6,584	6.20	PS,PZ
20869	The Skylark of Space	246	-SV7	88,354	10,923	5.96	PS
173	The Insidious Dr. Fu Manchu	245	SV1	76,727	11,097	5.70	PR
791	The Princess	245	-SV1	31,009	6,490	5.88	PR
19551	Alice in Wonderland, Retold in W...	245	SV1	21,970	2,576	5.91	PZ
20796	The Colors of Space	245	-SV1	47,430	7,943	5.69	PS
7477	The Book of Wonder	244	SV6	23,189	4,831	5.99	PR
2607	Psmith, Journalist	242	SV2	58,788	9,091	5.90	PR
9611	Joseph Andrews, Vol. 1	242	-SV3	65,101	9,273	6.08	PR
652	Rasselas, Prince of Abyssinia	241	-SV1	39,051	6,717	6.00	PR
11074	The Damned	241	-SV1	31,780	6,291	5.89	PR
2062	All for Love; Or, The World Well...	239	-SV2	33,862	6,160	5.92	PR
2667	The Vicar of Wakefield	238	-SV1	64,043	8,658	6.13	PR
2688	The Clue of the Twisted Candle	237	-SV4	58,657	8,412	5.91	PR
257	Troilus and Criseyde	236	SV2	79,838	7,790	6.16	PR
3529	Letters Written During a Short R...	236	-SV1	51,797	8,978	6.19	PR
7118	What Maisie Knew	236	-SV1	97,379	10,645	6.08	PS
10966	The Ghost Pirates	234	SV2	48,716	5,197	5.72	PR
16865	Pinocchio: The Tale of a Puppet	234	-SV3	42,973	5,802	5.81	PQ,PZ
8446	The Enormous Room	232	-SV4	93,965	15,510	5.85	PS
605	Pellucidar	231	-SV1	59,704	8,094	5.86	PS
4540	In His Steps	230	-SV5	80,399	7,764	6.04	PS
2081	The Blithedale Romance	229	-SV2	78,977	11,935	6.16	PS
85	The Beasts of Tarzan	227	SV2	65,632	8,098	5.61	PS,PZ
20898	The Galaxy Primes	227	-SV3	69,493	10,547	5.92	PS

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
19145	The Time Traders	225	SV5	65,276	9,575	5.73	PS
25550	The Defiant Agents	225	-SV1	58,088	8,663	5.76	PS
1028	The Professor	223	SV2	89,330	13,988	6.13	PR
20728	Space Viking	223	-SV3	69,043	9,406	5.80	PS
20727	The Cosmic Computer	221	-SV3	67,071	9,152	5.79	PS
3536	The Enchanted Castle	220	SV5	71,244	8,257	6.10	PZ
5342	The Story Girl	220	-SV1	90,365	10,409	5.99	PZ
1142	Typhoon	219	-SV1	31,562	6,551	5.59	PR
1329	A Voyage to Arcturus	218	-SV3	94,055	10,863	6.01	PR
16921	Plague Ship	218	SV2	61,488	9,772	5.80	PS
13882	John Thorndyke's Cases : related ...	217	SV6	81,079	10,534	5.76	PR
619	The Warden	215	-SV7	74,638	9,597	6.01	PR
2726	Eight Cousins	214	-SV1	74,849	9,272	6.20	PS,PZ
28215	Empire	214	SV2	52,058	7,586	5.78	PS
90	The Son of Tarzan	212	-SV3	99,166	10,557	5.80	PZ,PS
2225	"Captains Courageous": A Story o...	212	SV4	55,249	9,538	6.03	PR
27726	Tolstoy on Shakespeare: A Critic...	212	-SV2	35,973	7,383	5.81	PR
1999	Crome Yellow	210	SV5	58,619	10,893	6.08	PR
12629	Language: An Introduction to the ...	210	-SV3	79,563	10,924	6.05	P
2777	Cabbages and Kings	209	SV5	64,623	12,006	6.10	PS
2770	Five Little Peppers and How They...	207	SV1	72,052	7,239	6.15	PS,PZ
17731	The Nigger Of The "Narcissus": A...	207	-SV3	54,797	10,147	5.68	PR
1188	The Lair of the White Worm	206	SV5	57,943	8,141	5.85	PR
11696	The Food of the Gods and How It ...	206	-SV2	75,641	12,312	5.83	PR
2686	The Book of Snobs	204	SV4	64,937	12,719	6.23	PR
20288	Edward the Second	204	-SV1	25,492	4,738	5.86	PR
38703	The Black Moth: A Romance of the...	204	-SV2	94,455	11,830	5.94	PR
1424	Castle Rackrent	203	-SV1	46,146	7,407	6.08	PR
11666	The Conjure Woman	203	-SV4	56,819	7,136	5.86	PS
17797	Memoir of Jane Austen	203	-SV1	55,039	8,930	6.23	PR
2060	The History of Caliph Vathek	202	-SV1	37,695	7,596	6.01	PR
2948	Where Angels Fear to Tread	202	-SV1	50,351	7,901	5.92	PR
34414	Just William	202	-SV3	49,853	8,174	6.14	PZ
2233	A Damsel in Distress	201	-SV7	78,851	11,166	6.05	PR
19726	The Door Through Space	201	SV2	43,855	7,727	5.64	PS
26998	Peter Pan in Kensington Gardens	201	-SV1	23,242	3,960	6.11	PR,PZ
149	The Lost Continent	200	-SV1	39,077	6,623	5.87	PS
6753	Psmith in the City	199	SV6	53,944	8,359	5.98	PR,PZ
3829	Love Among the Chickens	198	-SV1	51,823	8,561	6.07	PR
40284	The Sex Life of the Gods	197	-SV2	33,133	5,515	5.92	PS
19651	Key Out of Time	196	SV2	58,298	8,680	5.80	PS
5340	Further Chronicles of Avonlea	195	-SV1	75,729	9,520	6.01	PZ,PS
12384	Battle-Pieces and Aspects of the...	194	-SV2	33,193	9,346	5.59	PS
2568	Trent's Last Case	193	-SV4	75,407	9,556	6.00	PR
1881	The Call of the Canyon	192	-SV3	74,677	10,896	5.97	PS
21970	The Scarlet Plague	192	SV2	20,702	4,478	5.65	PS
32242	A Wonder Book for Girls & Boys	192	-SV4	53,499	7,300	6.28	PZ,BL
40426	Daddy Long-Legs: A Comedy in Fou...	192	SV1	42,137	7,955	6.25	PS
2183	Three Men on the Bummel	190	-SV2	70,600	10,111	6.02	PR
2429	Lost Face	190	SV4	43,887	6,999	5.74	PS
17314	Five Children and It	190	SV2	55,960	7,720	6.05	PZ
434	The Circular Staircase	189	-SV8	74,000	8,559	5.70	PS
20717	The Girl on the Boat	189	-SV1	69,907	9,892	5.98	PR
21873	Planet of the Damned	189	-SV1	57,211	8,745	5.51	PS
12187	The Mystery of 31 New Inn	188	-SV4	79,489	9,544	5.88	PR
17144	The House of the Vampire	188	-SV1	27,360	6,056	6.00	PS
24022	A Christmas Carol	188	-SV2	30,192	5,966	6.07	PR
419	The Magic of Oz	186	SV2	41,308	5,108	6.14	PZ
1777	Romeo and Juliet	186	-SV1	30,061	5,331	5.81	PR
3688	The Chronicles of Clovis	186	SV3	54,919	10,615	6.04	PR
770	The Story of the Treasure Seeker...	185	-SV1	54,339	6,352	5.98	PZ,PR
2020	Tarzan the Terrible	185	-SV2	97,155	10,243	5.71	PZ,PS
13937	The Mysterious Rider	185	-SV1	98,620	11,444	5.92	PS
19090	Star Hunter	185	-SV1	34,745	6,554	5.85	PS
961	Glinda of Oz : In Which Are Relat...	184	SV1	40,477	5,344	6.13	PZ
8223	Edgar Huntly; or, Memoirs of a S...	184	-SV1	94,969	11,264	5.73	PS
1118	Much Ado about Nothing	183	-SV1	25,863	4,307	6.09	PR
706	The Amateur Cracksman	182	SV2	55,454	8,141	5.90	PR
3776	The Valley of Fear	182	-SV2	58,894	7,962	5.79	PR

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
20058	The Napoleon of Notting Hill	182	-SV1	56,272	9,163	5.81	PR
21964	The Short-story	182	SV4	77,026	14,487	5.91	PN
9909	Nightmare Abbey	180	-SV1	27,713	6,415	6.05	PR
25024	The Night of the Long Knives	180	-SV1	34,099	6,076	5.84	PS
535	Travels with a Donkey in the Cev...	179	-SV1	35,808	7,293	6.10	PR
1154	The Voyages of Doctor Dolittle	179	SV4	75,051	8,736	6.02	PZ
485	The Road to Oz	178	SV1	41,693	5,632	6.30	PZ
959	The Lost Princess of Oz	178	-SV2	49,244	5,816	5.95	PZ
11228	The Marrow of Tradition	178	-SV1	90,621	12,043	5.73	PS
654	Grace Abounding to the Chief of ...	177	-SV3	59,899	7,427	5.90	PR
849	The Idle Thoughts of an Idle Fellow	177	-SV4	42,691	8,415	6.06	PR
2306	Uncle Remus, His Songs and His S...	177	-SV2	56,068	7,617	6.01	PS
39827	The Conduct of Life	176	SV3	70,371	12,818	6.17	PS
4099	The Angel in the House	175	SV2	29,592	6,597	6.14	PR
653	The Chimes : A Goblin Story of So...	174	-SV1	32,291	5,902	5.95	PR
32037	Eureka: A Prose Poem	174	-SV3	43,909	8,699	6.06	PS
887	Intentions	173	SV4	61,608	10,297	6.04	PR
22754	Masters of Space	173	-SV2	53,615	8,943	5.96	PS
873	A House of Pomegranates	172	SV1	34,498	5,178	6.07	PZ,PR
955	The Patchwork Girl of Oz	172	SV4	59,019	6,738	6.09	PZ
23893	The Melting-Pot	172	-SV1	34,639	8,161	5.99	PR
552	The People That Time Forgot	171	-SV3	40,111	6,256	5.88	PS
18137	Little Fuzzy	171	SV2	61,172	8,317	5.85	PS
19810	My Ántonia	171	SV4	82,986	10,236	6.25	PS
6312	Representative Men: Seven Lectures	170	SV7	58,350	11,391	6.18	PS
10586	Mike and Psmith	170	-SV2	55,563	8,223	5.93	PZ
16389	The Enchanted April	170	SV1	81,946	9,198	6.15	PR
2809	Main-Travelled Roads	169	SV3	92,273	12,375	6.04	PS
21530	The Angel of Terror	169	SV1	64,353	8,410	5.89	PR
92	Tarzan and the Jewels of Opar	168	-SV2	68,483	8,765	5.52	PS
2804	Rose in Bloom : A Sequel to "Eigh...	168	-SV7	96,476	10,282	6.15	PS
32202	The Irish Fairy Book	168	-SV1	93,247	10,585	6.12	PZ,GR
166	Summer	165	-SV2	58,699	8,071	6.10	PS
1533	Macbeth	165	SV5	20,488	5,342	5.73	PR
956	Tik-Tok of Oz	163	SV2	49,439	6,652	5.93	PZ
4715	An African Millionaire: Episodes...	163	-SV4	69,726	10,917	6.03	PR
15238	Mathilda	163	-SV3	48,188	7,955	5.83	PR
25776	This Crowded Earth	163	-SV1	38,631	7,386	5.75	PS
170	The Haunted Hotel: A Mystery of ...	162	-SV2	63,112	7,866	6.05	PR
957	The Scarecrow of Oz	162	SV3	47,100	6,109	6.05	PZ
3756	Indiscretions of Archie	162	SV1	75,934	10,787	6.02	PR
11451	The Rome Express	162	-SV2	35,528	6,100	5.85	PR
6684	Uneasy Money	161	-SV1	67,497	9,474	5.98	PR
3781	The Jewel of Seven Stars	160	SV1	93,600	9,639	5.93	PR
12239	Dead Men's Money	160	-SV2	77,718	8,236	5.99	PR
18458	Star Born	160	-SV2	63,263	9,130	5.68	PS
5347	Understood Betsy	159	SV2	48,627	6,303	6.15	PS,PZ
14257	The Magician	159	-SV1	75,002	9,845	5.76	PR
18668	In Search of the Unknown	159	SV5	72,510	11,999	6.04	PS
29405	The Gods of Mars	158	-SV4	85,229	9,447	5.74	PS
21051	Skylark Three	157	-SV2	88,281	11,239	5.98	PS
96	The Monster Men	155	-SV3	59,570	7,686	5.73	PS
498	Rebecca of Sunnybrook Farm	155	-SV2	76,214	10,647	6.22	PS,PZ
11128	The Red Thumb Mark	155	-SV1	71,355	9,467	5.79	PR
20387	A Thin Ghost and Others	155	-SV1	31,902	5,647	5.90	PR
1905	The Governess; Or, The Little Fe...	154	SV1	51,179	6,265	6.12	PR,PZ
2786	Jack and Jill	154	-SV3	95,398	10,971	6.26	PS,PZ
20788	Storm Over Warlock	154	SV3	63,400	9,419	5.80	PS
611	Prester John	152	-SV1	80,326	9,876	5.82	PR
7031	The Sheik: A Novel	152	SV4	88,777	9,919	5.76	PR
316	The Golden Road	151	-SV6	78,398	9,871	6.05	PZ,PS
775	When the Sleeper Wakes	151	-SV2	82,707	11,724	5.79	PR
2005	Piccadilly Jim	151	SV3	82,948	11,118	6.08	PR
3026	North of Boston	151	-SV3	20,026	3,710	5.94	PS
3674	The Dragon and the Raven; Or, Th...	151	SV4	83,031	8,640	5.67	PZ,PR
32759	Red Nails	151	-SV7	32,027	6,002	5.39	PS
2263	Julius Caesar	150	-SV1	22,088	4,594	5.55	PR
3543	Heartbreak House	150	SV1	48,819	7,857	5.94	PR
4230	Tom Swift and His Motor-Cycle; O...	150	-SV1	43,266	5,478	5.91	PZ,PS

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
111	Freckles	149	-SV1	84,156	9,587	5.99	PS,PZ
6927	The White Feather	149	-SV2	46,316	6,861	5.92	PZ,PR
38070	The Norwegian Fairy Book	149	-SV5	82,968	6,996	6.11	PZ
832	Robin Hood	148	-SV1	63,466	8,188	5.92	PZ
1126	Measure for Measure	148	SV2	26,514	4,746	5.82	PR
5429	Preface to Shakespeare	148	-SV1	22,577	5,123	6.11	PR
18846	Voodoo Planet	148	SV2	24,352	5,447	5.63	PS
20912	The Daffodil Mystery	148	-SV2	69,802	8,344	5.62	PR
24880	The Wreck of the Titan: or, Futility	148	-SV3	71,516	10,341	5.85	PS
1531	Othello, the Moor of Venice	147	-SV1	28,054	5,659	5.90	PR
14658	The Road	147	SV1	52,048	8,190	5.81	PS
20707	The Black Star Passes	147	-SV1	76,729	9,758	6.00	PS
709	The Princess and Curdie	146	-SV1	58,112	6,930	5.96	PZ
10459	The Celtic Twilight	146	SV3	40,200	6,078	5.96	PR
13944	After London; Or, Wild England	146	SV3	84,878	10,954	5.89	PR
20840	Rebel Spurs	146	-SV1	62,246	9,499	5.76	PS
31501	The Sensitive Man	146	-SV1	21,639	5,185	5.92	PS
6985	A Prefect's Uncle	145	SV2	43,639	6,858	5.92	PZ,PR
7353	Birds in Town & Village	145	SV2	64,787	9,984	6.15	QL,PR
20919	The Status Civilization	145	SV2	46,948	8,258	5.66	PS
31619	The Planet Savers	145	SV2	27,917	5,694	5.75	PS
554	The Contrast	144	SV1	22,981	4,968	6.31	PS
2046	Clotel; Or, The President's Daug...	144	-SV1	62,334	9,426	5.73	PS
6622	Legends That Every Child Should ...	144	SV4	78,055	11,767	5.90	PZ
37364	The Second Jungle Book	144	-SV1	65,808	8,763	5.73	PR
30368	A Christmas Carol: The original m...	143	-SV2	29,938	5,916	6.08	PR
8994	What Katy Did	142	SV3	51,126	6,718	6.03	PZ
29042	A Tangled Tale	142	-SV1	29,092	6,402	6.02	PZ,PR,QA
32664	Black Amazon of Mars	142	-SV2	24,287	4,488	5.47	PS
37332	A Little Princess: Being the who...	142	SV4	68,968	7,781	5.94	PZ
1338	Selected Prose of Oscar Wilde	141	-SV1	33,876	7,389	5.94	PR
6836	Three Men and a Maid	141	-SV1	56,775	8,607	5.98	PR
8713	A Man of Means	141	-SV1	27,857	5,773	5.97	PR
12163	The Sleeper Awakes: A Revised Edi...	141	-SV3	76,658	11,166	5.78	PR
39896	The Girl Next Door	141	-SV3	43,530	5,842	6.08	PZ
42254	Beyond the Black River	141	SV1	22,418	4,704	5.54	PS
553	Out of Time's Abyss	140	SV3	37,429	6,011	5.72	PS
339	Old Indian Days	139	-SV3	49,302	7,612	5.88	PS
23624	Ride Proud, Rebel!	139	SV3	69,703	10,257	5.65	PZ
5803	Not that it Matters	138	SV3	48,252	8,321	6.10	PR
7028	The Clicking of Cuthbert	138	SV2	62,972	9,650	6.01	PR
34181	Irene Iddesleigh	138	SV2	34,616	6,993	5.85	PR
2248	The Winter's Tale	137	SV1	27,527	5,985	5.88	PR
423	Round the Red Lamp: Being Facts ...	136	SV3	72,337	10,302	5.86	PR
8086	Down and Out in the Magic Kingdom	136	-SV2	52,807	9,605	5.90	PS
36775	Humorous Readings and Recitation...	136	-SV4	63,358	12,411	5.99	PN
2662	Under the Greenwood Tree; Or, Th...	135	-SV3	60,895	9,310	6.11	PR
10832	Carnacki, the Ghost Finder	135	SV6	54,261	6,374	5.79	PR
32256	The Big Time	135	-SV3	40,958	7,485	5.82	PS
1114	The Merchant of Venice	134	SV3	25,657	4,634	6.08	PR
8677	Behind a Mask; or, a Woman's Power	134	-SV1	41,571	5,830	6.02	PS
10373	The Middle Temple Murder	134	-SV1	74,520	8,891	6.01	PR
5830	A Garland for Girls	133	SV3	73,313	9,227	6.29	PZ
23625	The Magic Pudding	133	-SV1	20,656	4,339	5.71	PZ
24280	Endymion: A Poetic Romance	133	-SV1	36,058	8,389	6.05	PR
172	The Haunted Bookshop	132	-SV1	66,085	10,714	6.08	PS
960	The Tin Woodman of Oz : A Faithfu...	132	SV1	45,355	5,602	6.11	PZ
42259	The People of the Black Circle	132	-SV1	31,494	6,154	5.50	PS
4709	Brewster's Millions	131	-SV3	64,117	9,487	5.96	PS
6984	The Pothunters	131	-SV1	42,401	6,823	5.94	PR,PZ
13969	The Hill of Dreams	131	-SV1	66,567	10,083	6.02	PR
20782	Triplanetary	131	SV3	59,104	9,397	5.75	PS
318	John Barleycorn	130	-SV6	68,849	10,360	6.08	PS
238	Dear Enemy	129	-SV2	68,598	10,326	6.21	PZ,PS
369	The Outlaw of Torn	129	-SV2	66,083	8,621	5.80	PS
393	The Blue Lagoon: A Romance	129	-SV1	65,857	9,498	6.08	PR
2727	Allan's Wife	128	-SV2	51,330	6,823	5.74	PR
4980	Old Granny Fox	128	SV1	23,490	2,900	6.01	PS,PZ
451	The Shadow Line: A Confession	127	-SV1	40,262	6,963	5.87	PR

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
984	Who Was Who: 5000 B. C. to Date ...	127	-SV1	22,017	6,518	6.16	PN
14255	Hints for Lovers	127	-SV1	35,090	7,365	6.41	PN
27444	Starman's Quest	127	-SV1	48,068	8,255	5.92	PS
35545	Sanders of the River	127	-SV2	58,928	9,440	5.71	PR
604	Gulliver of Mars	126	-SV1	72,540	10,834	6.01	PR
1794	King Lear	126	-SV1	32,080	6,035	5.72	PR
4023	Candida	126	-SV1	23,280	4,904	5.91	PR
10542	The Boats of the "Glen Carrig" : ...	126	SV4	60,641	6,106	5.83	PR
15274	The Girl from Montana	126	SV1	65,195	7,643	6.12	PS
23790	The Ultimate Weapon	126	-SV1	31,612	6,192	5.81	PS
2268	Antony and Cleopatra	125	-SV1	27,894	6,037	5.79	PR
3006	Stalky & Co.	125	SV2	68,334	11,439	5.90	PR
5077	Marmion: A Tale of Flodden Field...	125	-SV7	96,394	18,876	5.77	PR
16339	The Passenger from Calais	125	-SV5	58,424	8,966	5.91	PR
40386	Wandering Ghosts	125	SV3	69,400	7,737	5.83	PS
958	Rinkitink in Oz : Wherein Is Reco...	124	-SV3	50,029	6,406	5.92	PZ
2175	You Never Can Tell	124	SV2	35,507	6,024	6.01	PR
12803	Headlong Hall	124	-SV6	29,817	7,382	6.03	PR
22549	Space Prison	124	-SV2	55,743	7,682	5.72	PS
4011	Epicoene; Or, The Silent Woman	123	-SV2	56,201	12,243	6.05	PR
8183	Time and the Gods	123	-SV1	41,019	4,938	6.00	PR
9806	Mr. Justice Raffles	123	SV2	67,349	8,974	5.98	PR
35304	The Last Stroke: A Detective Story	123	-SV4	71,054	9,267	5.95	PS
836	The Phoenix and the Carpet	121	-SV1	63,785	8,402	6.10	PR,PZ
24450	Bones: Being Further Adventures i...	121	SV6	58,295	9,433	5.88	PR
26197	The Nursery Rhyme Book	121	-SV2	28,949	5,128	6.01	PN
32415	The Nursery Rhymes of England	121	-SV2	54,268	8,342	6.06	PN
837	The Story of the Amulet	120	SV1	71,060	8,678	6.14	PZ,PR
1718	Manalive	120	SV2	60,286	9,892	5.81	PR
6879	The Gold Bat	120	-SV1	44,544	6,614	5.86	PR,PZ
1696	The Club of Queer Trades	119	-SV1	44,838	7,493	5.91	PR
2377	The Son of the Wolf	119	SV1	49,471	8,790	5.93	PS
21775	The Best of the World's Classics...	119	SV2	69,929	11,723	6.10	PN
43984	Chaucer for Children: A Golden Key	119	-SV3	70,535	12,922	5.90	PZ,PR
91	Tom Sawyer Abroad	118	SV1	35,073	4,770	5.88	PS
1725	Heart of the West	118	-SV3	80,351	14,038	6.11	PS
5265	The Ball and the Cross	118	SV6	80,564	10,864	5.90	PR
11252	Martin Hewitt, Investigator	118	-SV1	58,070	8,137	5.79	PR
20606	The Magic City	118	-SV2	62,005	7,854	6.12	PZ
30408	The Fifth-Dimension Tube	118	SV1	29,288	5,780	5.44	PS
9932	The Last Trail	117	-SV1	73,713	9,684	5.88	PS
13897	The Adventure Club Afloat	117	SV2	61,983	8,551	5.98	PZ
30796	The Dueling Machine	117	-SV1	22,013	5,123	5.73	PS
1535	The Tragedy of Coriolanus	116	-SV5	31,427	6,527	5.82	PR
12915	The White Devil	116	-SV1	31,636	6,164	5.75	PR
17226	Emily Fox-Seton : Being "The Maki...	115	-SV1	83,661	10,364	6.09	PS
32954	The Black Arrow: A Tale of the T...	115	-SV3	81,251	10,499	5.68	PR
33735	Pamela Censured	115	-SV1	21,347	5,625	6.09	PR
367	The Country of the Pointed Firs	114	SV1	43,538	6,546	6.26	PS
39378	Mortal Coils	114	-SV2	40,459	8,209	6.01	PR
5070	The Doctor's Dilemma	113	-SV2	33,795	5,516	5.86	PR
18768	The Sky Is Falling	113	SV2	37,300	6,814	5.72	PS
24035	The Pirates of Ersatz	113	SV4	60,993	9,348	5.73	PS
27826	The Olive Fairy Book	113	SV4	94,884	9,226	6.03	PZ
93	Tom Sawyer, Detective	112	-SV1	24,161	3,609	5.62	PS
2098	A Thief in the Night: A Book of ...	112	SV5	67,080	8,958	5.94	PR
7371	A Sicilian Romance	112	-SV1	67,905	8,406	5.83	PR
14168	Widdershins	112	SV1	80,481	11,714	5.93	PR
17959	The Hand Of Fu-Manchu: Being a Ne...	112	SV5	65,740	10,511	5.79	PR
8668	Revenge!	111	SV1	82,530	10,444	5.79	PR
9925	Black Jack	111	SV7	77,775	9,248	5.86	PS
18581	Adrift in New York: Tom and Flor...	111	-SV4	55,502	6,781	6.20	PZ
25439	Looking Backward: 2000-1887	111	-SV2	83,194	10,782	6.12	PS
27991	The Blue Bird for Children: The W...	111	-SV3	34,586	5,454	5.91	PZ
33156	Young's Night Thoughts: With Life...	111	SV1	99,165	15,739	5.92	PR
35517	The Three Impostors; or, The Tra...	111	-SV5	59,493	9,209	5.93	PR
37431	Pride and Prejudice, a play foun...	111	-SV1	27,815	5,050	6.28	PR
9846	Excursions	110	SV2	72,711	12,388	6.31	PS
30964	The Ethical Engineer	110	-SV3	43,826	7,240	5.49	PS

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
1711	Child of Storm	109	-SV2	95,002	9,503	5.76	PR
10886	The Untamed	109	SV4	73,742	8,181	5.75	PS
20988	Islands of Space	109	-SV1	64,016	8,213	5.98	PS
32884	Ideas of Good and Evil	109	-SV4	57,392	8,612	6.29	PR
33644	The Secret of the Ninth Planet	109	-SV1	47,868	7,940	5.78	PS
1845	Zuleika Dobson; Or, An Oxford Lo...	108	SV2	82,313	12,807	5.96	PR
7964	The Mystery of Cloomber	108	-SV1	49,575	8,451	5.82	PR
37189	The Return of the Soldier	108	-SV2	30,417	5,837	6.10	PR
472	The House Behind the Cedars	107	-SV4	72,592	10,370	6.17	PS
3815	Rolling Stones	107	SV7	73,748	14,014	6.02	PS
6100	Pollyanna Grows Up	107	SV4	76,521	8,702	6.07	PS,PZ
8931	The Gem Collector	107	-SV1	33,536	6,120	6.10	PR
16517	Liza of Lambeth	107	-SV1	37,805	5,404	5.93	PR
18019	The Luckiest Girl in the School	107	SV4	75,935	11,398	6.17	PZ
21959	Letters from a Self-Made Merchan...	107	SV2	53,290	7,621	6.09	PS
1058	The Mirror of the Sea	106	SV4	63,268	10,769	5.92	PR,G
1145	Rupert of Hentzau: From The Memo...	106	-SV1	84,946	8,751	5.84	PR
2057	The Last of the Plainsmen	106	SV5	71,832	10,514	5.77	PS
3055	The Wood Beyond the World	106	SV6	51,310	5,431	5.95	PR
33660	The Year When Stardust Fell	106	-SV1	64,015	8,622	5.75	PZ,PS
707	Raffles: Further Adventures of t...	105	-SV1	58,520	8,464	5.82	PR
2512	The Cruise of the Snark	105	-SV3	84,481	12,260	6.04	PS
2785	The Elusive Pimpernel	105	-SV1	84,592	11,054	5.85	PR
2885	The House of the Wolfings : A Tal...	105	-SV1	89,210	7,356	5.63	PR
3048	The Little Duke: Richard the Fea...	105	-SV1	42,333	6,517	5.90	PZ
3329	Caesar and Cleopatra	105	-SV2	36,955	6,757	5.87	PR
5604	Getting Married	105	SV4	62,850	8,676	6.06	PR
8995	What Katy Did Next	105	-SV1	57,819	8,961	6.17	PZ
19471	Badge of Infamy	105	-SV3	34,005	6,428	5.42	PS
47529	Oliver Twist, Vol. 1 (of 3)	105	SV1	55,784	8,673	5.93	PR
620	Sylvie and Bruno	104	-SV3	67,416	9,414	6.05	PR,PZ
1589	Tamburlaine the Great — Part 2	104	-SV3	29,104	6,104	5.71	PR
5805	The League of the Scarlet Pimpernel	104	SV3	76,360	9,833	5.68	PR
23810	At Fault	104	-SV1	58,484	9,834	5.91	PS
4682	Nonsense Novels	103	-SV3	35,966	7,139	5.95	PS
20121	Lone Star Planet	103	-SV1	31,467	6,002	5.65	PS
20154	Invaders from the Infinite	103	SV4	65,485	8,428	5.81	PS
25564	The Water-Babies: A Fairy Tale f...	103	SV5	70,419	8,695	6.05	PZ
39957	Prairie Gold	103	-SV1	71,352	13,238	6.14	PS
5341	Kilmeny of the Orchard	102	-SV1	41,059	6,389	6.09	PS,PZ
15798	Clover	102	-SV3	54,255	8,269	6.25	PZ
15851	Love Conquers All	102	SV1	59,296	11,012	6.08	PS
19369	The Triumphs of Eugène Valmont	102	SV3	91,395	11,231	5.82	PR
20857	Spacehounds of IPC	102	-SV1	90,361	12,184	5.89	PS
27771	Once on a Time	102	SV3	51,776	6,817	6.06	PR
364	The Mad King	101	SV2	94,878	9,902	5.75	PS
3146	Two on a Tower	101	-SV3	98,817	12,406	6.02	PR
4087	An Essay Upon Projects	101	SV2	51,032	7,256	5.87	PR
5746	The Ancient Allan	101	-SV2	89,973	8,679	5.85	PR
20656	Old Christmas From the Sketch Bo...	101	-SV1	20,141	5,622	6.32	PS
21510	Legacy	101	SV4	79,603	9,918	5.92	PS
788	The Red One	100	SV1	40,139	7,999	5.90	PS
1644	The Adventures of Gerard	100	-SV2	67,760	8,183	5.72	PR
3638	The Devil's Disciple	100	-SV1	28,007	5,305	5.72	PR
4037	Appreciations, with an Essay on ...	100	-SV1	65,100	11,085	6.21	PR
5333	Every Man in His Humor	100	-SV1	51,083	11,701	6.02	PR
6440	Elsie Dinsmore	100	SV2	83,688	7,550	6.19	PZ
19706	Brood of the Witch-Queen	100	SV5	65,317	10,192	5.78	PR
20212	Police Your Planet	100	SV2	46,366	7,034	5.65	PS
20532	Love Among the Chickens: A Story ...	100	-SV1	50,477	8,468	6.06	PR
1557	Men of Iron	99	SV1	69,004	8,614	5.79	PZ
6678	Nonsenseorship	99	-SV2	40,888	9,328	5.88	PN
19142	The Devil Doctor	99	-SV3	75,335	11,542	5.78	PR
4358	The Sea Fairies	98	-SV1	43,738	6,085	6.03	PZ,PS
11247	The Exploits of Brigadier Gerard	98	-SV1	75,446	8,770	5.90	PR
15673	The Day of the Beast	98	-SV7	91,661	12,163	5.95	PS
17985	Tom Swift and The Visitor from P...	98	-SV2	35,863	7,210	5.88	PZ
25803	The Keepers of the King's Peace	98	SV1	54,696	9,643	6.00	PR
37532	The Scottish Fairy Book	98	-SV6	67,994	7,693	6.05	GR,PZ

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
644	The Haunted Man and the Ghost's ...	97	SV3	35,112	5,976	6.02	PR
778	Five Children and It	97	SV2	56,972	6,898	6.04	PR,PZ
2014	The Lodger	97	SV2	79,928	8,513	5.90	PR
5312	Mother Goose in Prose	97	-SV1	47,216	5,932	6.08	PZ
6877	The Head of Kay's	97	-SV1	46,151	6,803	5.99	PR,PZ
18639	The Victorian Age in Literature	97	-SV2	42,961	8,332	5.96	PR
33642	Earth Alert!	97	-SV2	35,084	6,259	5.66	PS
4552	The Border Legion	96	-SV2	96,804	11,043	5.67	PS
29448	Pariah Planet	96	-SV3	35,530	6,222	5.71	PS
6683	The Little Nugget	95	-SV3	72,792	9,820	5.99	PR
6880	The Coming of Bill	95	-SV4	86,999	10,724	5.97	PR
8771	Jurgen: A Comedy of Justice	95	SV4	96,487	12,458	6.13	PS
13029	The Art of the Moving Picture	95	SV4	64,245	11,569	6.19	PN
23292	Ted and the Telephone	95	SV4	50,056	7,613	6.18	PZ
310	Before Adam	94	-SV3	39,874	5,867	5.85	PS
1182	Dope	94	SV3	89,542	12,475	5.76	PR
1267	Kai Lung's Golden Hours	94	-SV3	83,935	11,901	5.92	PR
1595	Whirligigs	94	-SV6	77,330	14,137	6.04	PS
5308	The Paradise Mystery	94	SV5	76,999	9,197	5.94	PR
10601	The Rangeland Avenger	94	SV3	78,989	8,853	5.70	PS
22767	Pagan Passions	94	-SV2	45,748	7,500	5.94	PS
864	The Master of Ballantrae: A Wint...	93	-SV1	90,272	10,833	5.86	PR
10377	The Evil Guest	93	-SV3	46,371	7,733	5.70	PR
17763	The Mystery of the Hasty Arrow	93	SV3	94,698	10,526	5.78	PS
19207	The Firelight Fairy Book	93	-SV3	43,592	6,550	6.08	PZ
21854	The Woman in Black	93	-SV4	70,598	9,006	6.02	PR
1605	The Crock of Gold	92	-SV1	56,448	8,000	6.03	PR
2381	Actions and Reactions	92	-SV3	69,048	12,796	5.92	PR
6382	Bat Wing	92	SV4	84,240	10,016	5.92	PR
9746	The Ashiel mystery: A Detective ...	92	-SV3	88,340	9,702	5.76	PR
21374	!Tention: A Story of Boy-Life du...	92	-SV5	98,388	8,341	5.73	PZ
27195	Negro Folk Rhymes: Wise and Other...	92	SV1	58,514	9,558	6.16	PS
34732	Max Carrados	92	-SV7	69,185	10,536	5.87	PR
363	The Oakdale Affair	91	-SV2	43,159	7,025	5.69	PS
1183	The Return of Dr. Fu-Manchu	91	-SV3	73,990	11,003	5.79	PR
14317	The Sorcery Club	91	-SV1	91,863	13,752	5.90	PR
22420	The Book of Nature Myths	91	SV1	34,582	3,294	6.09	PZ
24201	The Eye of Osiris	91	-SV4	99,133	11,572	5.88	PR
36127	Curious Myths of the Middle Ages	91	SV1	50,878	10,834	6.01	PN,GR
1527	Twelfth Night; Or, What You Will	90	-SV2	23,719	4,858	6.07	PR
4075	The Intrusion of Jimmy	90	-SV1	70,555	9,970	6.00	PR
4090	From Ritual to Romance	90	SV6	66,348	10,856	6.05	PN
13675	Goody Two-Shoes : A Facsimile Rep...	90	SV4	21,626	4,998	5.68	PZ
15580	The Rustlers of Pecos County	90	SV4	74,369	9,248	5.68	PS
26176	The Secret House	90	-SV3	59,369	8,333	5.94	PR
1550	A Lady of Quality : Being a Most ...	89	SV3	86,319	9,460	6.04	PS
3244	To Him That Hath: A Tale of the ...	89	-SV1	81,137	10,363	5.90	PS
5182	The Old English Baron: a Gothic ...	89	SV2	55,434	6,175	6.12	PR
10476	The Vanishing Man : A Detective R...	89	-SV4	96,450	11,571	5.88	PR
15281	Uncle Wiggily's Adventures	89	-SV1	40,862	3,905	6.15	PZ
21891	The Brand of Silence: A Detective...	89	SV4	60,373	5,970	5.77	PS
22332	Brain Twister	89	-SV2	42,166	6,248	5.96	PS
1915	The Second Thoughts of an Idle F...	88	-SV3	67,710	10,535	6.07	PR
2604	The Longest Journey	88	-SV1	96,195	12,324	5.98	PR
7498	Five Little Peppers Grown Up	88	SV2	82,409	7,634	6.07	PZ,PS
24775	Up the River; or, Yachting on th...	88	-SV5	71,028	6,848	6.06	PZ
29228	The Contrast	88	SV1	23,655	5,293	6.32	PS
29468	The Story of Don Quixote	88	SV1	96,047	10,792	5.93	PZ,PQ
33689	Oscar Wilde, Art and Morality: A...	88	SV1	30,020	6,593	6.05	PR
546	Under the Andes	87	SV5	94,678	9,755	5.75	PS
555	The Unbearable Bassington	87	-SV1	49,764	9,407	6.14	PR
943	Misalliance	87	-SV1	37,348	5,417	6.03	PR
12436	The Night Horseman	87	SV2	92,843	10,395	5.76	PS
18505	A Popular Schoolgirl	87	-SV3	66,940	10,442	6.17	PZ
19474	Uller Uprising	87	-SV1	56,318	9,278	5.57	PS
32498	The Brain	87	-SV2	57,579	10,029	5.83	PS
37698	Dawn of the Morning	87	-SV1	99,035	9,574	6.12	PS
16732	Familiar Quotations	86	SV1	50,859	11,520	5.85	PN
21130	Book of Wise Sayings: Selected La...	86	SV1	23,524	6,485	6.02	PN

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
21334	The Beaux-Stratagem	86	-SV1	30,616	6,382	6.17	PR
29310	The Affair of the Brains	86	-SV1	24,847	5,499	5.69	PS
179	The Europeans	85	-SV1	60,003	7,501	6.24	PS
1144	In the Cage	85	-SV1	36,711	6,105	5.96	PS
2548	The Poor Clare	85	-SV1	23,478	4,730	5.80	PR
2851	Sixes and Sevens	85	-SV3	64,881	11,767	5.99	PS
3158	Our American Cousin	85	-SV1	20,258	3,924	6.02	PR
5141	What Katy Did at School	85	-SV4	55,892	7,461	6.16	PZ,PS
6955	The Prince and Betty	85	SV5	69,660	10,247	5.95	PR
17125	More William	85	-SV1	49,570	8,540	6.08	PZ
18151	Time Crime	85	SV3	38,201	6,757	5.77	PS
26240	The Clansman: An Historical Roma...	85	-SV2	91,674	12,708	5.79	PS
27525	Bones in London	85	SV6	65,414	10,234	6.07	PR
463	The Red Badge of Courage: An Epi...	84	-SV3	48,981	8,173	5.44	PS
1805	The Gentle Grafter	84	-SV5	45,478	8,724	6.19	PS
9902	The Middle of Things	84	-SV5	74,618	8,234	5.96	PR
14107	The Lost Stradivarius	84	-SV1	50,541	7,239	6.06	PR
18613	The Golden Scorpion	84	-SV5	67,147	10,022	5.75	PR
21927	Short Cruises	84	-SV1	41,057	6,011	6.08	PR
25770	The Dragon's Secret	84	-SV1	41,550	5,907	5.98	PZ
942	Green Mansions: A Romance of the...	83	-SV1	89,629	10,529	5.94	PR
1953	A Book of Strife in the Form of ...	83	-SV1	25,170	5,022	6.03	PR
8188	The Mysterious Key and What It O...	83	-SV1	20,093	4,032	6.14	PS
18492	Star Surgeon	83	-SV1	52,343	7,138	5.60	PS
18800	Last Enemy	83	-SV1	24,857	4,923	5.93	PS
22064	Tess of the Storm Country	83	-SV2	96,743	10,645	5.85	PS
22495	The New Pun Book	83	SV1	24,742	5,522	6.02	PN
25438	The Airlords of Han	83	-SV1	30,904	6,236	5.86	PS
37660	Of All Things	83	-SV1	44,177	9,370	6.10	PS
39281	Dictionary of English Proverbs a...	83	SV3	31,556	10,059	5.77	PN
311	Bunner Sisters	82	-SV1	31,612	5,941	6.02	PS
5758	Many Cargoes	82	SV6	69,637	9,188	6.08	PR
8681	The Face and the Mask	82	-SV5	72,288	9,658	5.82	PR
10736	Children of the Frost	82	SV2	51,252	7,626	5.65	PS
17221	History of the Plague in London	82	-SV1	95,676	10,368	5.54	PR
20730	For the Sake of the School	82	-SV1	62,777	9,967	6.16	PZ
32934	The Young Colonists: A Story of ...	82	-SV3	78,959	8,441	5.62	PZ
37172	In a Glass Darkly, v. 1/3	82	-SV1	40,539	7,628	5.70	PR
38562	The Big Book of Nursery Rhymes	82	-SV2	22,021	4,327	6.06	PZ
5317	Through the Magic Door	81	SV3	47,696	9,239	6.01	PR,Z
6840	Queen Lucia	81	-SV6	88,117	10,588	6.23	PR
18361	Operation: Outer Space	81	-SV1	61,178	9,430	6.01	PS
26862	Howard Pyle's Book of Pirates : F...	81	SV3	87,360	10,963	5.86	PS
34426	The Enchanted Barn	81	-SV2	99,579	9,502	6.27	PS
3797	In the Days of the Comet	80	SV1	81,975	12,781	5.91	PR
10234	Old Creole Days: A Story of Creo...	80	SV4	69,210	11,858	5.92	PS
18970	Caves of Terror	80	-SV1	45,847	7,534	5.83	PR
19717	The Bostonians, Vol. I (of II)	80	SV3	80,403	9,990	6.21	PS
20431	The Tale of Beowulf, Sometime Ki...	80	-SV2	39,504	6,395	5.36	PR
22145	A Book of Burlesques	80	-SV1	30,934	8,261	5.89	PS
32953	Quest of the Golden Ape	80	-SV1	34,938	6,090	5.80	PS
545	At the Earth's Core	79	-SV1	51,251	7,425	5.94	PS
4731	Seven Little Australians	79	-SV4	46,663	7,814	6.02	PZ,PR
7308	The History of Mr. Polly	79	-SV1	70,301	12,092	5.93	PR
11045	The Ghost Ship	79	-SV2	53,173	8,091	6.04	PR
37858	Leaves in the Wind	79	-SV5	68,208	11,265	5.85	PR
40852	Instigations: Together with An Es...	79	SV8	99,318	22,573	6.14	PN
1327	Elizabeth and Her German Garden	78	-SV1	49,047	8,018	6.18	PR
2250	Richard II	78	SV1	23,824	5,612	5.55	PR
2431	Is Shakespeare Dead? : From My Au...	78	SV1	22,237	5,354	6.08	PR
10066	Gunman's Reckoning	78	SV2	81,339	9,128	5.80	PS
10671	The Botanic Garden. Part II.: Con...	78	SV2	55,964	12,356	6.18	PR
20726	A Slave is a Slave	78	SV1	21,980	4,760	5.47	PS
26027	Puck of Pook's Hill	78	-SV6	60,322	9,644	5.94	PR,PZ
35612	Three Philosophical Poets: Lucre...	78	SV1	49,476	9,967	6.16	PN
678	The Cricket on the Hearth: A Fai...	77	-SV1	33,105	5,887	6.13	PR
1106	The Tragedy of Titus Andronicus	77	SV2	25,230	4,868	5.78	PR
3479	The Metal Monster	77	-SV2	82,479	12,777	5.86	PS
19141	Edison's Conquest of Mars	77	-SV1	65,532	8,799	5.97	PS

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
520	The Life and Adventures of Santa...	76	SV2	31,025	4,917	6.24	PS,PZ
2251	Henry IV, Part 1	76	-SV2	27,041	5,967	5.75	PR
2861	The Sleuth of St. James's Square	76	SV4	81,152	9,906	5.87	PS
3326	The Well-Beloved: A Sketch of a ...	76	-SV6	64,400	9,764	6.09	PR
12245	The Defendant	76	SV1	27,244	6,015	6.14	PR
24353	Wired Love: A Romance of Dots and...	76	-SV1	51,185	7,741	6.08	PS
28118	The Great Gray Plague	76	SV2	26,569	5,216	5.79	PS
31308	Orientations	76	-SV4	50,779	8,068	6.00	PR
1020	Sword Blades and Poppy Seed	75	-SV5	31,494	7,182	5.95	PS
1262	The Heritage of the Desert: A Novel	75	SV5	82,671	10,548	5.86	PS
12590	The Shadow of the Rope	75	SV3	78,654	9,902	5.95	PR
18817	Ralestone Luck	75	-SV1	61,073	9,379	5.96	PS
22287	'Smiles': A Rose of the Cumberlands	75	SV2	97,058	12,873	6.06	PS
2266	King Lear	74	-SV1	26,625	5,922	5.67	PR
2324	A House to Let	74	-SV1	35,682	6,070	5.99	PR
2540	Father and Son: A Study of Two T...	74	-SV2	79,858	12,619	6.11	PR
9834	The Talleyrand Maxim	74	-SV2	73,007	9,116	6.00	PR
18761	The Circular Study	74	SV1	56,589	7,715	5.80	PS
19709	Danger in Deep Space	74	SV4	51,039	6,758	5.85	PZ
28071	The Red Triangle: Being Some Furt...	74	SV1	62,741	7,910	5.73	PR
30427	The Lost Kafoozalum	74	-SV1	22,873	4,663	5.67	PR
35425	The Mad Planet	74	SV2	21,751	4,665	5.54	PS
213	The Man from Snowy River	73	SV2	27,548	5,498	5.73	PR
687	A Personal Record	73	SV2	45,965	9,069	6.12	PR
1539	The Winter's Tale	73	SV1	26,632	5,810	5.97	PR
1795	Macbeth	73	-SV3	21,594	4,716	5.75	PR
2028	The Yellow Claw	73	-SV4	90,799	12,941	5.91	PR
3075	The Return	73	-SV8	81,784	10,698	5.90	PR
5210	The Borough	73	-SV2	69,781	13,105	5.70	PR
9297	The Orange-Yellow Diamond	73	SV2	76,745	9,127	6.02	PR
2186	"Captains Courageous": A Story o...	72	-SV3	55,256	9,619	6.02	PR
3795	Under the Lilacs	72	SV2	84,389	10,400	6.13	PZ,PS
5148	Rodney Stone	72	-SV1	95,455	11,327	5.98	PR
25051	Space Platform	72	-SV2	59,572	8,955	5.67	PS
33066	The Garden of Eden	72	SV8	76,124	8,763	5.94	PS
402	Penrod	71	SV2	59,209	11,242	6.02	PS
586	Religio Medici, Hydriotaphia, an...	71	-SV1	62,378	11,407	5.73	PR
1244	Love for Love: A Comedy	71	-SV1	35,581	5,612	6.10	PR
4253	Dramatic Romances	71	-SV4	42,767	9,449	6.06	PN,PR
7947	The Diary of a U-boat Commander: ...	71	-SV1	50,526	8,128	5.68	PR
12491	Twelve Types	71	SV4	26,567	5,926	6.05	PR,CT
32587	The Ambassador	71	-SV4	20,074	4,932	5.85	PS
39592	Princess Mary's Gift Book : All p...	71	SV3	52,891	9,080	5.92	PN
1115	The First Part of King Henry the...	70	-SV2	29,644	5,446	5.91	PR
3005	Tom Swift and His Airship	70	SV3	44,134	5,888	5.95	PZ
3777	Tom Swift and His Electric Rifle...	70	-SV1	42,586	5,543	5.73	PZ,PS
6093	Far Away and Long Ago: A History...	70	-SV2	97,738	11,935	6.04	QL,PR
6340	Literary Lapses	70	SV2	42,832	8,507	6.07	PS
13372	The Gloved Hand	70	SV6	71,710	7,949	5.79	PS
17958	Warlord of Kor	70	-SV1	33,091	5,405	5.66	PS
19258	Tom Swift and the Electronic Hyd...	70	-SV2	33,037	6,810	5.99	PZ
20877	Mother West Wind's Children	70	SV1	30,730	3,321	6.07	PZ
27690	Nobody's Girl: (En Famille)	70	-SV7	75,505	7,624	5.99	PQ,PZ
1239	The Spirit of the Border: A Roma...	69	-SV1	89,991	11,175	5.77	PS
16199	Memoirs of the Author of a Vindi...	69	-SV1	25,056	5,030	6.26	PR
17412	The Bobbsey Twins : Or, Merry Day...	69	-SV2	35,035	5,051	6.05	PZ
19330	An Apache Princess: A Tale of th...	69	SV4	83,490	11,775	5.71	PS
22234	Aunt Jo's Scrap-Bag, Vol. 5: Jimm...	69	SV5	40,473	7,344	6.26	PZ
556	Rewards and Fairies	68	-SV1	76,118	10,963	5.89	PZ,PR
4006	Yesterdays	68	SV1	22,180	4,646	5.99	PS
7230	Not George Washington — an Autob...	68	-SV1	53,845	9,655	6.08	PR
9807	Scarhaven Keep	68	SV5	75,422	9,595	6.04	PR
9903	Way of the Lawless	68	-SV4	71,853	8,427	5.82	PS
16096	A Man's Woman	68	SV2	77,180	10,379	5.72	PS
18095	Successful Methods of Public Spe...	68	-SV1	20,886	5,212	6.29	PN
22354	The Adventures of Maya the Bee	68	-SV3	38,782	6,249	6.04	PZ,PT
24933	The Man Who Knew	68	SV2	53,696	7,630	5.97	PR
26624	The Road to Oz	68	SV1	41,600	5,975	6.31	PZ
40723	The Battle of Life. A Love Story	68	SV1	30,952	5,990	6.20	PR

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
329	Island Nights' Entertainments	67	-SV3	51,115	6,252	5.86	PR
720	Almayer's Folly: A Story of an E...	67	-SV1	66,011	8,756	5.89	PR
1537	Pericles, Prince of Tyre	67	-SV2	22,020	5,170	6.08	PR
2815	Democracy, an American novel	67	-SV1	72,151	9,565	5.94	PS
8914	England, My England	67	-SV2	65,871	9,483	5.87	PR
8920	The Light of Asia	67	SV2	40,105	8,282	5.95	PR
10869	The Abandoned Room	67	-SV1	79,681	8,156	5.58	PS
17513	St. Nicholas Magazine for Boys a...	67	SV1	52,200	10,229	6.21	PZ
18346	Null-ABC	67	-SV1	36,320	6,748	5.67	PS
18420	The Bobbsey Twins at Home	67	-SV1	44,743	4,142	6.15	PZ
24302	The Highest Treason	67	-SV1	23,369	4,708	5.52	PS
24723	Final Weapon	67	SV2	20,670	4,173	5.89	PS
24749	Adaptation	67	-SV1	24,213	5,107	5.78	PS
27903	The Magic World	67	-SV1	61,067	7,869	6.06	PZ
27924	Mugby Junction	67	-SV1	52,266	8,203	5.98	PR
37820	Chronicles of Martin Hewitt	67	-SV1	68,420	10,266	5.85	PR
41667	The Emerald City of Oz	67	-SV4	55,434	7,131	6.11	PZ
1869	The Man in Lower Ten	66	SV1	65,944	9,095	5.77	PS
2509	The Lani People	66	-SV3	60,766	9,632	5.78	PS
12170	The Wolf Hunters: A Tale of Adve...	66	-SV1	53,581	6,974	5.74	PS
12352	Iola Leroy; Or, Shadows Uplifted	66	SV1	75,790	9,672	5.90	PS
23845	Talents, Incorporated	66	-SV3	52,612	7,966	5.58	PS
24459	The Lost Princess of Oz	66	-SV2	48,468	5,924	5.96	PS,PZ
35117	Lord Tony's Wife: An Adventure o...	66	-SV4	91,093	11,740	5.80	PR
296	The Cash Boy	65	-SV1	28,355	4,051	6.15	PS,PZ
1472	In a German Pension	65	-SV1	31,065	6,199	6.16	PR
2244	As You Like It	65	SV2	24,200	5,015	6.12	PR
6995	Ghosts I Have Met and Some Others	65	-SV2	36,301	6,973	5.83	PS
11068	The Spirit of the Age; Or, Conte...	65	SV3	82,318	14,398	6.04	PR
11935	Mysticism in English Literature	65	-SV1	39,912	7,647	6.27	PR
13815	The Talking Beasts: A Book of Fa...	65	-SV2	78,095	10,644	5.90	PZ
15323	The Green Eyes of Bâst	65	SV3	75,766	10,244	5.88	PR
19066	Brigands of the Moon	65	SV5	71,019	8,526	5.68	PS
20630	The Borough Treasurer	65	SV1	78,858	9,969	5.93	PR
22182	The Best of the World's Classics...	65	-SV3	68,431	13,500	5.99	PR
24499	The Green Carnation	65	-SV5	46,138	8,383	6.07	PR
30324	The Pathless Trail	65	SV5	76,745	11,341	5.67	PS
34971	Among the Forest People	65	SV1	31,880	3,835	6.21	PZ
37503	Gammer Gurton's Needle	65	-SV1	23,264	5,001	5.54	PR
39143	The Making of a Saint	65	-SV1	73,742	9,187	5.85	PR
267	The Touchstone	64	-SV1	27,270	5,620	6.07	PS
330	Where There's a Will	64	SV3	68,246	7,529	6.01	PS
1204	Cabin Fever	64	SV4	57,450	9,115	6.00	PS
1547	Sir Thomas More	64	SV1	23,934	5,195	6.00	PR
2317	The Story of My Heart: An Autobi...	64	-SV2	33,401	5,716	6.29	PR
5776	100%: the Story of a Patriot	64	-SV9	98,446	10,978	5.70	PS
7884	In the Fog	64	-SV2	22,623	3,950	5.74	PS
9656	Alarms and Discursions	64	-SV3	50,945	9,397	6.03	PR
11153	No Hero	64	SV1	39,412	6,062	6.01	PR
17180	The Riddle of the Frozen Flame	64	-SV1	66,009	9,317	5.91	PS
18172	This World Is Taboo	64	-SV3	37,649	6,489	5.72	PS
19027	The Revolt on Venus	64	-SV1	54,941	7,369	5.72	PZ
19526	Stand by for Mars!	64	SV4	58,320	7,547	6.02	PZ
25067	The Planet Strappers	64	-SV1	67,425	12,119	5.88	PS
30537	The Royal Book of Oz : In which t...	64	-SV2	43,385	6,680	6.10	PZ
1583	Options	63	SV2	66,078	11,924	6.15	PS
1790	Troilus and Cressida	63	SV1	32,213	5,996	5.96	PR
3464	Tish: The Chronicle of Her Escap...	63	-SV2	91,009	10,697	5.92	PS
4381	The Aran Islands	63	-SV2	50,695	6,255	6.04	PR
5071	The Philanderer	63	-SV1	25,680	4,466	5.94	PR
6574	Watchers of the Sky	63	-SV1	33,018	6,964	6.07	PR
7365	If I May	63	-SV2	46,730	8,050	6.10	PR
14034	King Alfred's Viking: A Story of ...	63	-SV1	78,775	6,892	5.92	PZ
14154	The Tale of Terror: A Study of t...	63	-SV4	75,973	15,061	5.74	PN
18719	Space Tug	63	SV4	56,963	8,400	5.79	PS
27567	Aunt Jo's Scrap-Bag VI: An Old-Fa...	63	SV2	56,723	7,878	6.14	PZ
32563	The Lost Warship	63	-SV1	27,106	4,701	5.60	PS
32597	Accidental Flight	63	SV1	22,365	4,491	5.69	PS
34215	Shadowings	63	-SV2	43,506	10,051	6.23	PS

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
36281	The Slayer of Souls	63	-SV2	63,000	9,989	5.89	PS
41718	Dave Dawson on the Russian Front	63	SV1	50,697	6,407	5.86	PZ
764	Hans Brinker; Or, The Silver Skates	62	-SV1	85,043	11,700	6.03	PZ
966	Maid Marian	62	-SV1	36,838	7,162	5.99	PR
1163	Adventure	62	-SV2	76,325	10,405	5.80	PS
5311	Parnassus on Wheels	62	-SV1	36,880	6,444	6.09	PS
5676	A Double Story	62	SV1	34,563	4,828	5.85	PR,PZ
6359	The English Mail-Coach and Joan ...	62	-SV1	45,492	10,693	5.81	PR
9867	Riders of the Silences	62	-SV3	64,857	7,928	5.80	PS
11620	My Brilliant Career	62	-SV1	90,607	12,511	6.19	PR
16259	The Surprising Adventures of the...	62	-SV1	33,039	4,886	6.00	PZ
18217	Chambers's Elementary Science Re...	62	SV3	21,590	3,694	6.41	Q,PZ
18824	Fairies and Folk of Ireland	62	-SV1	74,241	5,535	6.06	PZ
21665	A Brief History of the English L...	62	SV1	78,236	15,998	6.21	PE,PN
27198	The Explorer	62	SV6	80,205	9,337	5.86	PR
30905	The Boarded-Up House	62	-SV2	34,944	5,682	6.01	PZ
35247	That Affair at Elizabeth	62	SV3	62,452	8,084	5.88	PS
35330	The Spanish Tragedy	62	SV1	26,640	5,113	5.74	PR
436	The Master Key: An Electrical Fa...	61	-SV1	36,417	6,242	5.92	PZ
2015	A Miscellany of Men	61	SV1	53,088	9,641	5.92	PR
2876	The Light That Failed	61	-SV1	74,177	10,326	5.90	PR
4993	A Texas Ranger	61	-SV2	72,903	9,330	5.88	PS
5747	Do and Dare — a Brave Boy's Figh...	61	-SV1	50,698	6,450	6.01	PZ
10443	The Rayner-Slade Amalgamation	61	-SV5	80,402	9,662	6.11	PR
11435	Small Means and Great Ends	61	SV2	31,081	6,087	6.02	PZ
17854	The Sport of the Gods	61	-SV2	41,937	6,572	5.94	PS
19023	A Daughter of the Sioux: A Tale ...	61	-SV3	77,282	11,415	5.74	PS
21865	King Arthur and His Knights	61	-SV1	43,484	4,920	5.71	PZ
22031	The Airplane Boys among the Clou...	61	SV2	49,500	6,496	5.97	PZ
22342	Supermind	61	SV7	73,369	9,148	5.85	PS
23028	Greylorn	61	-SV3	20,318	4,472	5.63	PS
28164	The Big Bow Mystery	61	-SV2	44,956	8,744	5.64	PR
30334	Ultima Thule	61	-SV12	26,476	5,094	5.92	PS
31356	The Man Who Staked the Stars	61	SV3	26,606	5,742	5.77	PS
32486	The Legion of Lazarus	61	-SV1	21,925	4,383	5.69	PS
34219	The Enchanted Castle	61	SV5	72,019	9,467	6.10	PZ
35617	The Terror: A Mystery	61	SV1	38,210	6,221	5.67	PR
36958	A Child of the Jago	61	-SV4	63,674	10,480	5.76	PR
1478	A Parody Outline of History : Whe...	60	-SV3	23,588	5,505	6.10	PN
1585	The Wrong Box	60	-SV2	58,971	9,665	5.87	PR
5083	The Man of Feeling	60	SV2	38,372	7,002	6.05	PR
10110	The Postmaster's Daughter	60	SV1	67,192	10,752	5.87	PR
18934	My Lady Nicotine: A Study in Smoke	60	-SV2	52,469	7,952	5.92	PR
19381	Among the Farmyard People	60	-SV1	40,239	4,250	6.11	PZ
19478	Four-Day Planet	60	-SV1	59,954	7,788	5.69	PS
19535	George Bernard Shaw	60	-SV3	53,197	9,069	5.94	PR
25102	Nobody's Boy: Sans Famille	60	SV2	92,076	8,041	5.98	PQ,PZ
26999	Peter Pan in Kensington Gardens	60	-SV2	24,535	3,990	6.12	PZ,PR
27063	The Hero	60	-SV7	73,181	9,185	5.98	PR
32440	Dave Dawson at Dunkirk	60	SV3	52,993	6,488	5.76	PZ
33505	The Trembling of the Veil	60	-SV3	71,863	10,742	6.08	PR
33623	The Inventions of the Idiot	60	-SV1	27,697	5,284	6.17	PS
41049	The Onslaught from Rigel	60	-SV3	68,709	9,387	5.81	PS
1076	The Wallet of Kai Lung	59	SV1	80,921	9,857	5.93	PR
1446	Perfect Behavior: A Guide for La...	59	-SV3	38,230	7,901	6.26	PN
1460	The Black Dwarf	59	-SV1	58,461	10,507	6.01	PR
1882	The Young Forester	59	-SV1	51,323	7,232	5.76	PS
5660	Mary Louise	59	-SV2	47,424	7,310	6.07	PZ,PS
12345	Friday, the Thirteenth : A Novel	59	-SV1	48,415	7,671	5.88	PS
17870	Operation Terror	59	-SV3	47,352	7,069	5.75	PS
21687	The Youngest Girl in the Fifth: ...	59	SV7	75,351	10,291	6.07	PZ
23641	The Forsaken Inn: A Novel	59	-SV4	66,098	8,716	5.89	PS
25767	Picture and Text: 1893	59	-SV1	32,974	6,836	6.27	PS
27595	Eight Keys to Eden	59	SV2	54,529	8,308	5.84	PS
36869	The Real Man	59	SV3	94,115	11,403	5.92	PS
94	Alexander's Bridge	58	-SV1	29,576	5,211	6.15	PS
1263	The Glimpses of the Moon	58	-SV8	84,375	11,356	6.16	PS
1282	Tom Swift Among the Diamond Make...	58	SV2	43,526	5,598	5.97	PZ
2305	A Set of Six	58	SV3	85,490	12,412	5.87	PR

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
2524	My Lady Ludlow	58	SV2	80,667	9,495	5.97	PR
2722	Morning Star	58	SV2	85,928	8,230	5.82	PR
2865	Otto of the Silver Hand	58	-SV1	28,161	4,517	5.72	PZ
5632	Five Little Peppers Midway	58	-SV3	72,449	7,747	6.14	PZ,PS
8435	The Sturdy Oak : A composite Nove...	58	-SV1	60,759	10,043	6.06	PS
9963	Elsie's Girlhood: A Sequel to "El...	58	-SV1	96,936	9,609	6.28	PZ
10324	Bull Hunter	58	-SV3	53,488	7,322	5.80	PS
12028	The Uttermost Farthing: A Savant'...	58	SV4	53,811	9,084	5.72	PR
14902	Deadwood Dick, the Prince of the...	58	SV1	35,479	7,613	5.87	PS
15422	Israel Potter : His Fifty Years o...	58	-SV1	66,586	12,774	5.83	PS
24370	Mercenary	58	-SV1	21,976	4,735	6.09	PS
24436	Anything You Can Do ...	58	SV2	55,504	8,734	5.74	PS
25728	Desert Conquest; or, Precious Wa...	58	-SV6	97,231	13,140	5.90	PS
28700	Robin Hood	58	SV2	94,173	10,221	5.96	PZ
32208	The Star Lord	58	-SV1	21,646	4,555	5.92	PS
34592	Behind the Green Door	58	-SV1	44,256	6,588	5.93	PZ
38777	Lad: A Dog	58	-SV5	83,311	12,631	5.81	PZ
1146	The Journal of a Voyage to Lisbon	57	SV3	45,588	7,422	6.14	PR
2154	Around the World in Eighty Days....	57	-SV3	63,910	9,642	5.89	PZ,PQ
2487	Cross Roads	57	-SV3	26,095	5,064	6.11	PS
4922	Bar-20 Days	57	-SV5	71,505	9,892	5.68	PS
7239	Men, Women, and Boats	57	-SV1	54,316	9,802	5.79	PS
9609	Joseph Andrews, Vol. 2	57	SV2	72,846	9,091	6.09	PR
11626	The Dawn of All	57	-SV4	89,297	10,893	5.93	PR
14427	True Love's Reward : A Sequel to ...	57	-SV1	70,493	7,636	6.05	PS
18881	The Idiot	57	SV1	22,941	4,767	6.16	PS
30836	Seven Keys to Baldpate	57	SV7	74,942	9,803	6.02	PS
32117	Eleven Possible Cases	57	-SV2	60,086	9,757	5.97	PS
32226	The Flower Princess	57	-SV1	21,907	3,938	6.17	PZ
36612	The Princess and Curdie	57	-SV1	57,852	7,439	5.96	PZ
37193	The Swedish Fairy Book	57	-SV2	51,321	5,610	6.05	PZ
1109	Love's Labour's Lost	56	-SV2	26,608	5,382	6.04	PR
1604	The Ebb-Tide: A Trio And Quartette	56	-SV1	48,277	8,357	5.89	PR
1611	Seventeen : A Tale of Youth and S...	56	-SV6	68,908	10,464	6.11	PS
1897	The Seventh Man	56	-SV3	78,397	9,114	5.74	PS
2713	Maiwa's Revenge; Or, The War of ...	56	-SV2	34,909	5,414	5.60	PR
4282	Don Rodriguez; Chronicles of Sha...	56	SV1	74,129	8,303	5.96	PR
5795	The Secret Rose	56	-SV2	22,480	4,147	5.87	PR
6936	Robinson Crusoe — in Words of On...	56	-SV1	27,568	2,425	5.79	PR,PZ
8673	A Columbus of Space	56	-SV8	71,372	9,309	5.97	PS
9791	Harrigan	56	-SV1	67,059	8,219	5.79	PS
9990	Brave and Bold; Or, The Fortunes...	56	-SV3	56,686	6,844	6.08	PZ
11377	The Man Whom the Trees Loved	56	SV1	26,393	5,365	6.02	PR
12793	Cobwebs from an Empty Skull	56	-SV2	53,936	11,664	5.89	PS
14667	A Christmas Garland	56	-SV2	29,981	8,075	6.09	PR
21632	Fame and Fortune; or, The Progre...	56	-SV1	45,855	5,764	6.25	PZ
21656	The Princess of the School	56	SV5	65,548	10,190	6.27	PZ
22338	The Impossibles	56	-SV2	52,808	6,920	5.90	PS
24929	The Green Rust	56	SV4	84,006	10,911	5.87	PR
25866	The Search	56	-SV5	64,628	7,748	6.03	PS
26019	Europa's Fairy Book	56	SV3	59,855	7,879	6.20	PZ
27922	David and the Phoenix	56	SV1	30,748	5,994	5.97	PZ
31343	The Invaders	56	-SV1	22,168	4,683	5.78	PS
751	The Autocrat of the Breakfast-Table	55	SV5	93,970	16,700	6.17	PS
2126	The Quest of the Sacred Slipper	55	-SV3	53,235	8,334	5.65	PR
4020	Arcadian Adventures with the Idl...	55	-SV1	67,329	9,609	6.05	PS
7088	The Pilgrim's Progress in Words ...	55	SV1	27,072	2,578	5.90	PR,PZ
14744	Different Girls	55	SV2	56,778	9,060	6.10	PS
19355	A Book of Prefaces	55	-SV2	61,149	12,820	5.97	PS
20559	R. Holmes & Co. : Being the Remar...	55	SV2	33,956	6,479	5.97	PS
27174	Captain Jim	55	-SV1	90,723	11,003	6.05	PZ
28267	Venus in Boston: A Romance of Ci...	55	SV1	59,354	10,376	6.03	PS
37995	The Diamond Fairy Book	55	SV1	57,643	8,945	6.07	PZ
1358	Enoch Arden, &c.	54	-SV5	26,435	6,080	5.92	PR
2515	Stepping Heavenward	54	SV2	97,434	9,061	6.03	PS
10067	The Mystery of the Boule Cabinet...	54	-SV6	69,880	7,994	5.81	PS
12431	The Coquette, or, The History of...	54	-SV1	59,565	8,390	6.31	PS
12986	The Card, a Story of Adventure i...	54	SV5	76,379	11,439	6.05	PR
13888	Bacon	54	SV2	72,536	10,588	5.94	PR

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
16721	A Place so Foreign	54	-SV4	21,982	4,826	6.07	PS
21092	On the Trail of the Space Pirates	54	-SV1	54,158	6,935	5.87	PZ
21633	The Man of the Desert	54	-SV2	61,623	7,650	6.08	PS
32746	The Revival of Irish Literature ...	54	SV2	33,517	8,215	6.14	PR
34313	Literature in the Making, by Som...	54	-SV3	54,321	8,459	6.24	PN
35204	Sense of Obligation	54	-SV1	54,726	8,435	5.50	PS
1127	The Tragedy of Othello, Moor of ...	53	-SV1	32,617	5,445	5.90	PR
1143	Notes on Life & Letters	53	-SV3	81,978	12,718	5.95	PR
3475	The Efficiency Expert	53	-SV1	51,038	6,758	6.01	PS
6987	Five Little Peppers Abroad	53	-SV4	84,392	8,095	6.24	PZ,PS
9156	Life and Remains of John Clare, ...	53	-SV3	69,387	11,805	6.21	PR
10337	Lady into Fox	53	SV2	24,701	4,060	5.79	PR
10847	The Maids Tragedy	53	-SV1	32,291	5,048	5.66	PR
11195	Alcatraz	53	SV2	70,059	9,099	5.72	PS
11371	The Moorland Cottage	53	-SV1	44,584	6,794	6.05	PR
15585	Humorous Masterpieces from Ameri...	53	SV3	59,148	11,782	6.07	PN
20551	The White Invaders	53	-SV3	32,428	5,644	5.78	PS
20698	The Story of Glass	53	-SV2	39,063	6,299	6.30	PZ
21715	Away in the Wilderness	53	SV4	28,138	4,929	5.95	PZ
22057	Kid Wolf of Texas : A Western Story	53	-SV2	66,418	8,981	5.57	PS
22463	Chivalry	53	SV2	60,217	11,228	5.85	PS
35533	The Haunted Room: A Tale	53	-SV2	71,695	9,634	5.90	PR
39868	Glinda of Oz : In which are Relat...	53	SV1	40,697	5,410	6.13	PZ
949	Tom Swift and His Submarine Boat...	52	-SV1	44,612	5,539	6.04	PZ
1283	Tom Swift and His Wizard Camera;...	52	-SV1	45,898	5,938	5.93	PZ
2260	Titus Andronicus	52	SV2	22,535	5,119	5.62	PR
2273	Tom Swift and His Motor-Boat; Or...	52	SV2	45,964	5,255	5.90	PZ
2295	Waifs and Strays: Part 1	52	-SV1	30,834	7,421	6.14	PS
13783	The Boy Inventors' Radio Telephone	52	SV2	46,496	7,559	5.94	PZ
14632	The Mystery of Mary	52	SV4	37,618	5,479	6.09	PS
30333	Daddy's Girl	52	-SV2	77,411	7,813	6.25	PZ
33979	Miscellaneous Aphorisms; The Sou...	52	-SV1	33,159	5,315	5.93	PR,HX
1625	The Frozen Deep	51	-SV3	28,550	4,900	6.04	PR
4268	Cousin Phillis	51	-SV1	40,955	6,050	6.18	PR
5815	The Great Impersonation	51	-SV1	76,505	9,387	6.00	PR
6418	Five Little Peppers and their Fr...	51	-SV1	88,606	8,260	6.02	PS,PZ
9196	The Clockmaker; Or, the Sayings ...	51	SV5	75,395	10,028	5.91	PS
10551	Affair in Araby	51	SV1	54,814	8,213	5.88	PR
14654	A Daughter of the Snows	51	-SV1	93,032	13,136	5.85	PS
15625	The Lookout Man	51	SV4	75,499	10,423	6.03	PS
16740	The Busie Body	51	-SV1	26,653	5,854	5.90	PR
17667	Dialogues of the Dead	51	SV5	64,892	9,719	6.01	PR
19370	Ullr Uprising	51	-SV1	42,036	7,215	5.52	PS
29466	Lords of the Stratosphere	51	-SV1	23,283	4,634	5.78	PS
38245	Atlantic Classics, Second Series	51	SV1	76,638	13,453	6.17	PS
38887	How to Write a Novel: A Practica...	51	SV1	36,995	8,040	6.11	PN
41027	The Revolt of the Star Men	51	-SV3	28,402	5,952	5.66	PS
557	Puck of Pook's Hill	50	-SV6	62,110	8,938	5.93	PR,PZ
794	The Wouldbegoods: Being the Furt...	50	SV5	80,570	9,339	5.97	PR,PZ
953	Tom Swift and His Big Tunnel; Or...	50	-SV1	46,119	6,273	5.83	PZ
1159	Fire-Tongue	50	SV2	64,506	9,211	5.81	PR
1515	The Merchant of Venice	50	SV3	24,592	5,105	6.06	PR
2246	All's Well That Ends Well	50	-SV1	25,941	5,352	6.04	PR
2911	Justice	50	SV2	26,466	4,241	5.75	PR
5066	The Whole Family: a Novel by Twe...	50	SV2	79,321	9,742	6.16	PS
5090	I Will Repay	50	-SV1	64,465	9,401	5.70	PR
5606	Guns of the Gods: A Story of Yas...	50	-SV3	91,027	12,068	6.01	PR
8197	India's Love Lyrics	50	SV1	25,289	4,960	6.03	PR
10850	Philaster; Or, Love Lies a Bleeding	50	-SV3	40,406	6,894	5.87	PR
15454	Imperium in Imperio: A Study of ...	50	-SV1	54,829	8,820	5.93	PS
22132	Giants on the Earth	50	-SV1	25,009	4,267	5.71	PS
30214	The Red Hell of Jupiter	50	-SV1	24,884	5,054	5.34	PS
30852	The Tin Woodman of Oz: A Faithful...	50	SV1	45,503	5,712	6.11	PZ
32542	Dave Dawson on Guadalcanal	50	-SV1	50,526	6,264	5.83	PZ
32730	The Heart of a Woman	50	-SV1	81,281	10,534	5.84	PR
33582	Rhyme? And Reason?	50	-SV1	21,700	5,779	5.83	PR
40241	Hieroglyphics	50	-SV4	45,287	7,771	6.28	PN
295	The Early Short Fiction of Edith...	49	SV3	43,439	8,000	5.97	PS
872	Reprinted Pieces	49	-SV8	95,667	15,086	6.02	PR

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
897	The Rose and the Ring	49	SV2	30,041	5,888	6.05	PR,PZ
1508	The Taming of the Shrew	49	-SV1	24,596	5,184	6.12	PR
1809	Bucky O'Connor: A Tale of the Un...	49	SV2	79,879	10,627	5.97	PS
2257	Richard III	49	-SV2	32,096	6,461	5.53	PR
2906	The Silver Box: A Comedy in Thre...	49	-SV3	21,352	3,358	5.86	PR
4082	The Barrier	49	-SV1	90,906	11,246	5.86	PS
8730	A Little Bush Maid	49	SV5	62,805	9,160	6.07	PZ,PR
15883	The London-Bawd: With Her Charac...	49	SV1	38,798	6,282	6.06	PR
18753	The Space Pioneers	49	-SV1	53,296	6,711	6.01	PZ
19246	The Young Pitcher	49	-SV3	56,278	7,583	5.93	PZ,PS
20859	Wandl the Invader	49	-SV3	48,422	6,885	5.76	PS
21073	A Pirate of the Caribbees	49	-SV4	87,347	10,576	5.89	PZ
24313	Once a Week	49	SV5	78,356	11,024	6.05	PR
25870	A World of Girls: The Story of a...	49	-SV1	83,267	9,687	6.13	PZ
25919	Miss Mapp	49	-SV5	87,493	11,284	6.09	PR
26853	Vice Versa; or, A Lesson to Fathers	49	SV4	96,507	12,652	5.95	PR
30759	Exit Betty	49	SV3	55,532	7,223	6.11	PS
32420	A Yankee Flier with the R.A.F.	49	-SV1	39,240	5,898	5.67	PZ
35920	The Sea Lady	49	-SV1	40,745	7,242	6.07	PR
37012	The Recruiting Officer	49	SV3	21,516	4,396	6.15	PR
37758	Atlantic Classics	49	SV2	67,889	13,413	6.11	PS
41231	The Life and Beauties of Fanny Fern	49	SV4	57,760	12,433	6.20	PS
41715	Dave Dawson with the R.A.F.	49	SV2	48,629	6,587	5.75	PZ
496	The Little Lame Prince	48	-SV4	45,846	6,554	6.09	PZ
980	Alice Adams	48	-SV4	88,191	9,855	6.11	PS
1375	New Chronicles of Rebecca	48	SV1	63,357	9,754	6.08	PZ,PS
1457	Mistress Wilding	48	-SV1	90,095	11,156	5.73	PR
1751	Twilight Land	48	SV5	73,912	6,176	6.10	PZ,PS
1814	The Agony Column	48	-SV1	25,353	4,938	5.99	PS
2024	Diary of a Pilgrimage	48	SV3	43,744	7,806	6.07	PR
2245	The Taming of the Shrew	48	-SV1	23,611	5,008	6.06	PR
2389	Bardelys the Magnificent : Being ...	48	-SV1	76,982	10,070	5.81	PR
4025	Anna Christie	48	-SV1	25,963	4,958	5.92	PS
8899	Three Weeks	48	SV5	52,847	8,067	6.23	PR
14280	Holidays at Roselands : A Sequel ...	48	SV2	93,880	8,107	6.04	PZ
17028	Eastern Standard Tribe	48	SV1	55,676	10,347	5.86	PS
18520	Sabotage in Space	48	SV2	46,857	6,597	5.76	PZ
25003	The Nicest Girl in the School: A...	48	-SV3	55,712	7,364	6.15	PZ
26933	Visions and Revisions: A Book of ...	48	-SV1	53,770	11,101	5.95	PN
34943	Among the Meadow People	48	-SV5	28,691	3,813	6.17	PZ
40320	Mr. Punch Afloat: The Humours of...	48	SV1	23,660	6,980	6.14	PN
40493	The King of Diamonds: A Tale of ...	48	-SV2	82,673	12,226	5.96	PR
45658	The Mystery of the Downs	48	-SV7	81,156	8,596	5.78	PR
618	Codex Junius 11	47	SV3	40,603	5,557	6.00	PR
888	The Lazy Tour of Two Idle Appren...	47	-SV2	42,002	7,799	5.99	PR
1264	The Wheels of Chance: A Bicyclin...	47	-SV2	56,518	10,028	6.03	PR
2019	The Bat	47	-SV1	65,484	8,953	5.80	PS
5977	Bound to Rise; Or, Up the Ladder	47	-SV1	48,353	5,861	6.25	PZ
7434	The Adventures of Joel Pepper	47	-SV4	83,873	6,946	5.98	PZ,PS
10119	Adonais	47	-SV1	50,656	9,930	5.68	PR
10882	The Eagle's Shadow	47	-SV1	50,415	9,069	6.04	PS
17112	Many Thoughts of Many Minds: A Tr...	47	-SV2	78,328	14,061	6.03	PN
19360	Six to Sixteen: A Story for Girls	47	-SV4	66,042	10,056	6.22	PZ
19672	The Holladay Case: A Tale	47	-SV3	45,710	6,713	5.86	PS
20739	Rebels of the Red Planet	47	-SV4	47,731	7,116	5.80	PS
20856	Ten From Infinity	47	-SV2	43,848	7,280	5.81	PS
21048	Just Patty	47	-SV2	51,109	9,304	6.14	PZ
26494	Vera; Or, The Nihilists	47	-SV1	20,085	3,981	5.82	PR
29965	Two Thousand Miles Below	47	-SV2	57,790	8,161	5.67	PS
32331	Dave Dawson at Casablanca	47	-SV3	47,663	6,278	5.86	PZ
37173	In a Glass Darkly, v. 2/3	47	-SV2	39,351	7,123	6.06	PR
39116	Unicorns	47	SV3	83,569	17,663	6.09	PS
356	Beyond the City	46	-SV1	40,171	6,536	6.10	PR
1671	When a Man Marries	46	-SV7	56,170	7,879	5.85	PS
4272	The Christian Year	46	SV2	55,144	8,338	5.96	PR
5829	The Moneychangers	46	-SV2	66,414	7,447	6.01	PS
6313	Masterpieces of American Wit and...	46	-SV1	42,501	9,593	5.99	PN
9931	K	46	-SV1	95,884	10,353	5.92	PS
13054	A Thane of Wessex : Being a Story...	46	-SV1	69,893	6,477	5.88	PZ

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
13694	Mince Pie	46	SV3	53,946	11,656	6.20	PS
14228	Bracebridge Hall	46	-SV1	47,893	9,082	6.26	PS
14360	The Dawn and the Day : Or, The Bu...	46	SV2	43,524	7,764	6.08	PS
15119	Handy Dictionary of Poetical Quo...	46	-SV10	79,135	19,987	5.84	PN
19111	Code Three	46	SV2	22,299	4,629	5.90	PS
26649	Terribly Intimate Portraits	46	-SV2	25,320	7,772	6.08	PR
28849	Smugglers' Reef: A Rick Brant Sc...	46	-SV4	56,101	7,305	6.07	PZ
29416	The Mind Master	46	-SV1	29,245	5,128	5.66	PS
31598	The Egyptian Cat Mystery: A Rick...	46	-SV1	41,234	6,830	6.07	PZ
32398	Brood of the Dark Moon : (A Seque...	46	-SV2	61,907	8,591	5.77	PS
32620	The Three Mulla-mulgars	46	-SV1	70,919	8,914	5.85	PZ
34020	The Window at the White Cat	46	-SV3	69,833	8,799	5.70	PS
37174	In a Glass Darkly, v. 3/3	46	-SV4	34,057	6,291	6.00	PR
42250	Dave Dawson with the Commandos	46	-SV1	50,488	6,642	5.78	PZ
875	The Duchess of Padua	45	-SV1	25,724	4,588	5.78	PR
1077	The Mirror of Kong Ho	45	SV5	51,688	9,446	5.98	PR
1281	Tom Swift and His Aerial Warship...	45	SV4	45,747	5,898	5.58	PZ
1423	No Thoroughfare	45	-SV1	50,599	7,678	6.05	PR
1721	The Trees of Pride	45	-SV1	25,426	4,876	5.87	PR
1948	The Story of a Bad Boy	45	-SV3	57,582	10,475	5.99	PS,PZ
2763	The World's Desire	45	-SV3	86,882	8,075	5.63	PR
3179	The American Claimant	45	-SV2	66,257	10,762	5.97	PS
3490	The Admirable Crichton	45	SV1	25,297	4,833	6.15	PR
5008	Katherine's Sheaves	45	SV1	92,258	10,979	6.17	PS
6120	Soldiers Three	45	-SV2	92,097	14,165	5.82	PR
8457	Frenzied Fiction	45	-SV3	49,577	8,240	6.07	PS
9871	The Avenger	45	-SV2	76,075	8,529	5.93	PR
13135	Pardners	45	-SV1	45,724	9,125	5.75	PS
20081	A Houseful of Girls	45	-SV3	94,698	12,317	6.10	PR,PZ
20204	The Storm-Cloud of the Nineteent...	45	SV1	25,388	6,112	6.16	PR
21407	Figures of Several Centuries	45	SV1	78,509	12,961	6.20	PN
22527	Beyond the Vanishing Point	45	-SV2	28,689	4,880	5.75	PS
24770	A Prisoner of Morro; Or, In the ...	45	-SV5	60,682	7,739	5.52	PS
25780	The Fire People	45	SV3	68,632	8,496	5.98	PS
30970	Miss Cayley's Adventures	45	-SV3	86,353	13,470	5.97	PR
40038	The Lone Ranger Rides	45	-SV1	64,363	8,289	5.57	PZ,PS
306	The Early Short Fiction of Edith...	44	SV3	45,131	8,730	6.04	PS
950	Tom Swift and His Electric Runab...	44	-SV3	44,261	5,831	5.89	PZ
954	Tom Swift and His War Tank; Or, ...	44	-SV1	46,396	5,793	5.98	PZ
1828	Chronicles of the Canongate, 1st...	44	-SV1	88,216	14,389	5.98	PR
5901	Dyke Darrel the Railroad Detecti...	44	SV2	57,369	7,619	5.74	PZ,PS
5962	Oh, Money! Money! A Novel	44	SV4	83,980	9,442	6.18	PS
9380	A Nonsense Anthology	44	-SV1	53,854	12,693	5.86	PN
9862	City of Endless Night	44	-SV1	85,809	11,235	6.00	PS
13716	A Trip to Venus: A Novel	44	-SV2	51,029	9,084	6.31	PR
14888	The Inheritors	44	-SV1	63,024	9,745	6.00	PR
16255	Dickey Downy: The Autobiography ...	44	SV3	33,123	6,265	6.13	PZ,QL
16551	The Girl of the Golden West	44	-SV1	75,368	10,075	5.97	PS
17047	The Half-Hearted	44	-SV1	93,430	12,220	5.94	PR
17189	Autumn Leaves : Original Pieces i...	44	SV1	36,797	8,633	6.10	PS
19527	The Yukon Trail: A Tale of the N...	44	SV7	70,079	9,744	5.92	PS
22278	A Master of Mysteries	44	SV1	51,260	7,059	5.83	PR
24025	The New Girl at St. Chad's: A St...	44	-SV1	72,364	10,018	6.10	PZ
25449	The Young Castellan: A Tale of t...	44	-SV1	99,058	9,030	5.79	PR
25496	New Treasure Seekers; Or, The Ba...	44	SV1	70,388	8,885	6.01	PZ
26348	Lisbeth Longfrock	44	SV1	33,962	4,880	6.16	PZ
26715	Victorian Songs: Lyrics of the A...	44	-SV1	28,183	6,314	6.20	PR
28434	The Astronomy of Milton's 'Parad...	44	-SV2	83,537	11,616	6.38	PR
32501	The Golden Age	44	-SV3	38,634	8,862	5.92	PR
33348	Reveries over Childhood and Youth	44	-SV3	32,480	5,725	6.19	PR
37667	Three Hours after Marriage	44	SV1	28,921	7,248	6.10	PR
38006	The Heatherford Fortune: a sequel...	44	-SV5	64,355	8,437	6.12	PS
39682	The Idiot at Home	44	-SV1	39,678	7,169	6.15	PS
40603	The Root of All Evil	44	-SV2	80,114	10,078	6.09	PR
534	An Inland Voyage	43	SV2	38,781	7,830	6.20	PR
753	Arizona Nights	43	SV1	69,316	10,153	5.94	PS
1532	The Tragedy of King Lear	43	-SV1	29,831	6,717	5.77	PR
1590	The Amazing Interlude	43	SV5	71,512	8,673	5.76	PS
1908	Her Prairie Knight	43	-SV1	34,425	6,045	5.96	PS

Table B.1: *All Project Gutenberg eBooks considered in this study, sorted by downloads.*

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h^{b_i}_{\text{avg}}$	LoCC
2454	The Silent Bullet	43	SV2	91,190	11,389	5.73	PS
4223	The Mystery of a Hansom Cab	43	-SV3	89,817	10,341	5.81	PR
5121	Dark Hollow	43	SV5	92,271	10,955	5.81	PS
5232	Sejanus: His Fall	43	-SV1	53,886	12,775	5.82	PR
11127	The Case of Jennie Brice	43	-SV2	36,514	4,996	5.77	PS
11583	The Runaway Asteroid	43	SV4	66,662	8,393	5.97	PS
14883	Grandmother Elsie	43	SV1	63,301	7,429	6.30	PZ
17393	Men and Women	43	SV1	38,018	9,312	6.08	PR
20163	The Jolliest School of All	43	SV2	77,103	11,504	6.23	PZ
20795	The Cricket on the Hearth	43	-SV1	33,041	5,995	6.14	PR
21626	Adrift in the Wilds; Or, The Adv...	43	-SV3	63,666	8,138	5.81	PZ
24680	The Martyr of the Catacombs: A Ta...	43	SV3	33,034	5,930	5.55	PS,BV
34403	The Clock Strikes Thirteen	43	-SV2	42,721	6,660	5.96	PZ
35027	Mr. Punch's Railway Book	43	SV1	20,128	5,650	5.89	PN
38052	Reynard the Fox	43	-SV1	28,302	6,412	5.92	PR
38551	The Crux: A Novel	43	-SV1	54,556	9,177	6.22	PS
40814	Ruth Hall: A Domestic Tale of th...	43	SV1	79,856	12,644	6.19	PS
291	The Golden Age	42	-SV7	36,954	8,529	5.91	PR
1122	The Tragedy of Hamlet, Prince of...	42	-SV1	37,349	6,692	5.96	PR
1125	All's Well That Ends Well	42	-SV1	28,367	5,007	6.01	PR
1785	Julius Caesar	42	-SV1	24,321	4,298	5.69	PR
2644	Isaac Bickerstaff, Physician and...	42	SV3	44,818	7,404	6.19	PR
3185	Those Extraordinary Twins	42	-SV1	22,002	4,906	5.86	PS
6428	The Surgeon's Daughter	42	-SV1	68,002	10,881	6.09	PR
9415	Olaf the Glorious: A Story of th...	42	SV5	89,911	8,901	5.77	PZ
10317	Betty Gordon at Boarding School;...	42	-SV1	43,866	7,245	6.17	PZ
14534	Christmas with Grandma Elsie	42	-SV1	69,231	7,538	6.36	PZ
14540	When William Came	42	SV2	52,347	9,911	6.16	PR
19718	The Bostonians, Vol. II (of II)	42	-SV1	84,314	9,793	6.04	PS
20104	The Cross-Cut	42	-SV1	82,591	10,003	5.79	PS
20519	Highways in Hiding	42	SV2	84,044	10,607	5.80	PS
21768	A Desert Drama: Being The Traged...	42	-SV3	47,742	7,718	5.79	PR
24283	Down the River; Or, Buck Bradfor...	42	SV4	61,510	7,965	6.00	PZ
25585	Shakespeare, Ben Jonson, Beaumon...	42	-SV1	70,721	13,088	6.04	PR
26732	Free Air	42	SV7	85,921	14,402	6.07	PS
27129	Lyrics from the Song-Books of th...	42	SV2	42,305	7,888	5.96	PR
29774	A Yankee Flier Over Berlin	42	SV2	36,975	5,533	5.77	PZ,PS
30339	Status Quo	42	SV2	26,538	5,150	6.04	PS
30431	Calumet 'K'	42	SV2	66,673	7,053	5.99	PS
32161	Tangle Hold	42	-SV1	20,381	4,254	5.81	PS
32351	Voyage To Eternity	42	-SV2	33,961	6,638	5.86	PS
41753	Dave Dawson at Truk	42	SV5	50,142	6,149	5.74	PZ
42710	Bizarre	42	-SV1	29,868	8,839	5.98	PS
1123	Twelfth Night; Or, What You Will	41	-SV1	24,541	4,431	6.06	PR
2496	Our Village	41	SV2	52,274	10,380	6.24	PR
2702	The Lion's Skin	41	-SV1	84,165	11,056	5.82	PR
2761	Benita, an African romance	41	-SV1	78,214	8,715	5.76	PR
4050	Mates at Billabong	41	-SV1	68,141	9,970	6.09	PR,PZ
4227	Tom Swift and His Wireless Messa...	41	SV2	41,816	5,663	5.94	PS,PZ
4531	The Secret Passage	41	-SV1	92,248	8,518	5.75	PR
5162	Agatha Webb	41	SV2	90,298	9,720	5.87	PS
9190	The Greater Inclination	41	-SV2	55,175	9,195	6.04	PS
10581	Uncle Bernac: A Memory of the Em...	41	-SV2	59,936	8,167	5.98	PR
12215	Odd Craft, Complete	41	-SV3	60,836	7,711	5.98	PR
14203	Varied Types	41	SV2	39,616	7,743	6.08	PR
14875	Elsie's children	41	SV4	72,168	8,880	6.11	PZ
14917	The Wings of the Morning	41	SV3	90,412	13,493	5.83	PR
15717	Books and Persons; Being Comment...	41	-SV3	64,047	11,512	6.13	PN
18614	At the Back of the North Wind	41	SV1	27,095	3,534	6.12	PZ
19307	The Lion of Petra	41	SV5	54,669	8,206	5.80	PR
19819	Milton's Comus	41	SV3	47,977	12,916	6.05	PR
20526	Short Story Writing: A Practical ...	41	-SV1	53,163	9,052	6.16	PN
20989	'A Comedy of Errors' in Seven Acts	41	SV2	23,916	5,762	5.84	PR
21639	When Patty Went to College	41	-SV3	38,793	6,379	6.04	PS
24160	The Basket of Flowers	41	SV4	27,425	4,503	6.08	PT,PZ
25388	The Herapath Property	41	SV4	76,286	9,558	6.01	PR
25581	Rinkitink in Oz	41	-SV3	50,118	6,475	5.92	PZ
30742	Anything You Can Do!	41	SV2	28,315	5,777	5.74	PS
33325	The Spoils of Poynton	41	-SV3	72,258	8,540	6.03	PS

Table B.1: All Project Gutenberg eBooks considered in this study, sorted by downloads.

ID	Title	DL's	Mode	N_w	Uniq(N_w)	$h_{\text{avg}}^{b_i}$	LoCC
37992	The King of Pirates : Being an Ac...	41	SV2	29,764	4,043	5.99	PR
38252	Fairies I Have Met	41	-SV1	20,377	2,936	6.33	PZ
38567	Eight Cousins; Or, The Aunt-Hill	41	-SV1	76,013	9,952	6.19	PZ
39782	Brownies and Bogles	41	-SV1	25,440	6,765	6.21	PZ
40263	Folly as It Flies; Hit at by Fan...	41	SV1	90,059	14,570	6.17	PS
40504	Ginger-Snaps	41	SV2	76,388	12,810	6.15	PS
1284	Tom Swift and His Air Scout; Or,...	40	-SV1	47,739	5,894	5.91	PZ
1461	A Legend of Montrose	40	-SV1	90,196	13,422	5.93	PR
1621	Miss or Mrs.?	40	-SV1	31,396	5,596	6.16	PR
1654	An Unsocial Socialist	40	SV2	91,348	12,238	5.92	PR
1987	The Outlet	40	SV6	95,966	10,228	6.02	PS
2013	The Pit Prop Syndicate	40	SV2	97,973	10,059	5.85	PR
2687	The Snare	40	-SV1	82,451	10,571	5.76	PR
3785	In the Reign of Terror: The Adve...	40	SV2	94,244	8,690	5.82	PR,PZ
4735	The Shepherd of the Hills	40	-SV1	76,964	7,916	6.05	PS
7052	Dr. Heidenhoff's Process	40	-SV1	32,887	5,934	5.87	PS
8394	The Doings of Raffles Haw	40	-SV1	38,635	6,651	6.15	PR
10422	Caesar Dies	40	-SV1	50,889	8,731	5.70	PR
10490	The Golden Legend	40	SV1	29,603	6,902	5.73	PS
10723	Betty's Bright Idea; Deacon Pitk...	40	SV3	28,464	5,922	6.20	PS
15976	Puck of Pook's Hill	40	-SV6	59,395	9,037	5.93	PR,PZ
19079	The Adventures of Lightfoot the ...	40	-SV2	21,937	2,930	5.72	PZ
19928	Sunset Pass; or, Running the Gau...	40	-SV1	30,536	5,307	5.79	PS
20147	Rip Foster Rides the Gray Planet	40	SV2	53,257	6,799	5.68	PZ
20472	Grace Harlowe's Plebe Year at Hi...	40	SV4	53,788	7,792	6.05	PZ
21932	Embarrassments	40	-SV1	62,724	8,760	6.10	PS
22892	The Best Made Plans	40	-SV3	44,144	6,329	5.90	PS
24197	The Tinted Venus: A Farcical Rom...	40	-SV3	61,018	8,954	5.98	PR
24767	Jack O' Judgment	40	SV2	73,954	9,201	5.87	PR
25472	Blackbeard: Buccaneer	40	SV1	73,029	11,371	5.83	PZ,PS
33028	Man and Maid	40	-SV2	61,379	9,145	6.03	PR
38053	The Coo-ee Reciter: Humorous, Pa...	40	SV1	27,909	6,950	5.87	PR,PS
47530	Oliver Twist, Vol. 2 (of 3)	40	SV2	51,446	8,271	5.90	PR

B.5 Principal Component Analysis (SVD)

In this section we provide a (1) more in-depth, intuitive explanation of the method and (2) more results from the SVD analysis.

In an effort to develop a better intuition for the results of the principal component analysis by way of SVD, we plot Eq. 3.1 along with representations of the matrices in Fig. B.5.

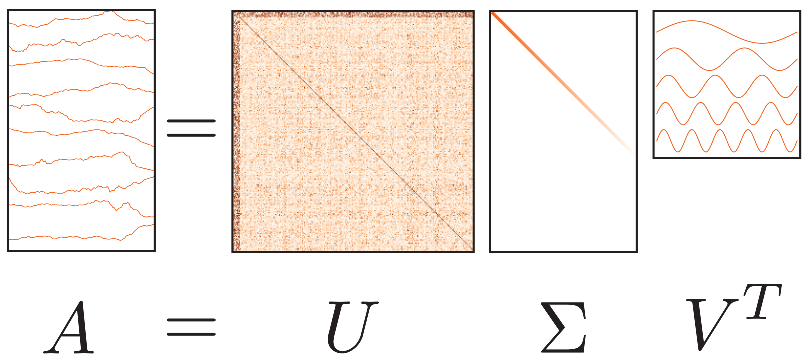


Figure B.5: Schematic of the Singular Value Decomposition applied to emotional arcs of Project Gutenberg books. Shown in A are 10 randomly chosen emotional arcs, in U a "spy" of the matrix, in Σ the decreasing singular values, and in V^T sinusoidal modes. We emphasize that this representation is purely for intuition, as only U is a image of the actual matrix, and A has only 10 of the 1,327 books.

Further, we considered in Eq. 3.1 the mode coefficient in the matrix W, and in Fig B.6 we plot the second line of the equation with W:

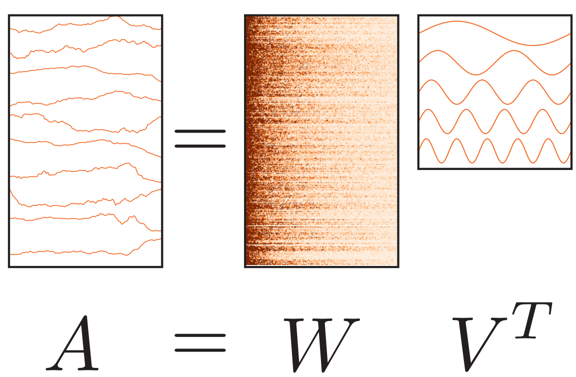


Figure B.6: Schematic of the Singular Value Decomposition applied to emotional arcs of Project Gutenberg books, with $W = U\Sigma$ containing the mode coefficients. Again shown in A are 10 randomly chosen emotional arcs, in W a "spy" of the matrix used in the analysis, and in V^T representative sinusoidal modes.

With A written as $W \cdot V^T$, the coefficients for each mode (row of V^T) for a book i are given as the rows of W. To reconstruct the emotional arc of book i, using mode j from V^T, we simply multiply $W[i,j] \cdot V^T[j,:]$. Shown below in Fig. B.7, we built the emotional arc for an example story using only the first mode through the first 12 modes.

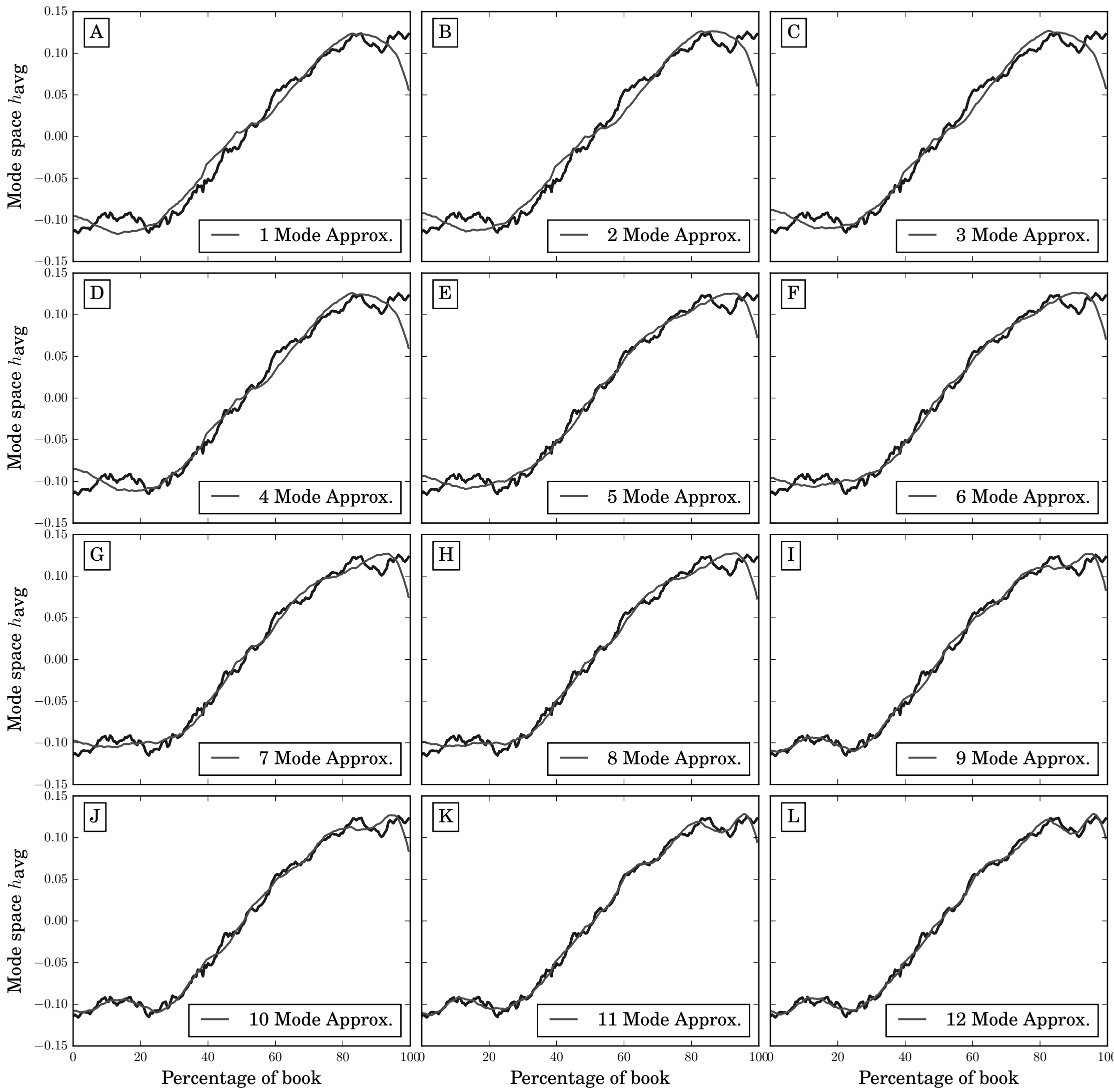


Figure B.7: Reconstruction of the emotional arc from Alice's Adventures Under Ground, by Lewis Carroll. The addition of more modes from the SVD more closely reconstructs the detailed emotional arc. This book is well represented by the first mode alone, with only minor corrections from modes 2-11, as we should expect for a book whose emotional arc so closely resembles the "Rags to Riches" arc.

B.5.1 Additional details for 40 download threshold

First, we consider modes 4–6 and their closest stories in Fig. B.8.

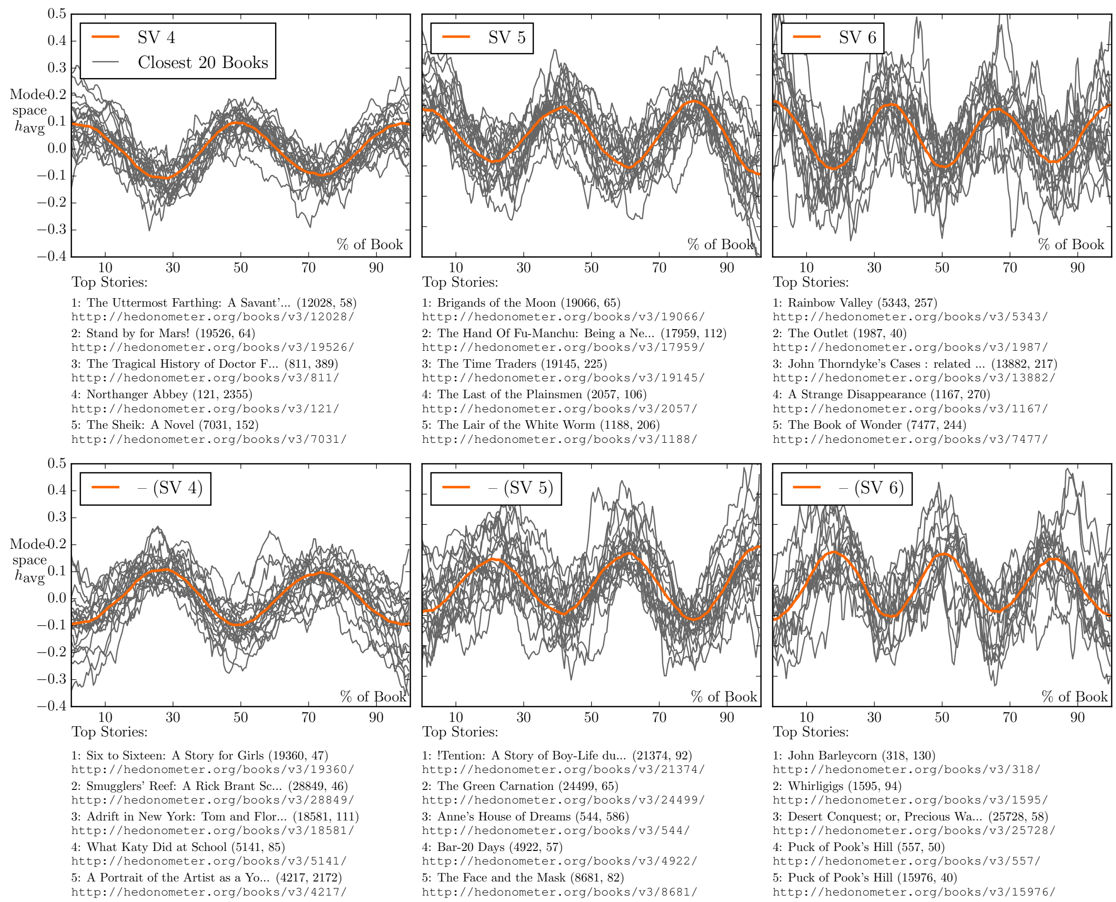


Figure B.8: SVD modes 4–6 (and their negation) with closest stories. Again, to show the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. Shown in parenthesis for each story is the Project Gutenberg ID and the number of downloads from the Project Gutenberg website, respectively. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

Next, we provide a full list of the books closest to each mode in the analysis, both sorted by downloads and support from the mode.

Table B.2: Stories which are are represented by core emotional arc 1, sorted by the variance explained in their emotional arc by core emotional arc 1.

ID	Title	DL's	max(W[i,:])	Arc
1539	The Winter's Tale	73	0.5217	
33689	Oscar Wilde, Art and Morality: A...	88	0.3804	
35617	The Terror: A Mystery	61	0.3734	
7088	The Pilgrim's Progress in Words ...	55	0.3522	
26624	The Road to Oz	68	0.3412	
21130	Book of Wise Sayings: Selected La...	86	0.3361	
2248	The Winter's Tale	137	0.3355	
485	The Road to Oz	178	0.3320	
17393	Men and Women	43	0.3030	
36127	Curious Myths of the Middle Ages	91	0.2965	
1121	As You Like It	355	0.2940	
17028	Eastern Standard Tribe	48	0.2935	
27195	Negro Folk Rhymes: Wise and Other...	92	0.2911	
1547	Sir Thomas More	64	0.2808	
2377	The Son of the Wolf	119	0.2755	
35330	The Spanish Tragedy	62	0.2726	
1905	The Governess; Or, The Little Fe...	154	0.2713	
960	The Tin Woodman of Oz : A Faithfu...	132	0.2690	
30852	The Tin Woodman of Oz: A Faithful...	50	0.2680	
20877	Mother West Wind's Children	70	0.2587	
14883	Grandmother Elsie	43	0.2575	
6043	The Spanish Tragedie	389	0.2558	
885	An Ideal Husband	1,303	0.2557	
1790	Troilus and Cressida	63	0.2506	
19551	Alice in Wonderland, Retold in W...	245	0.2489	
29228	The Contrast	88	0.2475	
25496	New Treasure Seekers; Or, The Ba...	44	0.2462	
20726	A Slave is a Slave	78	0.2462	
10002	The House on the Borderland	563	0.2452	
8197	India's Love Lyrics	50	0.2451	
18761	The Circular Study	74	0.2409	
34971	Among the Forest People	65	0.2407	
269	Beasts and Super-Beasts	804	0.2271	
22420	The Book of Nature Myths	91	0.2257	
91	Tom Sawyer Abroad	118	0.2244	
550	Silas Marner	780	0.2217	
14168	Widdershins	112	0.2191	
28521	The Power of Mesmerism: A Highly ...	643	0.2189	
38887	How to Write a Novel: A Practica...	51	0.2169	
15274	The Girl from Montana	126	0.2162	
18614	At the Back of the North Wind	41	0.2148	
11153	No Hero	64	0.2144	
41718	Dave Dawson on the Russian Front	63	0.2134	
19994	The Aesop for Children : With pic...	676	0.2120	
500	The Adventures of Pinocchio	863	0.2108	

Table B.2: Stories which are are represented by core emotional arc 1, sorted by the variance explained in their emotional arc by core emotional arc 1.

ID	Title	DL's	max(W[i,:])	Arc
17189	Autumn Leaves : Original Pieces i...	44	0.2107	
14658	The Road	147	0.2089	
40320	Mr. Punch Afloat: The Humours of...	48	0.2084	
30408	The Fifth-Dimension Tube	118	0.2078	
17513	St. Nicholas Magazine for Boys a...	67	0.2074	
788	The Red One	100	0.2068	
215	The Call of the Wild	2,439	0.2046	
14902	Deadwood Dick, the Prince of the...	58	0.2043	
4006	Yesterdays	68	0.2001	
12352	Iola Leroy; Or, Shadows Uplifted	66	0.1999	
32597	Accidental Flight	63	0.1991	
38053	The Coo-ee Reciter: Humorous, Pa...	40	0.1961	
3543	Heartbreak House	150	0.1953	
21084	Jokes For All Occasions: Selected...	281	0.1939	
10551	Affair in Araby	51	0.1915	
2431	Is Shakespeare Dead? : From My Au...	78	0.1883	
1557	Men of Iron	99	0.1877	
28071	The Red Triangle: Being Some Furt...	74	0.1868	
20204	The Storm-Cloud of the Nineteent...	45	0.1855	
367	The Country of the Pointed Firs	114	0.1834	
25803	The Keepers of the King's Peace	98	0.1821	
1375	New Chronicles of Rebecca	48	0.1808	
5676	A Double Story	62	0.1808	
22495	The New Pun Book	83	0.1785	
40263	Folly as It Flies; Hit at by Fan...	41	0.1758	
21407	Figures of Several Centuries	45	0.1745	
22278	A Master of Mysteries	44	0.1736	
18881	The Idiot	57	0.1734	
236	The Jungle Book	3,478	0.1728	
554	The Contrast	144	0.1722	
2250	Richard II	78	0.1698	
5008	Katherine's Sheaves	45	0.1694	
21665	A Brief History of the English L...	62	0.1663	
17157	Gulliver's Travels into Several ...	528	0.1643	
10150	Dracula's Guest	456	0.1631	
35997	The Jungle Book	1,370	0.1623	
27922	David and the Phoenix	56	0.1611	
11377	The Man Whom the Trees Loved	56	0.1606	
1869	The Man in Lower Ten	66	0.1595	
5348	Ragged Dick, Or, Street Life in ...	378	0.1588	
3490	The Admirable Crichton	45	0.1585	
20630	The Borough Treasurer	65	0.1577	
33156	Young's Night Thoughts: With Life...	111	0.1557	
25472	Blackbeard: Buccaneer	40	0.1552	
12245	The Defendant	76	0.1551	

Table B.2: *Stories which are are represented by core emotional arc 1, sorted by the variance explained in their emotional arc by core emotional arc 1.*

ID	Title	DL's	max(W[i,:])	Arc
4980	Old Granny Fox	128	0.1550	
29468	The Story of Don Quixote	88	0.1545	
21816	The Confidence-Man: His Masquerade	289	0.1503	
961	Glinda of Oz : In Which Are Relat...	184	0.1498	
26348	Lisbeth Longfrock	44	0.1487	
3781	The Jewel of Seven Stars	160	0.1486	
39868	Glinda of Oz : In which are Relat...	53	0.1476	
47529	Oliver Twist, Vol. 1 (of 3)	105	0.1469	
3011	The Lady of the Lake	399	0.1455	
21530	The Angel of Terror	169	0.1452	
8668	Revenge!	111	0.1442	
3797	In the Days of the Comet	80	0.1441	
753	Arizona Nights	43	0.1427	
3756	Indiscretions of Archie	162	0.1397	
35027	Mr. Punch's Railway Book	43	0.1391	
28267	Venus in Boston: A Romance of Ci...	55	0.1381	
2770	Five Little Peppers and How They...	207	0.1379	
2015	A Miscellany of Men	61	0.1369	
15883	The London-Bawd: With Her Charac...	49	0.1363	
40723	The Battle of Life. A Love Story	68	0.1350	
78	Tarzan of the Apes	1,272	0.1341	
837	The Story of the Amulet	120	0.1326	
37995	The Diamond Fairy Book	55	0.1318	
35612	Three Philosophical Poets: Lucre...	78	0.1299	
38245	Atlantic Classics, Second Series	51	0.1292	
501	The Story of Doctor Dolittle	504	0.1268	
32706	Triplanetary	309	0.1213	
40814	Ruth Hall: A Domestic Tale of th...	43	0.1212	
1076	The Wallet of Kai Lung	59	0.1205	
37667	Three Hours after Marriage	44	0.1181	
10490	The Golden Legend	40	0.1159	
16389	The Enchanted April	170	0.1153	
42254	Beyond the Black River	141	0.1152	
173	The Insidious Dr. Fu Manchu	245	0.1142	
779	The Tragical History of Doctor F...	2,133	0.1133	
4282	Don Rodriguez; Chronicles of Sha...	56	0.1103	
873	A House of Pomegranates	172	0.1092	
15851	Love Conquers All	102	0.0987	
844	The Importance of Being Earnest:...	9,373	0.0963	
40426	Daddy Long-Legs: A Comedy in Fou...	192	0.0926	
1929	The School for Scandal	417	0.0907	
16732	Familiar Quotations	86	0.0800	
10110	The Postmaster's Daughter	60	0.0766	

Table B.3: Top 10 stories which are are represented by core emotional arc 1, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
844	The Importance of Being Earnest:...	9,373	0.0963	
236	The Jungle Book	3,478	0.1728	
215	The Call of the Wild	2,439	0.2046	
779	The Tragical History of Doctor F...	2,133	0.1133	
35997	The Jungle Book	1,370	0.1623	
885	An Ideal Husband	1,303	0.2557	
78	Tarzan of the Apes	1,272	0.1341	
500	The Adventures of Pinocchio	863	0.2108	
269	Beasts and Super-Beasts	804	0.2271	

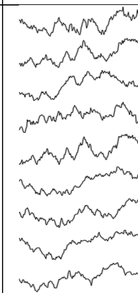


Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
946	Lady Susan	894	0.5006	
17958	Warlord of Kor	70	0.4664	
17144	The House of the Vampire	188	0.4657	
93	Tom Sawyer, Detective	112	0.4566	
159	The Island of Doctor Moreau	1,083	0.4538	
32208	The Star Lord	58	0.4504	
790	Lady Windermere's Fan	485	0.4257	
1777	Romeo and Juliet	186	0.4013	
875	The Duchess of Padua	45	0.3945	
4025	Anna Christie	48	0.3934	
39782	Brownies and Bogles	41	0.3913	
30796	The Dueling Machine	117	0.3891	
311	Bunner Sisters	82	0.3866	
38252	Fairies I Have Met	41	0.3835	
21632	Fame and Fortune; or, The Progre...	56	0.3744	
267	The Touchstone	64	0.3732	
1531	Othello, the Moor of Venice	147	0.3729	
42259	The People of the Black Circle	132	0.3706	
16517	Liza of Lambeth	107	0.3676	
30214	The Red Hell of Jupiter	50	0.3621	
16199	Memoirs of the Author of a Vindi...	69	0.3621	
2266	King Lear	74	0.3605	
23790	The Ultimate Weapon	126	0.3596	
35920	The Sea Lady	49	0.3591	
7052	Dr. Heidenhoff's Process	40	0.3564	
2265	Hamlet	1,051	0.3527	
451	The Shadow Line: A Confession	127	0.3483	
1719	The Ballad of the White Horse	394	0.3481	
20656	Old Christmas From the Sketch Bo...	101	0.3451	
21334	The Beaux-Stratagem	86	0.3449	
949	Tom Swift and His Submarine Boat...	52	0.3439	
1127	The Tragedy of Othello, Moor of ...	53	0.3438	

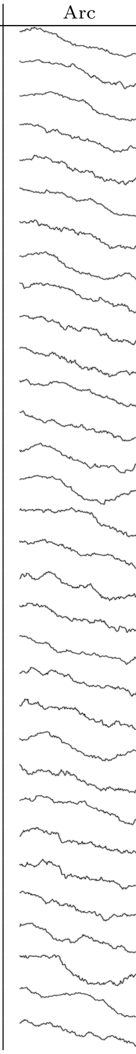


Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
4358	The Sea Fairies	98	0.3424	
38052	Reynard the Fox	43	0.3396	
2865	Otto of the Silver Hand	58	0.3383	
33660	The Year When Stardust Fell	106	0.3377	
2232	The Duchess of Malfi	534	0.3375	
14255	Hints for Lovers	127	0.3374	
18800	Last Enemy	83	0.3320	
20058	The Napoleon of Notting Hill	182	0.3319	
2263	Julius Caesar	150	0.3319	
5429	Preface to Shakespeare	148	0.3306	
24749	Adaptation	67	0.3298	
18420	The Bobbsey Twins at Home	67	0.3291	
1882	The Young Forester	59	0.3260	
20121	Lone Star Planet	103	0.3234	
42250	Dave Dawson with the Commandos	46	0.3217	
35	The Time Machine	3,732	0.3199	
24302	The Highest Treason	67	0.3176	
1621	Miss or Mrs.?	40	0.3156	
34592	Behind the Green Door	58	0.3138	
6879	The Gold Bat	120	0.3128	
1041	Shakespeare's Sonnets	831	0.3125	
2245	The Taming of the Shrew	48	0.3121	
94	Alexander's Bridge	58	0.3121	
19928	Sunset Pass; or, Running the Gau...	40	0.3073	
32154	The Variable Man	618	0.3059	
12915	The White Devil	116	0.3057	
172	The Haunted Bookshop	132	0.3052	
4023	Candida	126	0.3039	
5311	Parnassus on Wheels	62	0.3021	
24370	Mercenary	58	0.3011	
1123	Twelfth Night; Or, What You Will	41	0.3007	
22145	A Book of Burlesques	80	0.2992	
11074	The Damned	241	0.2966	
1122	The Tragedy of Hamlet, Prince of...	42	0.2958	
1787	Hamlet	361	0.2936	
1423	No Thoroughfare	45	0.2909	
1118	Much Ado about Nothing	183	0.2896	
179	The Europeans	85	0.2895	
5341	Kilmeny of the Orchard	102	0.2886	
25770	The Dragon's Secret	84	0.2867	
32226	The Flower Princess	57	0.2839	
1142	Typhoon	219	0.2833	
791	The Princess	245	0.2812	
25767	Picture and Text: 1893	59	0.2810	
15454	Imperium in Imperio: A Study of ...	50	0.2809	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
1814	The Agony Column	48	0.2803	
5071	The Philanderer	63	0.2799	
18492	Star Surgeon	83	0.2799	
4050	Mates at Billabong	41	0.2792	
5090	I Will Repay	50	0.2789	
7947	The Diary of a U-boat Commander: ...	71	0.2781	
60	The Scarlet Pimpernel	710	0.2777	
3777	Tom Swift and His Electric Rifle...	70	0.2748	
2267	Othello	760	0.2710	
33979	Miscellaneous Aphorisms; The Sou...	52	0.2705	
8931	The Gem Collector	107	0.2698	
25585	Shakespeare, Ben Jonson, Beaumon...	42	0.2692	
18346	Null-ABC	67	0.2663	
17226	Emily Fox-Seton : Being "The Maki...	115	0.2650	
2268	Antony and Cleopatra	125	0.2642	
5977	Bound to Rise; Or, Up the Ladder	47	0.2641	
1283	Tom Swift and His Wizard Camera;...	52	0.2632	
9931	K	46	0.2626	
6574	Watchers of the Sky	63	0.2620	
21865	King Arthur and His Knights	61	0.2616	
526	Heart of Darkness	4,362	0.2597	
586	Religio Medici, Hydriotaphia, an...	71	0.2592	
20387	A Thin Ghost and Others	155	0.2581	
11012	The Autobiography of an Ex-Color...	509	0.2580	
9932	The Last Trail	117	0.2572	
42710	Bizarre	42	0.2571	
23810	At Fault	104	0.2566	
39682	The Idiot at Home	44	0.2558	
22132	Giants on the Earth	50	0.2547	
37431	Pride and Prejudice, a play foun...	111	0.2541	
1107	The Taming of the Shrew	383	0.2540	
1953	A Book of Strife in the Form of ...	83	0.2534	
19381	Among the Farmyard People	60	0.2533	
2870	Washington Square	285	0.2511	
21927	Short Cruises	84	0.2498	
3529	Letters Written During a Short R...	236	0.2492	
19027	The Revolt on Venus	64	0.2482	
17797	Memoir of Jane Austen	203	0.2472	
37172	In a Glass Darkly, v. 1/3	82	0.2470	
3185	Those Extraordinary Twins	42	0.2468	
36612	The Princess and Curdie	57	0.2466	
3825	Pygmalion	3,580	0.2446	
8677	Behind a Mask; or, a Woman's Power	134	0.2444	
652	Rasselas, Prince of Abyssinia	241	0.2441	
6313	Masterpieces of American Wit and...	46	0.2426	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
15422	Israel Potter : His Fifty Years o...	58	0.2410	
32953	Quest of the Golden Ape	80	0.2406	
1244	Love for Love: A Comedy	71	0.2401	
18970	Caves of Terror	80	0.2400	
20730	For the Sake of the School	82	0.2393	
24280	Endymion: A Poetic Romance	133	0.2391	
2295	Waifs and Strays: Part 1	52	0.2387	
1239	The Spirit of the Border: A Roma...	69	0.2383	
1424	Castle Rackrent	203	0.2381	
24025	The New Girl at St. Chad's: A St...	44	0.2380	
42	The Strange Case of Dr. Jekyll a...	4,908	0.2379	
31501	The Sensitive Man	146	0.2369	
12431	The Coquette, or, The History of...	54	0.2349	
14654	A Daughter of the Snows	51	0.2345	
32486	The Legion of Lazarus	61	0.2344	
709	The Princess and Curdie	146	0.2336	
2019	The Bat	47	0.2320	
35204	Sense of Obligation	54	0.2320	
953	Tom Swift and His Big Tunnel; Or...	50	0.2295	
29310	The Affair of the Brains	86	0.2289	
5312	Mother Goose in Prose	97	0.2289	
1508	The Taming of the Shrew	49	0.2288	
11935	Mysticism in English Literature	65	0.2284	
18753	The Space Pioneers	49	0.2278	
8188	The Mysterious Key and What It O...	83	0.2270	
1461	A Legend of Montrose	40	0.2269	
653	The Chimes : A Goblin Story of So...	174	0.2268	
1532	The Tragedy of King Lear	43	0.2267	
4037	Appreciations, with an Essay on ...	100	0.2253	
20988	Islands of Space	109	0.2252	
1908	Her Prairie Knight	43	0.2249	
2604	The Longest Journey	88	0.2249	
3158	Our American Cousin	85	0.2242	
7964	The Mystery of Cloomber	108	0.2237	
30427	The Lost Kafoozalum	74	0.2232	
1604	The Ebb-Tide: A Trio And Quartette	56	0.2225	
31598	The Egyptian Cat Mystery: A Rick...	46	0.2224	
7239	Men, Women, and Boats	57	0.2218	
6936	Robinson Crusoe — in Words of On...	56	0.2214	
24353	Wired Love: A Romance of Dots and...	76	0.2208	
23625	The Magic Pudding	133	0.2206	
25438	The Airlords of Han	83	0.2205	
43	The Strange Case of Dr. Jekyll a...	599	0.2204	
26740	The Picture of Dorian Gray	257	0.2203	
18817	Ralestone Luck	75	0.2194	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
1524	Hamlet, Prince of Denmark	2,329	0.2193	
219	Heart of Darkness	3,243	0.2184	
18361	Operation: Outer Space	81	0.2184	
7371	A Sicilian Romance	112	0.2180	
95	The Prisoner of Zenda	339	0.2169	
171	Charlotte Temple	337	0.2168	
1338	Selected Prose of Oscar Wilde	141	0.2157	
20532	Love Among the Chickens: A Story ...	100	0.2145	
10119	Adonais	47	0.2130	
3244	To Him That Hath: A Tale of the ...	89	0.2128	
2702	The Lion's Skin	41	0.2124	
1828	Chronicles of the Canongate, 1st...	44	0.2123	
6359	The English Mail-Coach and Joan ...	62	0.2119	
558	The Thirty-Nine Steps	989	0.2118	
21873	Planet of the Damned	189	0.2116	
25016	The House of Souls	362	0.2114	
24761	The Rivals: A Comedy	408	0.2102	
13135	Pardners	45	0.2099	
19090	Star Hunter	185	0.2086	
1794	King Lear	126	0.2082	
1128	The Tragedy of King Lear	548	0.2082	
33644	The Secret of the Ninth Planet	109	0.2078	
1153	The Chessmen of Mars	409	0.2077	
25024	The Night of the Long Knives	180	0.2069	
9380	A Nonsense Anthology	44	0.2068	
28346	Deathworld	287	0.2063	
29042	A Tangled Tale	142	0.2061	
678	The Cricket on the Hearth: A Fai...	77	0.2057	
20795	The Cricket on the Hearth	43	0.2055	
720	Almayer's Folly: A Story of an E...	67	0.2053	
10882	The Eagle's Shadow	47	0.2051	
1144	In the Cage	85	0.2048	
943	Misalliance	87	0.2036	
32161	Tangle Hold	42	0.2035	
8713	A Man of Means	141	0.2034	
11252	Martin Hewitt, Investigator	118	0.2033	
2060	The History of Caliph Vathek	202	0.2031	
28520	Forbidden Fruit: Luscious and exc...	2,716	0.2028	
1125	All's Well That Ends Well	42	0.2022	
14888	The Inheritors	44	0.2019	
1120	The Tragedy of Julius Caesar	496	0.2019	
1785	Julius Caesar	42	0.2017	
2761	Benita, an African romance	41	0.2009	
1472	In a German Pension	65	0.2006	
32420	A Yankee Flier with the R.A.F.	49	0.2004	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
356	Beyond the City	46	0.2003	
19474	Uller Uprising	87	0.2000	
11371	The Moorland Cottage	53	0.1994	
8995	What Katy Did Next	105	0.1992	
12345	Friday, the Thirteenth : A Novel	59	0.1972	
17866	Murder in the Gunroom	267	0.1965	
14534	Christmas with Grandma Elsie	42	0.1961	
2852	The Hound of the Baskervilles	3,358	0.1951	
6984	The Pothunters	131	0.1943	
4075	The Intrusion of Jimmy	90	0.1940	
3070	The Hound of the Baskervilles	549	0.1940	
8435	The Sturdy Oak : A composite Nove...	58	0.1931	
4081	The Alchemist	744	0.1931	
32563	The Lost Warship	63	0.1927	
2389	Bardelys the Magnificent : Being ...	48	0.1922	
296	The Cash Boy	65	0.1922	
966	Maid Marian	62	0.1918	
16740	The Busie Body	51	0.1914	
3638	The Devil's Disciple	100	0.1904	
174	The Picture of Dorian Gray	7,652	0.1893	
25449	The Young Castellan: A Tale of t...	44	0.1893	
32530	Armageddon—2419 A.D.	313	0.1893	
11228	The Marrow of Tradition	178	0.1879	
7308	The History of Mr. Polly	79	0.1869	
555	The Unbearable Bassington	87	0.1868	
25776	This Crowded Earth	163	0.1867	
16259	The Surprising Adventures of the...	62	0.1866	
572	The Great Big Treasury of Beatri...	307	0.1864	
27174	Captain Jim	55	0.1854	
19478	Four-Day Planet	60	0.1853	
20707	The Black Star Passes	147	0.1853	
2667	The Vicar of Wakefield	238	0.1852	
29827	The Life and Amours of the Beaut...	678	0.1852	
27924	Mugby Junction	67	0.1850	
954	Tom Swift and His War Tank; Or, ...	44	0.1847	
4082	The Barrier	49	0.1845	
1605	The Crock of Gold	92	0.1841	
3829	Love Among the Chickens	198	0.1836	
984	Who Was Who: 5000 B. C. to Date ...	127	0.1833	
611	Prester John	152	0.1826	
17180	The Riddle of the Frozen Flame	64	0.1825	
10317	Betty Gordon at Boarding School;...	42	0.1822	
4020	Arcadian Adventures with the Idl...	55	0.1814	
4230	Tom Swift and His Motor-Cycle; O...	150	0.1810	
149	The Lost Continent	200	0.1810	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
17125	More William	85	0.1808	
2876	The Light That Failed	61	0.1807	
11247	The Exploits of Brigadier Gerard	98	0.1803	
6428	The Surgeon's Daughter	42	0.1797	
17047	The Half-Hearted	44	0.1796	
5747	Do and Dare — a Brave Boy's Figh...	61	0.1794	
1155	The Secret Adversary	2,070	0.1787	
19718	The Bostonians, Vol. II (of II)	42	0.1779	
8164	My Man Jeeves	1,317	0.1779	
2948	Where Angels Fear to Tread	202	0.1777	
436	The Master Key: An Electrical Fa...	61	0.1771	
2246	All's Well That Ends Well	50	0.1767	
37698	Dawn of the Morning	87	0.1742	
4078	The Picture of Dorian Gray	565	0.1742	
27903	The Magic World	67	0.1742	
7118	What Maisie Knew	236	0.1739	
545	At the Earth's Core	79	0.1733	
37503	Gammer Gurton's Needle	65	0.1723	
33623	The Inventions of the Idiot	60	0.1723	
8394	The Doings of Raffles Haw	40	0.1722	
32730	The Heart of a Woman	50	0.1719	
139	The Lost World	1,274	0.1717	
3475	The Efficiency Expert	53	0.1707	
14107	The Lost Stradivarius	84	0.1702	
836	The Phoenix and the Carpet	121	0.1700	
123	At the Earth's Core	296	0.1695	
3048	The Little Duke: Richard the Fea...	105	0.1693	
942	Green Mansions: A Romance of the...	83	0.1686	
6684	Uneasy Money	161	0.1683	
39957	Prairie Gold	103	0.1682	
39143	The Making of a Saint	65	0.1680	
19370	Ullr Uprising	51	0.1676	
33582	Rhyme? And Reason?	50	0.1675	
8183	Time and the Gods	123	0.1672	
4268	Cousin Phillis	51	0.1659	
16551	The Girl of the Golden West	44	0.1655	
1460	The Black Dwarf	59	0.1653	
1721	The Trees of Pride	45	0.1647	
6836	Three Men and a Maid	141	0.1643	
14257	The Magician	159	0.1632	
21092	On the Trail of the Space Pirates	54	0.1614	
5333	Every Man in His Humor	100	0.1612	
792	Wieland; Or, The Transformation:...	303	0.1608	
20857	Spacehounds of IPC	102	0.1606	
208	Daisy Miller: A Study	1,101	0.1603	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
25550	The Defiant Agents	225	0.1600	
23893	The Melting-Pot	172	0.1591	
1284	Tom Swift and His Air Scout; Or,...	40	0.1590	
18095	Successful Methods of Public Spe...	68	0.1589	
16328	Beowulf : An Anglo-Saxon Epic Poem	5,359	0.1584	
10422	Caesar Dies	40	0.1579	
10373	The Middle Temple Murder	134	0.1570	
24737	The Children of Odin: The Book o...	352	0.1568	
27444	Starman's Quest	127	0.1542	
29416	The Mind Master	46	0.1536	
20840	Rebel Spurs	146	0.1536	
40038	The Lone Ranger Rides	45	0.1531	
2046	Clotel; Or, The President's Daug...	144	0.1529	
4735	The Shepherd of the Hills	40	0.1529	
13969	The Hill of Dreams	131	0.1529	
102	The Tragedy of Pudd'nhead Wilson	1,140	0.1528	
10847	The Maids Tragedy	53	0.1526	
5342	The Story Girl	220	0.1521	
4531	The Secret Passage	41	0.1517	
1327	Elizabeth and Her German Garden	78	0.1515	
38551	The Crux: A Novel	43	0.1514	
4039	Volpone; Or, The Fox	558	0.1502	
14427	True Love's Reward : A Sequel to ...	57	0.1487	
770	The Story of the Treasure Seeker...	185	0.1487	
21932	Embarrassments	40	0.1485	
25870	A World of Girls: The Story of a...	49	0.1485	
20717	The Girl on the Boat	189	0.1485	
5232	Sejanus: His Fall	43	0.1476	
6418	Five Little Peppers and their Fr...	51	0.1475	
2815	Democracy, an American novel	67	0.1469	
20796	The Colors of Space	245	0.1466	
9909	Nightmare Abbey	180	0.1466	
14317	The Sorcery Club	91	0.1464	
9963	Elsie's Girlhood: A Sequel to "El...	58	0.1456	
805	This Side of Paradise	1,122	0.1455	
5815	The Great Impersonation	51	0.1450	
1457	Mistress Wilding	48	0.1444	
19141	Edison's Conquest of Mars	77	0.1436	
2548	The Poor Clare	85	0.1430	
31343	The Invaders	56	0.1429	
325	Phantastes: A Faerie Romance for...	461	0.1417	
5148	Rodney Stone	72	0.1413	
707	Raffles: Further Adventures of t...	105	0.1412	
1937	The Second Jungle Book	304	0.1402	
111	Freckles	149	0.1394	

Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

ID	Title	DL's	max(W[i,:])	Arc
9862	City of Endless Night	44	0.1383	
535	Travels with a Donkey in the Cev...	179	0.1381	
2726	Eight Cousins	214	0.1381	
20104	The Cross-Cut	42	0.1374	
5340	Further Chronicles of Avonlea	195	0.1372	
1696	The Club of Queer Trades	119	0.1363	
2042	Something New	384	0.1354	
981	Beowulf	718	0.1344	
37660	Of All Things	83	0.1340	
8223	Edgar Huntly; or, Memoirs of a S...	184	0.1337	
605	Pellucidar	231	0.1328	
479	Little Lord Fauntleroy	246	0.1321	
25067	The Planet Strappers	64	0.1315	
14228	Bracebridge Hall	46	0.1306	
17221	History of the Plague in London	82	0.1303	
13937	The Mysterious Rider	185	0.1296	
1091	On Heroes, Hero-Worship, and the...	622	0.1289	
2324	A House to Let	74	0.1286	
32620	The Three Mulla-mulgars	46	0.1279	
556	Rewards and Fairies	68	0.1276	
9791	Harrigan	56	0.1264	
11505	All Things Considered	485	0.1242	
7464	The Adventures of Sally	282	0.1237	
42243	The Hour of the Dragon	247	0.1237	
13054	A Thane of Wessex : Being a Story...	46	0.1235	
14034	King Alfred's Viking: A Story of ...	63	0.1230	
37820	Chronicles of Martin Hewitt	67	0.1229	
864	The Master of Ballantrae: A Wint...	93	0.1226	
10869	The Abandoned Room	67	0.1226	
11128	The Red Thumb Mark	155	0.1222	
37364	The Second Jungle Book	144	0.1211	
2885	The House of the Wolfings : A Tal...	105	0.1205	
2785	The Elusive Pimpernel	105	0.1202	
393	The Blue Lagoon: A Romance	129	0.1196	
20526	Short Story Writing: A Practical ...	41	0.1187	
38567	Eight Cousins; Or, The Aunt-Hill	41	0.1181	
20288	Edward the Second	204	0.1167	
764	Hans Brinker; Or, The Silver Skates	62	0.1163	
604	Gulliver of Mars	126	0.1158	
32542	Dave Dawson on Guadalcanal	50	0.1152	
1640	Lilith: A Romance	281	0.1143	
2687	The Snare	40	0.1137	
7230	Not George Washington — an Autob...	68	0.1131	
15281	Uncle Wiggily's Adventures	89	0.1127	
26998	Peter Pan in Kensington Gardens	201	0.1110	

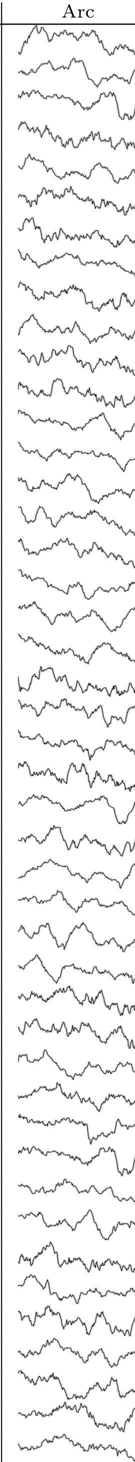


Table B.4: Stories which are are represented by core emotional arc 2, sorted by the variance explained in their emotional arc by core emotional arc 2.

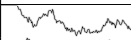
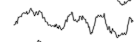
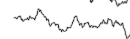
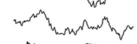
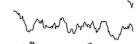
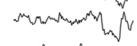
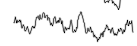
ID	Title	DL's	max(W[i,:])	Arc
26933	Visions and Revisions: A Book of ...	48	0.1102	
12170	The Wolf Hunters: A Tale of Adve...	66	0.1084	
832	Robin Hood	148	0.1077	
8092	Tremendous Trifles	352	0.1056	
1026	The Diary of a Nobody	329	0.1037	
33735	Pamela Censured	115	0.0998	
26494	Vera; Or, The Nihilists	47	0.0970	
1145	Rupert of Hentzau: From The Memo...	106	0.0947	
10556	The Old Man in the Corner	285	0.0917	
29466	Lords of the Stratosphere	51	0.0896	
26715	Victorian Songs: Lyrics of the A...	44	0.0894	
6877	The Head of Kay's	97	0.0867	
32202	The Irish Fairy Book	168	0.0759	
18824	Fairies and Folk of Ireland	62	0.0746	
11620	My Brilliant Career	62	0.0745	

Table B.5: Top 10 stories which are are represented by core emotional arc 2, sorted by downloads.

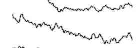
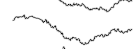
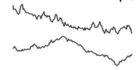
ID	Title	DL's	max(W[i,:])	Arc
174	The Picture of Dorian Gray	7,652	0.1893	
16328	Beowulf : An Anglo-Saxon Epic Poem	5,359	0.1584	
42	The Strange Case of Dr. Jekyll a...	4,908	0.2379	
526	Heart of Darkness	4,362	0.2597	
35	The Time Machine	3,732	0.3199	
3825	Pygmalion	3,580	0.2446	
2852	The Hound of the Baskervilles	3,358	0.1951	
219	Heart of Darkness	3,243	0.2184	
28520	Forbidden Fruit: Luscious and exc...	2,716	0.2028	

Table B.6: Stories which are are represented by core emotional arc 3, sorted by the variance explained in their emotional arc by core emotional arc 3.

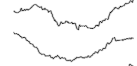
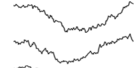
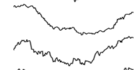
ID	Title	DL's	max(W[i,:])	Arc
419	The Magic of Oz	186	0.3150	
10736	Children of the Frost	82	0.3083	
1094	Tamburlaine the Great — Part 1	474	0.3081	
520	The Life and Adventures of Santa...	76	0.3041	
2911	Justice	50	0.2989	
18768	The Sky Is Falling	113	0.2978	
29774	A Yankee Flier Over Berlin	42	0.2864	
19726	The Door Through Space	201	0.2826	
4087	An Essay Upon Projects	101	0.2745	
28118	The Great Gray Plague	76	0.2706	
5083	The Man of Feeling	60	0.2664	
32746	The Revival of Irish Literature ...	54	0.2636	

Table B.6: Stories which are are represented by core emotional arc 3, sorted by the variance explained in their emotional arc by core emotional arc 3.

ID	Title	DL's	max(W[i,:])	Arc
37992	The King of Pirates : Being an Ac...	41	0.2635	
20559	R. Holmes & Co. : Being the Remar...	55	0.2625	
2814	Dubliners	4,742	0.2583	
21970	The Scarlet Plague	192	0.2512	
10671	The Botanic Garden. Part II.: Con...	78	0.2484	
18137	Little Fuzzy	171	0.2430	
28215	Empire	214	0.2391	
11435	Small Means and Great Ends	61	0.2390	
20519	Highways in Hiding	42	0.2356	
10337	Lady into Fox	53	0.2351	
687	A Personal Record	73	0.2334	
1282	Tom Swift Among the Diamond Make...	58	0.2288	
956	Tik-Tok of Oz	163	0.2275	
14280	Holidays at Roselands : A Sequel ...	48	0.2259	
534	An Inland Voyage	43	0.2216	
6440	Elsie Dinsmore	100	0.2207	
22031	The Airplane Boys among the Clou...	61	0.2197	
30742	Anything You Can Do!	41	0.2178	
6985	A Prefect's Uncle	145	0.2177	
24933	The Man Who Knew	68	0.2123	
20919	The Status Civilization	145	0.2097	
27129	Lyrics from the Song-Books of th...	42	0.2088	
901	The Jew of Malta	279	0.2082	
2607	Psmith, Journalist	242	0.2080	
27595	Eight Keys to Eden	59	0.2077	
19111	Code Three	46	0.2046	
47530	Oliver Twist, Vol. 2 (of 3)	40	0.2021	
16921	Plague Ship	218	0.2015	
1718	Manalive	120	0.2009	
20147	Rip Foster Rides the Gray Planet	40	0.1997	
9806	Mr. Justice Raffles	123	0.1983	
5347	Understood Betsy	159	0.1969	
1583	Options	63	0.1962	
24436	Anything You Can Do ...	58	0.1962	
126	The Poison Belt	268	0.1943	
4227	Tom Swift and His Wireless Messa...	41	0.1940	
4099	The Angel in the House	175	0.1914	
222	The Moon and Sixpence	352	0.1909	
9846	Excursions	110	0.1906	
18846	Voodoo Planet	148	0.1903	
37758	Atlantic Classics	49	0.1901	
5901	Dyke Darrel the Railroad Detecti...	44	0.1894	
1059	The World Set Free	343	0.1890	
897	The Rose and the Ring	49	0.1874	
2253	Henry V	266	0.1853	

Table B.6: Stories which are are represented by core emotional arc 3, sorted by the variance explained in their emotional arc by core emotional arc 3.

ID	Title	DL's	max(W[i,:])	Arc
27567	Aunt Jo's Scrap-Bag VI: An Old-Fa...	63	0.1831	
402	Penrod	71	0.1808	
17314	Five Children and It	190	0.1798	
2013	The Pit Prop Syndicate	40	0.1793	
2260	Titus Andronicus	52	0.1786	
55	The Wonderful Wizard of Oz	3,035	0.1770	
30431	Calumet 'K'	42	0.1762	
41715	Dave Dawson with the R.A.F.	49	0.1751	
1106	The Tragedy of Titus Andronicus	77	0.1750	
10966	The Ghost Pirates	234	0.1720	
20212	Police Your Planet	100	0.1697	
257	Troilus and Criseyde	236	0.1696	
13888	Bacon	54	0.1686	
14744	Different Girls	55	0.1683	
13897	The Adventure Club Afloat	117	0.1681	
14360	The Dawn and the Day : Or, The Bu...	46	0.1651	
778	Five Children and It	97	0.1647	
5182	The Old English Baron: a Gothic ...	89	0.1644	
3795	Under the Lilacs	72	0.1635	
16096	A Man's Woman	68	0.1631	
447	Maggie: A Girl of the Streets	1,295	0.1616	
1837	The Prince and the Pauper	1,389	0.1604	
13783	The Boy Inventors' Radio Telephone	52	0.1594	
14540	When William Came	42	0.1587	
21959	Letters from a Self-Made Merchan...	107	0.1582	
1654	An Unsocial Socialist	40	0.1581	
1845	Zuleika Dobson; Or, An Oxford Lo...	108	0.1555	
3006	Stalky & Co.	125	0.1552	
10066	Gunman's Reckoning	78	0.1551	
1013	The First Men in the Moon	348	0.1551	
24767	Jack O' Judgment	40	0.1544	
1809	Bucky O'Connor: A Tale of the Un...	49	0.1533	
18520	Sabotage in Space	48	0.1523	
2273	Tom Swift and His Motor-Boat; Or...	52	0.1513	
706	The Amateur Cracksman	182	0.1495	
20163	The Jolliest School of All	43	0.1489	
834	The Memoirs of Sherlock Holmes	2,164	0.1476	
34181	Irene Iddesleigh	138	0.1467	
22287	'Smiles': A Rose of the Cumberlands	75	0.1464	
9297	The Orange-Yellow Diamond	73	0.1457	
51	Anne of the Island	826	0.1451	
2014	The Lodger	97	0.1443	
1126	Measure for Measure	148	0.1436	
12436	The Night Horseman	87	0.1435	
1951	The Coming Race	350	0.1433	

Table B.6: Stories which are are represented by core emotional arc 3, sorted by the variance explained in their emotional arc by core emotional arc 3.

ID	Title	DL's	max(W[i,:])	Arc
2496	Our Village	41	0.1430	
1159	Fire-Tongue	50	0.1424	
113	The Secret Garden	1,153	0.1416	
5162	Agatha Webb	41	0.1412	
8920	The Light of Asia	67	0.1411	
25102	Nobody's Boy: Sans Famille	60	0.1405	
17396	The Secret Garden	716	0.1394	
4368	Flappers and Philosophers	287	0.1386	
7498	Five Little Peppers Grown Up	88	0.1383	
2454	The Silent Bullet	43	0.1380	
847	Lays of Ancient Rome	259	0.1378	
6340	Literary Lapses	70	0.1362	
19651	Key Out of Time	196	0.1361	
24723	Final Weapon	67	0.1353	
3785	In the Reign of Terror: The Adve...	40	0.1341	
2175	You Never Can Tell	124	0.1323	
31619	The Planet Savers	145	0.1322	
9609	Joseph Andrews, Vol. 2	57	0.1307	
40504	Ginger-Snaps	41	0.1300	
14203	Varied Types	41	0.1289	
420	Dorothy and the Wizard in Oz	385	0.1286	
2524	My Lady Ludlow	58	0.1277	
421	Kidnapped	1,132	0.1274	
5066	The Whole Family: a Novel by Twe...	50	0.1266	
20989	'A Comedy of Errors' in Seven Acts	41	0.1261	
5230	The Invisible Man: A Grotesque R...	1,011	0.1250	
486	Ozma of Oz	268	0.1243	
7353	Birds in Town & Village	145	0.1226	
2776	The Four Million	255	0.1225	
364	The Mad King	101	0.1216	
21775	The Best of the World's Classics...	119	0.1209	
2722	Morning Star	58	0.1185	
1028	The Professor	223	0.1151	
10743	Moonfleet	258	0.1099	
85	The Beasts of Tarzan	227	0.1097	
30339	Status Quo	42	0.1094	
4272	The Christian Year	46	0.1072	
11195	Alcatraz	53	0.1062	
35425	The Mad Planet	74	0.1024	
213	The Man from Snowy River	73	0.1013	
22463	Chivalry	53	0.0975	
2515	Stepping Heavenward	54	0.0929	
7028	The Clicking of Cuthbert	138	0.0874	
12	Through the Looking-Glass	2,892	0.0742	
28700	Robin Hood	58	0.0717	

Table B.6: Stories which are are represented by core emotional arc 3, sorted by the variance explained in their emotional arc by core emotional arc 3.

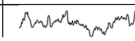
ID	Title	DL's	max(W[i,:])	Arc
2244	As You Like It	65	0.0711	

Table B.7: Top 10 stories which are are represented by core emotional arc 3, sorted by downloads.

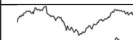
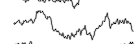
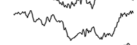
ID	Title	DL's	max(W[i,:])	Arc
2814	Dubliners	4,742	0.2583	
55	The Wonderful Wizard of Oz	3,035	0.1770	
12	Through the Looking-Glass	2,892	0.0742	
834	The Memoirs of Sherlock Holmes	2,164	0.1476	
1837	The Prince and the Pauper	1,389	0.1604	
447	Maggie: A Girl of the Streets	1,295	0.1616	
113	The Secret Garden	1,153	0.1416	
421	Kidnapped	1,132	0.1274	
5230	The Invisible Man: A Grotesque R...	1,011	0.1250	

Table B.8: Stories which are are represented by core emotional arc 4, sorted by the variance explained in their emotional arc by core emotional arc 4.

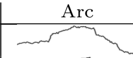
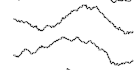
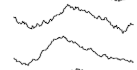
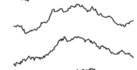
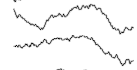
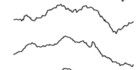
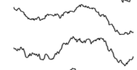
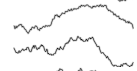
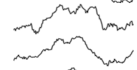
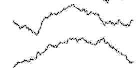
ID	Title	DL's	max(W[i,:])	Arc
34215	Shadowings	63	0.4141	
12384	Battle-Pieces and Aspects of the...	194	0.3546	
36281	The Slayer of Souls	63	0.3283	
17412	The Bobbsey Twins : Or, Merry Day...	69	0.3259	
2727	Allan's Wife	128	0.3124	
27761	Hamlet, Prince of Denmark	301	0.3057	
363	The Oakdale Affair	91	0.2968	
17854	The Sport of the Gods	61	0.2855	
27726	Tolstoy on Shakespeare: A Critic...	212	0.2764	
1292	The Way of the World	367	0.2722	
4381	The Aran Islands	63	0.2644	
19355	A Book of Prefaces	55	0.2621	
19	The Song of Hiawatha	297	0.2546	
1526	Twelfth Night; Or, What You Will	292	0.2531	
1527	Twelfth Night; Or, What You Will	90	0.2475	
11696	The Food of the Gods and How It ...	206	0.2469	
5829	The Moneychangers	46	0.2456	
20912	The Daffodil Mystery	148	0.2415	
551	The Land That Time Forgot	278	0.2412	
11045	The Ghost Ship	79	0.2411	
2040	Confessions of an English Opium-...	643	0.2382	
20431	The Tale of Beowulf, Sometime Ki...	80	0.2371	
3329	Caesar and Cleopatra	105	0.2360	
11127	The Case of Jennie Brice	43	0.2359	
24459	The Lost Princess of Oz	66	0.2345	
11451	The Rome Express	162	0.2314	

Table B.8: Stories which are are represented by core emotional arc 4, sorted by the variance explained in their emotional arc by core emotional arc 4.

ID	Title	DL's	max(W[i,:])	Arc
30368	A Christmas Carol: The original m...	143	0.2276	
13716	A Trip to Venus: A Novel	44	0.2267	
28164	The Big Bow Mystery	61	0.2248	
959	The Lost Princess of Oz	178	0.2242	
170	The Haunted Hotel: A Mystery of ...	162	0.2222	
6678	Nonsenseorship	99	0.2219	
2240	Much Ado about Nothing	289	0.2180	
39378	Mortal Coils	114	0.2180	
19337	A Christmas Carol	622	0.2171	
5210	The Borough	73	0.2169	
1163	Adventure	62	0.2163	
20856	Ten From Infinity	47	0.2162	
22332	Brain Twister	89	0.2149	
40603	The Root of All Evil	44	0.2147	
2317	The Story of My Heart: An Autobi...	64	0.2139	
32664	Black Amazon of Mars	142	0.2124	
775	When the Sleeper Wakes	151	0.2123	
22549	Space Prison	124	0.2122	
2713	Maiwa's Revenge; Or, The War of ...	56	0.2099	
18639	The Victorian Age in Literature	97	0.2072	
2062	All for Love; Or, The World Well...	239	0.2053	
24	O Pioneers!	371	0.2048	
4011	Epicoene; Or, The Silent Woman	123	0.2038	
1103	King Richard III	384	0.2014	
1264	The Wheels of Chance: A Bicyclin...	47	0.1989	
35533	The Haunted Room: A Tale	53	0.1989	
5131	Childe Harold's Pilgrimage	1,481	0.1987	
6995	Ghosts I Have Met and Some Others	65	0.1985	
19860	The Arabian Nights Entertainments	470	0.1981	
5070	The Doctor's Dilemma	113	0.1978	
29965	Two Thousand Miles Below	47	0.1975	
1115	The First Part of King Henry the...	70	0.1964	
2257	Richard III	49	0.1963	
33642	Earth Alert!	97	0.1958	
28434	The Astronomy of Milton's 'Parad...	44	0.1931	
134	Maria; Or, The Wrongs of Woman	271	0.1929	
8914	England, My England	67	0.1927	
46	A Christmas Carol in Prose; Bein...	4,602	0.1917	
13650	Nonsense Books	282	0.1913	
1537	Pericles, Prince of Tyre	67	0.1909	
22767	Pagan Passions	94	0.1899	
5795	The Secret Rose	56	0.1893	
888	The Lazy Tour of Two Idle Appren...	47	0.1888	
131	The Pilgrim's Progress from this...	1,126	0.1888	
3776	The Valley of Fear	182	0.1884	

Table B.8: Stories which are are represented by core emotional arc 4, sorted by the variance explained in their emotional arc by core emotional arc 4.

ID	Title	DL's	max(W[i,:])	Arc
37173	In a Glass Darkly, v. 2/3	47	0.1879	
24022	A Christmas Carol	188	0.1873	
22057	Kid Wolf of Texas : A Western Story	53	0.1869	
34426	The Enchanted Barn	81	0.1864	
1240	The Playboy of the Western World...	317	0.1864	
28522	Laura Middleton; Her Brother and...	1,097	0.1851	
3289	The Valley of Fear	1,228	0.1845	
18934	My Lady Nicotine: A Study in Smoke	60	0.1845	
13815	The Talking Beasts: A Book of Fa...	65	0.1797	
1585	The Wrong Box	60	0.1797	
22527	Beyond the Vanishing Point	45	0.1796	
32415	The Nursery Rhymes of England	121	0.1789	
68	Warlord of Mars	571	0.1779	
4737	A Tale of a Tub	277	0.1775	
8086	Down and Out in the Magic Kingdom	136	0.1769	
30537	The Royal Book of Oz : In which t...	64	0.1745	
21633	The Man of the Desert	54	0.1719	
2081	The Blithedale Romance	229	0.1719	
20606	The Magic City	118	0.1700	
40493	The King of Diamonds: A Tale of ...	48	0.1699	
26649	Terribly Intimate Portraits	46	0.1680	
7884	In the Fog	64	0.1680	
1097	Mrs. Warren's Profession	780	0.1679	
3479	The Metal Monster	77	0.1660	
33028	Man and Maid	40	0.1659	
22338	The Impossibles	56	0.1655	
40284	The Sex Life of the Gods	197	0.1651	
14667	A Christmas Garland	56	0.1645	
3179	The American Claimant	45	0.1639	
6120	Soldiers Three	45	0.1636	
1720	The Man Who Knew Too Much	310	0.1628	
30905	The Boarded-Up House	62	0.1627	
19258	Tom Swift and the Electronic Hyd...	70	0.1624	
30333	Daddy's Girl	52	0.1623	
4552	The Border Legion	96	0.1614	
22754	Masters of Space	173	0.1604	
17112	Many Thoughts of Many Minds: A Tr...	47	0.1579	
32498	The Brain	87	0.1548	
32398	Brood of the Dark Moon : (A Seque...	46	0.1542	
34403	The Clock Strikes Thirteen	43	0.1527	
25051	Space Platform	72	0.1523	
1711	Child of Storm	109	0.1511	
54	The Marvelous Land of Oz	419	0.1482	
20	Paradise Lost	2,522	0.1475	
7365	If I May	63	0.1474	

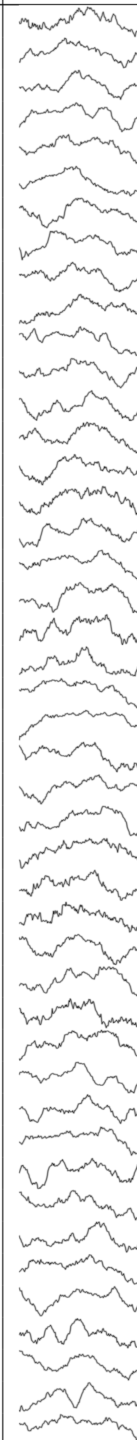


Table B.8: Stories which are are represented by core emotional arc 4, sorted by the variance explained in their emotional arc by core emotional arc 4.

ID	Title	DL's	max(W[i,:])	Arc
854	A Woman of No Importance	729	0.1455	
21048	Just Patty	47	0.1451	
6927	The White Feather	149	0.1427	
92	Tarzan and the Jewels of Opar	168	0.1404	
369	The Outlaw of Torn	129	0.1380	
2251	Henry IV, Part 1	76	0.1375	
2183	Three Men on the Bummel	190	0.1362	
32117	Eleven Possible Cases	57	0.1359	
9871	The Avenger	45	0.1355	
238	Dear Enemy	129	0.1354	
20698	The Story of Glass	53	0.1350	
3464	Tish: The Chronicle of Her Escap...	63	0.1349	
35545	Sanders of the River	127	0.1338	
32351	Voyage To Eternity	42	0.1333	
21051	Skylark Three	157	0.1331	
10581	Uncle Bernac: A Memory of the Em...	41	0.1322	
5746	The Ancient Allan	101	0.1319	
26197	The Nursery Rhyme Book	121	0.1301	
5660	Mary Louise	59	0.1282	
2306	Uncle Remus, His Songs and His S...	177	0.1275	
26654	Peter and Wendy	1,068	0.1275	
37193	The Swedish Fairy Book	57	0.1273	
1109	Love's Labour's Lost	56	0.1269	
3618	Arms and the Man	536	0.1252	
9834	The Talleyrand Maxim	74	0.1242	
26999	Peter Pan in Kensington Gardens	60	0.1238	
470	Heretics	395	0.1236	
10586	Mike and Psmith	170	0.1231	
37189	The Return of the Soldier	108	0.1223	
19079	The Adventures of Lightfoot the ...	40	0.1172	
12793	Cobwebs from an Empty Skull	56	0.1136	
12753	The Legends of King Arthur and H...	640	0.1125	
38562	The Big Book of Nursery Rhymes	82	0.1113	
1644	The Adventures of Gerard	100	0.1103	
2020	Tarzan the Terrible	185	0.1096	
498	Rebecca of Sunnybrook Farm	155	0.1083	
9190	The Greater Inclination	41	0.1068	
12239	Dead Men's Money	160	0.1063	
376	A Journal of the Plague Year : Wr...	461	0.1057	
25439	Looking Backward: 2000-1887	111	0.1044	
62	A Princess of Mars	2,515	0.1032	
2540	Father and Son: A Study of Two T...	74	0.1021	
166	Summer	165	0.1011	
18458	Star Born	160	0.0989	
38703	The Black Moth: A Romance of the...	204	0.0970	

Table B.8: *Stories which are are represented by core emotional arc 4, sorted by the variance explained in their emotional arc by core emotional arc 4.*

ID	Title	DL's	max(W[i,:])	Arc
72	Thuvia, Maid of Mars	437	0.0939	
972	The Devil's Dictionary	1,257	0.0936	
804	A Sentimental Journey Through Fr...	261	0.0921	
157	Daddy-Long-Legs	531	0.0895	
4993	A Texas Ranger	61	0.0868	
17985	Tom Swift and The Visitor from P...	98	0.0832	
6093	Far Away and Long Ago: A History...	70	0.0818	
308	Three Men in a Boat	2,059	0.0815	
22064	Tess of the Storm Country	83	0.0814	
26240	The Clansman: An Historical Roma...	85	0.0728	

Table B.9: *Top 10 stories which are are represented by core emotional arc 4, sorted by downloads.*

ID	Title	DL's	max(W[i,:])	Arc
46	A Christmas Carol in Prose; Bein...	4,602	0.1917	
20	Paradise Lost	2,522	0.1475	
62	A Princess of Mars	2,515	0.1032	
308	Three Men in a Boat	2,059	0.0815	
5131	Childe Harold's Pilgrimage	1,481	0.1987	
972	The Devil's Dictionary	1,257	0.0936	
3289	The Valley of Fear	1,228	0.1845	
131	The Pilgrim's Progress from this...	1,126	0.1888	
28522	Laura Middleton; Her Brother and...	1,097	0.1851	

Table B.10: *Stories which are are represented by core emotional arc 5, sorted by the variance explained in their emotional arc by core emotional arc 5.*

ID	Title	DL's	max(W[i,:])	Arc
17763	The Mystery of the Hasty Arrow	93	0.2301	
5317	Through the Magic Door	81	0.2299	
13944	After London; Or, Wild England	146	0.2259	
12590	The Shadow of the Rope	75	0.2202	
35247	That Affair at Elizabeth	62	0.2179	
39827	The Conduct of Life	176	0.2041	
295	The Early Short Fiction of Edith...	49	0.2012	
10459	The Celtic Twilight	146	0.1974	
957	The Scarecrow of Oz	162	0.1972	
19819	Milton's Comus	41	0.1969	
644	The Haunted Man and the Ghost's ...	97	0.1915	
23624	Ride Proud, Rebel!	139	0.1882	
27771	Once on a Time	102	0.1858	
13694	Mince Pie	46	0.1811	
40386	Wandering Ghosts	125	0.1809	
32440	Dave Dawson at Dunkirk	60	0.1786	
1146	The Journal of a Voyage to Lisbon	57	0.1776	

Table B.10: Stories which are are represented by core emotional arc 5, sorted by the variance explained in their emotional arc by core emotional arc 5.

ID	Title	DL's	max(W[i,:])	Arc
25780	The Fire People	45	0.1768	
306	The Early Short Fiction of Edith...	44	0.1738	
27780	Treasure Island	281	0.1712	
8994	What Katy Did	142	0.1700	
20788	Storm Over Warlock	154	0.1687	
24680	The Martyr of the Catacombs: A Ta...	43	0.1664	
30759	Exit Betty	49	0.1664	
1212	Love and Freindship [sic]	611	0.1657	
2166	King Solomon's Mines	788	0.1652	
120	Treasure Island	4,402	0.1620	
20782	Triplanetary	131	0.1599	
11068	The Spirit of the Age; Or, Conte...	65	0.1570	
330	Where There's a Will	64	0.1566	
618	Codex Junius 11	47	0.1553	
39592	Princess Mary's Gift Book : All p...	71	0.1541	
5805	The League of the Scarlet Pimpernel	104	0.1505	
16255	Dickey Downy: The Autobiography ...	44	0.1496	
14917	The Wings of the Morning	41	0.1478	
19369	The Triumphs of Eugène Valmont	102	0.1477	
696	The Castle of Otranto	1,663	0.1465	
15585	Humorous Masterpieces from Ameri...	53	0.1462	
1164	The Iron Heel	506	0.1395	
26862	Howard Pyle's Book of Pirates : F...	81	0.1391	
423	Round the Red Lamp: Being Facts ...	136	0.1382	
1515	The Merchant of Venice	50	0.1353	
15323	The Green Eyes of Bâst	65	0.1329	
553	Out of Time's Abyss	140	0.1325	
18217	Chambers's Elementary Science Re...	62	0.1320	
1280	Spoon River Anthology	671	0.1297	
10723	Betty's Bright Idea; Deacon Pitk...	40	0.1281	
1550	A Lady of Quality : Being a Most ...	89	0.1261	
34339	The Princess and the Goblin	268	0.1254	
1114	The Merchant of Venice	134	0.1233	
2644	Isaac Bickerstaff, Physician and...	42	0.1191	
5830	A Garland for Girls	133	0.1185	
10601	The Rangeland Avenger	94	0.1177	
3005	Tom Swift and His Airship	70	0.1159	
18151	Time Crime	85	0.1158	
39281	Dictionary of English Proverbs a...	83	0.1149	
26019	Europa's Fairy Book	56	0.1144	
1182	Dope	94	0.1138	
37012	The Recruiting Officer	49	0.1114	
2809	Main-Travelled Roads	169	0.1105	
2024	Diary of a Pilgrimage	48	0.1071	
31356	The Man Who Staked the Stars	61	0.1054	

Table B.10: Stories which are are represented by core emotional arc 5, sorted by the variance explained in their emotional arc by core emotional arc 5.

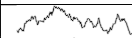
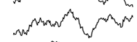
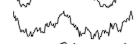
ID	Title	DL's	max(W[i,:])	Arc
2005	Piccadilly Jim	151	0.1039	
394	Cranford	285	0.1029	
36869	The Real Man	59	0.1027	
708	The Princess and the Goblin	579	0.0996	
39116	Unicorns	47	0.0969	
5803	Not that it Matters	138	0.0956	
22693	A Book of Myths	248	0.0875	
2305	A Set of Six	58	0.0862	
19717	The Bostonians, Vol. I (of II)	80	0.0817	
3688	The Chronicles of Clovis	186	0.0812	
2243	The Merchant of Venice	506	0.0787	

Table B.11: Top 10 stories which are are represented by core emotional arc 5, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
120	Treasure Island	4,402	0.1620	
696	The Castle of Otranto	1,663	0.1465	
2166	King Solomon's Mines	788	0.1652	
1280	Spoon River Anthology	671	0.1297	
1212	Love and Freindship [sic]	611	0.1657	
708	The Princess and the Goblin	579	0.0996	
1164	The Iron Heel	506	0.1395	
2243	The Merchant of Venice	506	0.0787	
394	Cranford	285	0.1029	

Table B.12: Stories which are are represented by core emotional arc 6, sorted by the variance explained in their emotional arc by core emotional arc 6.

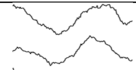
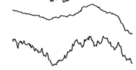
ID	Title	DL's	max(W[i,:])	Arc
18172	This World Is Taboo	64	0.3281	
339	Old Indian Days	139	0.3148	
10377	The Evil Guest	93	0.2868	
29448	Pariah Planet	96	0.2864	
289	The Wind in the Willows	1,475	0.2750	
20727	The Cosmic Computer	221	0.2716	
27805	The Wind in the Willows	543	0.2702	
27991	The Blue Bird for Children: The W...	111	0.2677	
96	The Monster Men	155	0.2346	
9156	Life and Remains of John Clare, ...	53	0.2299	
41027	The Revolt of the Star Men	51	0.2276	
34313	Literature in the Making, by Som...	54	0.2253	
209	The Turn of the Screw	2,175	0.2251	
32256	The Big Time	135	0.2229	
22354	The Adventures of Maya the Bee	68	0.2212	
654	Grace Abounding to the Chief of ...	177	0.2209	

Table B.12: Stories which are are represented by core emotional arc 6, sorted by the variance explained in their emotional arc by core emotional arc 6.

ID	Title	DL's	max(W[i,:])	Arc
20859	Wandl the Invader	49	0.2106	
310	Before Adam	94	0.2101	
958	Rinkitink in Oz : Wherein Is Reco...	124	0.2092	
10850	Philaster; Or, Love Lies a Bleeding	50	0.2091	
1725	Heart of the West	118	0.2090	
15798	Clover	102	0.2069	
25581	Rinkitink in Oz	41	0.2054	
1906	Erewhon; Or, Over the Range	251	0.2051	
20781	Heidi: (Gift Edition)	642	0.2033	
17731	The Nigger Of The "Narcissus": A...	207	0.2027	
9990	Brave and Bold; Or, The Fortunes...	56	0.2020	
22182	The Best of the World's Classics...	65	0.1972	
1915	The Second Thoughts of an Idle F...	88	0.1956	
244	A Study in Scarlet	4,535	0.1955	
9611	Joseph Andrews, Vol. 1	242	0.1939	
863	The Mysterious Affair at Styles	3,112	0.1937	
10324	Bull Hunter	58	0.1918	
1329	A Voyage to Arcturus	218	0.1897	
21768	A Desert Drama: Being The Traged...	42	0.1841	
34020	The Window at the White Cat	46	0.1837	
43984	Chaucer for Children: A Golden Key	119	0.1826	
1143	Notes on Life & Letters	53	0.1821	
17870	Operation Terror	59	0.1809	
12163	The Sleeper Awakes: A Revised Edi...	141	0.1759	
223	The Wisdom of Father Brown	563	0.1757	
20898	The Galaxy Primes	227	0.1727	
6683	The Little Nugget	95	0.1723	
1448	Heidi	268	0.1710	
30970	Miss Cayley's Adventures	45	0.1707	
950	Tom Swift and His Electric Runab...	44	0.1706	
1446	Perfect Behavior: A Guide for La...	59	0.1704	
90	The Son of Tarzan	212	0.1700	
23028	Greylorn	61	0.1691	
1129	The Tragedy of Macbeth	449	0.1690	
25003	The Nicest Girl in the School: A...	48	0.1689	
1948	The Story of a Bad Boy	45	0.1683	
15238	Mathilda	163	0.1673	
32037	Eureka: A Prose Poem	174	0.1669	
22892	The Best Made Plans	40	0.1664	
1450	Pollyanna	349	0.1646	
2154	Around the World in Eighty Days....	57	0.1642	
1795	Macbeth	73	0.1637	
2126	The Quest of the Sacred Slipper	55	0.1626	
19023	A Daughter of the Sioux: A Tale ...	61	0.1621	
84	Frankenstein; Or, The Modern Pro...	11,699	0.1602	

Table B.12: Stories which are are represented by core emotional arc 6, sorted by the variance explained in their emotional arc by core emotional arc 6.

ID	Title	DL's	max(W[i,:])	Arc
1874	The Railway Children	437	0.1595	
2512	The Cruise of the Snark	105	0.1583	
3790	Major Barbara	416	0.1581	
23845	Talents, Incorporated	66	0.1567	
30964	The Ethical Engineer	110	0.1566	
42324	Frankenstein; Or, The Modern Pro...	313	0.1565	
32934	The Young Colonists: A Story of ...	82	0.1565	
33505	The Trembling of the Veil	60	0.1561	
21639	When Patty Went to College	41	0.1558	
5632	Five Little Peppers Midway	58	0.1554	
19471	Badge of Infamy	105	0.1541	
32954	The Black Arrow: A Tale of the T...	115	0.1530	
2763	The World's Desire	45	0.1526	
2381	Actions and Reactions	92	0.1522	
463	The Red Badge of Courage: An Epi...	84	0.1514	
848	The Black Arrow: A Tale of Two R...	292	0.1496	
73	The Red Badge of Courage: An Epi...	1,163	0.1494	
32	Herland	1,013	0.1485	
329	Island Nights' Entertainments	67	0.1483	
1625	The Frozen Deep	51	0.1467	
2906	The Silver Box: A Comedy in Thre...	49	0.1463	
16865	Pinocchio: The Tale of a Puppet	234	0.1451	
41445	Frankenstein; Or, The Modern Pro...	786	0.1448	
39896	The Girl Next Door	141	0.1433	
20728	Space Viking	223	0.1375	
383	She Stoops to Conquer; Or, The M...	903	0.1359	
2509	The Lani People	66	0.1358	
15717	Books and Persons; Being Comment...	41	0.1356	
12215	Odd Craft, Complete	41	0.1355	
2851	Sixes and Sevens	85	0.1336	
10007	Carmilla	1,416	0.1333	
33348	Reveries over Childhood and Youth	44	0.1320	
620	Sylvie and Bruno	104	0.1318	
9746	The Ashiel mystery: A Detective ...	92	0.1303	
552	The People That Time Forgot	171	0.1302	
9656	Alarms and Discursions	64	0.1285	
3146	Two on a Tower	101	0.1284	
32331	Dave Dawson at Casablanca	47	0.1280	
24197	The Tinted Venus: A Farcical Rom...	40	0.1271	
4223	The Mystery of a Hansom Cab	43	0.1268	
12629	Language: An Introduction to the ...	210	0.1259	
2186	"Captains Courageous": A Story o...	72	0.1259	
33325	The Spoils of Poynton	41	0.1256	
19207	The Firelight Fairy Book	93	0.1211	
19672	The Holladay Case: A Tale	47	0.1204	

Table B.12: Stories which are are represented by core emotional arc 6, sorted by the variance explained in their emotional arc by core emotional arc 6.

ID	Title	DL's	max(W[i,:])	Arc
2097	The Sign of the Four	2,283	0.1199	
2487	Cross Roads	57	0.1198	
41049	The Onslaught from Rigel	60	0.1178	
18505	A Popular Schoolgirl	87	0.1168	
5606	Guns of the Gods: A Story of Yas...	50	0.1145	
26176	The Secret House	90	0.1139	
24880	The Wreck of the Titan: or, Futility	148	0.1138	
389	The Great God Pan	807	0.1127	
2276	The Private Memoirs and Confessi...	260	0.1121	
8457	Frenzied Fiction	45	0.1115	
4682	Nonsense Novels	103	0.1108	
4709	Brewster's Millions	131	0.1083	
20551	The White Invaders	53	0.1080	
36	The War of the Worlds	2,496	0.1060	
3026	North of Boston	151	0.1045	
10554	Right Ho, Jeeves	896	0.1004	
2662	Under the Greenwood Tree; Or, Th...	135	0.0993	
1881	The Call of the Canyon	192	0.0992	
32501	The Golden Age	44	0.0988	
1897	The Seventh Man	56	0.0976	
1183	The Return of Dr. Fu-Manchu	91	0.0976	
19142	The Devil Doctor	99	0.0971	
21626	Adrift in the Wilds; Or, The Adv...	43	0.0965	
20081	A Houseful of Girls	45	0.0947	
1354	Chronicles of Avonlea	258	0.0944	
2786	Jack and Jill	154	0.0940	
9867	Riders of the Silences	62	0.0905	
1267	Kai Lung's Golden Hours	94	0.0903	
1478	A Parody Outline of History : Whe...	60	0.0874	
19535	George Bernard Shaw	60	0.0870	
19246	The Young Pitcher	49	0.0867	
1589	Tamburlaine the Great — Part 2	104	0.0851	
34414	Just William	202	0.0716	

Table B.13: Top 10 stories which are are represented by core emotional arc 6, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
84	Frankenstein; Or, The Modern Pro...	11,699	0.1602	
244	A Study in Scarlet	4,535	0.1955	
863	The Mysterious Affair at Styles	3,112	0.1937	
36	The War of the Worlds	2,496	0.1060	
2097	The Sign of the Four	2,283	0.1199	
209	The Turn of the Screw	2,175	0.2251	
289	The Wind in the Willows	1,475	0.2750	
10007	Carmilla	1,416	0.1333	
73	The Red Badge of Courage: An Epi...	1,163	0.1494	

Table B.14: Stories which are are represented by core emotional arc 7, sorted by the variance explained in their emotional arc by core emotional arc 7.

ID	Title	DL's	max(W[i,:])	Arc
12028	The Uttermost Farthing: A Savant'...	58	0.2612	
19526	Stand by for Mars!	64	0.2425	
811	The Tragical History of Doctor F...	389	0.2245	
121	Northanger Abbey	2,355	0.2196	
7031	The Sheik: A Novel	152	0.2071	
24035	The Pirates of Ersatz	113	0.2017	
2429	Lost Face	190	0.1975	
74	The Adventures of Tom Sawyer	9,454	0.1906	
1058	The Mirror of the Sea	106	0.1885	
3674	The Dragon and the Raven; Or, Th...	151	0.1791	
26853	Vice Versa; or, A Lesson to Fathers	49	0.1787	
19709	Danger in Deep Space	74	0.1751	
20472	Grace Harlowe's Plebe Year at Hi...	40	0.1729	
18719	Space Tug	63	0.1721	
24929	The Green Rust	56	0.1709	
27826	The Olive Fairy Book	113	0.1670	
5962	Oh, Money! Money! A Novel	44	0.1660	
10886	The Untamed	109	0.1658	
18019	The Luckiest Girl in the School	107	0.1649	
6622	Legends That Every Child Should ...	144	0.1612	
2686	The Book of Snobs	204	0.1589	
19330	An Apache Princess: A Tale of th...	69	0.1576	
21891	The Brand of Silence: A Detective...	89	0.1568	
15580	The Rustlers of Pecos County	90	0.1537	
24283	Down the River; Or, Buck Bradfor...	42	0.1529	
6382	Bat Wing	92	0.1511	
10542	The Boats of the "Glen Carrig" : ...	126	0.1482	
20154	Invaders from the Infinite	103	0.1429	
21964	The Short-story	182	0.1414	
8771	Jurgen: A Comedy of Justice	95	0.1407	
11583	The Runaway Asteroid	43	0.1384	
8492	The King in Yellow	1,504	0.1354	
13029	The Art of the Moving Picture	95	0.1317	
13675	Goody Two-Shoes : A Facsimile Rep...	90	0.1307	
10234	Old Creole Days: A Story of Creo...	80	0.1277	
1281	Tom Swift and His Aerial Warship...	45	0.1261	
1695	The Man Who Was Thursday: A Nigh...	796	0.1238	
2225	"Captains Courageous": A Story o...	212	0.1232	
910	White Fang	1,178	0.1223	
4517	Ethan Frome	2,895	0.1220	
16	Peter Pan	5,789	0.1197	
146	A Little Princess : Being the who...	825	0.1189	
6100	Pollyanna Grows Up	107	0.1140	
24160	The Basket of Flowers	41	0.1095	
1154	The Voyages of Doctor Dolittle	179	0.1081	

Table B.14: *Stories which are are represented by core emotional arc 7, sorted by the variance explained in their emotional arc by core emotional arc 7.*

ID	Title	DL's	max(W[i,:])	Arc
14632	The Mystery of Mary	52	0.1075	
21510	Legacy	101	0.1069	
37332	A Little Princess: Being the who...	142	0.1066	
955	The Patchwork Girl of Oz	172	0.1030	
242	My Antonia	847	0.1024	
1376	The Little White Bird; Or, Adven...	250	0.1014	
47	Anne of Avonlea	803	0.1000	
12491	Twelve Types	71	0.0997	
19810	My Ántonia	171	0.0962	
1204	Cabin Fever	64	0.0947	
23292	Ted and the Telephone	95	0.0942	
15625	The Lookout Man	51	0.0935	
25388	The Herapath Property	41	0.0929	
41231	The Life and Beauties of Fanny Fern	49	0.0907	
887	Intentions	173	0.0890	
23661	The Book of Dragons	254	0.0883	
81	The Return of Tarzan	384	0.0864	
14875	Elsie's children	41	0.0861	
2861	The Sleuth of St. James's Square	76	0.0848	
21715	Away in the Wilderness	53	0.0789	
5604	Getting Married	105	0.0727	

Table B.15: *Top 10 stories which are are represented by core emotional arc 7, sorted by downloads.*

ID	Title	DL's	max(W[i,:])	Arc
74	The Adventures of Tom Sawyer	9,454	0.1906	
16	Peter Pan	5,789	0.1197	
4517	Ethan Frome	2,895	0.1220	
121	Northanger Abbey	2,355	0.2196	
8492	The King in Yellow	1,504	0.1354	
910	White Fang	1,178	0.1223	
242	My Antonia	847	0.1024	
146	A Little Princess : Being the who...	825	0.1189	
47	Anne of Avonlea	803	0.1000	

Table B.16: *Stories which are are represented by core emotional arc 8, sorted by the variance explained in their emotional arc by core emotional arc 8.*

ID	Title	DL's	max(W[i,:])	Arc
19360	Six to Sixteen: A Story for Girls	47	0.2415	
28849	Smugglers' Reef: A Rick Brant Sc...	46	0.2320	
18581	Adrift in New York: Tom and Flor...	111	0.2182	
5141	What Katy Did at School	85	0.2065	
4217	A Portrait of the Artist as a Yo...	2,172	0.2039	
31308	Orientations	76	0.1924	

Table B.16: Stories which are are represented by core emotional arc 8, sorted by the variance explained in their emotional arc by core emotional arc 8.

ID	Title	DL's	max(W[i,:])	Arc
35117	Lord Tony's Wife: An Adventure o...	66	0.1748	
4715	An African Millionaire: Episodes...	163	0.1714	
4731	Seven Little Australians	79	0.1679	
16721	A Place so Foreign	54	0.1654	
32242	A Wonder Book for Girls & Boys	192	0.1588	
496	The Little Lame Prince	48	0.1577	
14154	The Tale of Terror: A Study of t...	63	0.1570	
26	Paradise Lost	730	0.1565	
20739	Rebels of the Red Planet	47	0.1511	
4253	Dramatic Romances	71	0.1507	
23641	The Forsaken Inn: A Novel	59	0.1496	
64	The Gods of Mars	628	0.1481	
37174	In a Glass Darkly, v. 3/3	46	0.1473	
12187	The Mystery of 31 New Inn	188	0.1464	
29405	The Gods of Mars	158	0.1441	
2688	The Clue of the Twisted Candle	237	0.1402	
7434	The Adventures of Joel Pepper	47	0.1376	
204	The Innocence of Father Brown	800	0.1357	
980	Alice Adams	48	0.1338	
11626	The Dawn of All	57	0.1319	
2028	The Yellow Claw	73	0.1286	
849	The Idle Thoughts of an Idle Fellow	177	0.1281	
9903	Way of the Lawless	68	0.1275	
4352	Laughter: An Essay on the Meanin...	365	0.1251	
2568	Trent's Last Case	193	0.1248	
472	The House Behind the Cedars	107	0.1246	
36775	Humorous Readings and Recitation...	136	0.1238	
21073	A Pirate of the Caribbees	49	0.1230	
35304	The Last Stroke: A Detective Story	123	0.1214	
8446	The Enormous Room	232	0.1195	
36958	A Child of the Jago	61	0.1193	
41667	The Emerald City of Oz	67	0.1163	
24201	The Eye of Osiris	91	0.1103	
517	The Emerald City of Oz	266	0.1097	
21854	The Woman in Black	93	0.1085	
5670	Jacob's Room	403	0.1013	
10476	The Vanishing Man : A Detective R...	89	0.0992	
11666	The Conjure Woman	203	0.0898	
225	At the Back of the North Wind	288	0.0889	
40241	Hieroglyphics	50	0.0871	
32587	The Ambassador	71	0.0865	
6987	Five Little Peppers Abroad	53	0.0829	
6880	The Coming of Bill	95	0.0799	
32884	Ideas of Good and Evil	109	0.0737	

Table B.17: *Top 10 stories which are are represented by core emotional arc 8, sorted by downloads.*

ID	Title	DL's	max(W[i,:])	Arc
4217	A Portrait of the Artist as a Yo...	2,172	0.2039	
204	The Innocence of Father Brown	800	0.1357	
26	Paradise Lost	730	0.1565	
64	The Gods of Mars	628	0.1481	
5670	Jacob's Room	403	0.1013	
4352	Laughter: An Essay on the Meanin...	365	0.1251	
225	At the Back of the North Wind	288	0.0889	
517	The Emerald City of Oz	266	0.1097	
2688	The Clue of the Twisted Candle	237	0.1402	

Table B.18: *Stories which are are represented by core emotional arc 9, sorted by the variance explained in their emotional arc by core emotional arc 9.*

ID	Title	DL's	max(W[i,:])	Arc
19066	Brigands of the Moon	65	0.2113	
17959	The Hand Of Fu-Manchu: Being a Ne...	112	0.1861	
19145	The Time Traders	225	0.1801	
2057	The Last of the Plainsmen	106	0.1681	
1188	The Lair of the White Worm	206	0.1632	
546	Under the Andes	87	0.1544	
18668	In Search of the Unknown	159	0.1489	
6955	The Prince and Betty	85	0.1488	
19706	Brood of the Witch-Queen	100	0.1484	
8899	Three Weeks	48	0.1457	
2098	A Thief in the Night: A Book of ...	112	0.1426	
1872	The Red House Mystery	422	0.1421	
41753	Dave Dawson at Truk	42	0.1405	
9415	Olaf the Glorious: A Story of th...	42	0.1365	
1999	Crome Yellow	210	0.1319	
24313	Once a Week	49	0.1311	
9807	Scarhaven Keep	68	0.1309	
34219	The Enchanted Castle	61	0.1296	
8730	A Little Bush Maid	49	0.1273	
25344	The Scarlet Letter	386	0.1266	
17667	Dialogues of the Dead	51	0.1232	
3536	The Enchanted Castle	220	0.1227	
33	The Scarlet Letter	3,045	0.1195	
25564	The Water-Babies: A Fairy Tale f...	103	0.1193	
3328	Man and Superman: A Comedy and a...	312	0.1181	
12986	The Card, a Story of Adventure i...	54	0.1169	
2775	The Good Soldier	426	0.1144	
21656	The Princess of the School	56	0.1096	
1533	Macbeth	165	0.1088	
1051	Sartor Resartus: The Life and Op...	347	0.1074	
30324	The Pathless Trail	65	0.1069	
19307	The Lion of Petra	41	0.1044	

Table B.18: *Stories which are are represented by core emotional arc 9, sorted by the variance explained in their emotional arc by core emotional arc 9.*

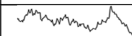
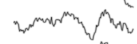
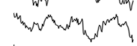
ID	Title	DL's	max(W[i,:])	Arc
11	Alice's Adventures in Wonderland	17,366	0.0988	
1751	Twilight Land	48	0.0985	
5121	Dark Hollow	43	0.0978	
22234	Aunt Jo's Scrap-Bag, Vol. 5: Jimm...	69	0.0975	
751	The Autocrat of the Breakfast-Table	55	0.0960	
794	The Wouldbegoods: Being the Furt...	50	0.0956	
2777	Cabbages and Kings	209	0.0919	
1077	The Mirror of Kong Ho	45	0.0884	
271	Black Beauty	780	0.0870	
28885	Alice's Adventures in Wonderland...	1,051	0.0862	
5308	The Paradise Mystery	94	0.0862	
1262	The Heritage of the Desert: A Novel	75	0.0841	
1590	The Amazing Interlude	43	0.0823	
9196	The Clockmaker; Or, the Sayings ...	51	0.0708	

Table B.19: *Top 10 stories which are are represented by core emotional arc 9, sorted by downloads.*

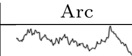
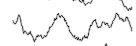
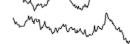
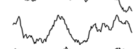
ID	Title	DL's	max(W[i,:])	Arc
11	Alice's Adventures in Wonderland	17,366	0.0988	
33	The Scarlet Letter	3,045	0.1195	
28885	Alice's Adventures in Wonderland...	1,051	0.0862	
271	Black Beauty	780	0.0870	
2775	The Good Soldier	426	0.1144	
1872	The Red House Mystery	422	0.1421	
25344	The Scarlet Letter	386	0.1266	
1051	Sartor Resartus: The Life and Op...	347	0.1074	
3328	Man and Superman: A Comedy and a...	312	0.1181	

Table B.20: *Stories which are are represented by core emotional arc 10, sorted by the variance explained in their emotional arc by core emotional arc 10.*

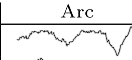
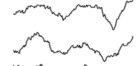
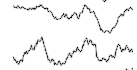
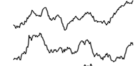
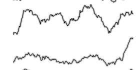
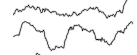
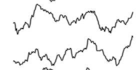
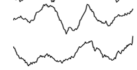
ID	Title	DL's	max(W[i,:])	Arc
21374	!Tention: A Story of Boy-Life du...	92	0.2195	
24499	The Green Carnation	65	0.1986	
544	Anne's House of Dreams	586	0.1676	
4922	Bar-20 Days	57	0.1648	
8681	The Face and the Mask	82	0.1528	
1805	The Gentle Grafter	84	0.1505	
10443	The Rayner-Slade Amalgamation	61	0.1486	
16339	The Passenger from Calais	125	0.1454	
24775	Up the River; or, Yachting on th...	88	0.1449	
9902	The Middle of Things	84	0.1327	
25305	Memoirs Of Fanny Hill: A New and ...	2,222	0.1296	
37858	Leaves in the Wind	79	0.1244	
24770	A Prisoner of Morro; Or, In the ...	45	0.1243	

Table B.20: Stories which are are represented by core emotional arc 10, sorted by the variance explained in their emotional arc by core emotional arc 10.

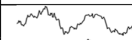
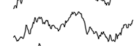
ID	Title	DL's	max(W[i,:])	Arc
25866	The Search	56	0.1209	
25919	Miss Mapp	49	0.1156	
35517	The Three Impostors; or, The Tra...	111	0.1148	
1020	Sword Blades and Poppy Seed	75	0.1141	
1535	The Tragedy of Coriolanus	116	0.1137	
38006	The Heatherford Fortune: a sequel...	44	0.1122	
38777	Lad: A Dog	58	0.1059	
38070	The Norwegian Fairy Book	149	0.1044	
1358	Enoch Arden, &c.	54	0.1042	
18613	The Golden Scorpion	84	0.0954	
4540	In His Steps	230	0.0902	
34943	Among the Meadow People	48	0.0727	

Table B.21: Top 10 stories which are are represented by core emotional arc 10, sorted by downloads.

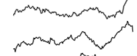
ID	Title	DL's	max(W[i,:])	Arc
25305	Memoirs Of Fanny Hill: A New and ...	2,222	0.1296	
544	Anne's House of Dreams	586	0.1676	
4540	In His Steps	230	0.0902	
38070	The Norwegian Fairy Book	149	0.1044	
16339	The Passenger from Calais	125	0.1454	
1535	The Tragedy of Coriolanus	116	0.1137	
35517	The Three Impostors; or, The Tra...	111	0.1148	
21374	!Tention: A Story of Boy-Life du...	92	0.2195	
24775	Up the River; or, Yachting on th...	88	0.1449	

Table B.22: Stories which are are represented by core emotional arc 11, sorted by the variance explained in their emotional arc by core emotional arc 11.

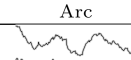
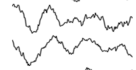
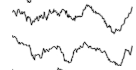
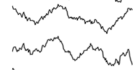
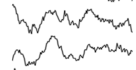
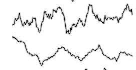
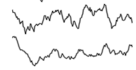
ID	Title	DL's	max(W[i,:])	Arc
5343	Rainbow Valley	257	0.1533	
1987	The Outlet	40	0.1444	
13882	John Thorndyke's Cases : related ...	217	0.1343	
1167	A Strange Disappearance	270	0.1282	
7477	The Book of Wonder	244	0.1276	
6753	Psmith in the City	199	0.1245	
5265	The Ball and the Cross	118	0.1229	
974	The Secret Agent: A Simple Tale	1,142	0.1224	
27525	Bones in London	85	0.1157	
24450	Bones: Being Further Adventures i...	121	0.1072	
5758	Many Cargoes	82	0.1048	
4090	From Ritual to Romance	90	0.1037	
27198	The Explorer	62	0.1035	
10832	Carnacki, the Ghost Finder	135	0.1032	
13372	The Gloved Hand	70	0.0942	

Table B.22: Stories which are are represented by core emotional arc 11, sorted by the variance explained in their emotional arc by core emotional arc 11.

ID	Title	DL's	max(W[i,:])	Arc
767	Agnes Grey	287	0.0856	
3055	The Wood Beyond the World	106	0.0799	
2641	A Room with a View	1,354	0.0681	

Table B.23: Top 10 stories which are are represented by core emotional arc 11, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
2641	A Room with a View	1,354	0.0681	
974	The Secret Agent: A Simple Tale	1,142	0.1224	
767	Agnes Grey	287	0.0856	
1167	A Strange Disappearance	270	0.1282	
5343	Rainbow Valley	257	0.1533	
7477	The Book of Wonder	244	0.1276	
13882	John Thorndyke's Cases : related ...	217	0.1343	
6753	Psmith in the City	199	0.1245	
10832	Carnacki, the Ghost Finder	135	0.1032	

Table B.24: Stories which are are represented by core emotional arc 12, sorted by the variance explained in their emotional arc by core emotional arc 12.

ID	Title	DL's	max(W[i,:])	Arc
318	John Barleycorn	130	0.1909	
1595	Whirligigs	94	0.1563	
25728	Desert Conquest; or, Precious Wa...	58	0.1543	
557	Puck of Pook's Hill	50	0.1482	
15976	Puck of Pook's Hill	40	0.1376	
26027	Puck of Pook's Hill	78	0.1357	
3326	The Well-Beloved: A Sketch of a ...	76	0.1226	
1611	Seventeen : A Tale of Youth and S...	56	0.1145	
12803	Headlong Hall	124	0.1113	
610	Idylls of the King	494	0.1075	
624	Looking Backward, 2000 to 1887	679	0.1060	
37532	The Scottish Fairy Book	98	0.1036	
10067	The Mystery of the Boule Cabinet...	54	0.0938	
316	The Golden Road	151	0.0864	
6840	Queen Lucia	81	0.0801	

Table B.25: Top 10 stories which are are represented by core emotional arc 12, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
624	Looking Backward, 2000 to 1887	679	0.1060	
610	Idylls of the King	494	0.1075	
316	The Golden Road	151	0.0864	
318	John Barleycorn	130	0.1909	
12803	Headlong Hall	124	0.1113	

Table B.25: Top 10 stories which are are represented by core emotional arc 12, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
37532	The Scottish Fairy Book	98	0.1036	
1595	Whirligigs	94	0.1563	
6840	Queen Lucia	81	0.0801	
26027	Puck of Pook's Hill	78	0.1357	

Table B.26: Stories which are are represented by core emotional arc 13, sorted by the variance explained in their emotional arc by core emotional arc 13.

ID	Title	DL's	max(W[i,:])	Arc
3815	Rolling Stones	107	0.1377	
26732	Free Air	42	0.1295	
30836	Seven Keys to Baldpate	57	0.1245	
22342	Supermind	61	0.1185	
9925	Black Jack	111	0.1174	
21687	The Youngest Girl in the Fifth: ...	59	0.1097	
19527	The Yukon Trail: A Tale of the N...	44	0.1085	
6312	Representative Men: Seven Lectures	170	0.0986	
33391	Bill Nye's Cordwood	248	0.0888	

Table B.27: Top 10 stories which are are represented by core emotional arc 13, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
33391	Bill Nye's Cordwood	248	0.0888	
6312	Representative Men: Seven Lectures	170	0.0986	
9925	Black Jack	111	0.1174	
3815	Rolling Stones	107	0.1377	
22342	Supermind	61	0.1185	
21687	The Youngest Girl in the Fifth: ...	59	0.1097	
30836	Seven Keys to Baldpate	57	0.1245	
19527	The Yukon Trail: A Tale of the N...	44	0.1085	
26732	Free Air	42	0.1295	

Table B.28: Stories which are are represented by core emotional arc 14, sorted by the variance explained in their emotional arc by core emotional arc 14.

ID	Title	DL's	max(W[i,:])	Arc
5077	Marmion: A Tale of Flodden Field...	125	0.1476	
20869	The Skylark of Space	246	0.1457	
15272	Spenser's The Faerie Queene, Book I	978	0.1302	
1027	The Lone Star Ranger: A Romance ...	253	0.1277	
2233	A Damsel in Distress	201	0.1256	
27690	Nobody's Girl: (En Famille)	70	0.1219	
32759	Red Nails	151	0.1158	
27063	The Hero	60	0.1155	
2804	Rose in Bloom : A Sequel to "Eigh...	168	0.1083	
45658	The Mystery of the Downs	48	0.1030	

Table B.28: Stories which are are represented by core emotional arc 14, sorted by the variance explained in their emotional arc by core emotional arc 14.

ID	Title	DL's	max(W[i,:])	Arc
105	Persuasion	2,535	0.0980	
619	The Warden	215	0.0946	
34732	Max Carrados	92	0.0941	
291	The Golden Age	42	0.0854	
15673	The Day of the Beast	98	0.0792	
1671	When a Man Marries	46	0.0651	

Table B.29: Top 10 stories which are are represented by core emotional arc 14, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
105	Persuasion	2,535	0.0980	
15272	Spenser's The Faerie Queene, Book I	978	0.1302	
1027	The Lone Star Ranger: A Romance ...	253	0.1277	
20869	The Skylark of Space	246	0.1457	
619	The Warden	215	0.0946	
2233	A Damsel in Distress	201	0.1256	
2804	Rose in Bloom : A Sequel to "Eigh...	168	0.1083	
32759	Red Nails	151	0.1158	
5077	Marmion: A Tale of Flodden Field...	125	0.1476	

Table B.30: Stories which are are represented by core emotional arc 15, sorted by the variance explained in their emotional arc by core emotional arc 15.

ID	Title	DL's	max(W[i,:])	Arc
33066	The Garden of Eden	72	0.1165	
40852	Instigations: Together with An Es...	79	0.0806	

Table B.31: Top 10 stories which are are represented by core emotional arc 15, sorted by downloads.

ID	Title	DL's	max(W[i,:])	Arc
40852	Instigations: Together with An Es...	79	0.0806	
33066	The Garden of Eden	72	0.1165	

Table B.32: Stories which are are represented by core emotional arc 16, sorted by the variance explained in their emotional arc by core emotional arc 16.

ID	Title	DL's	max(W[i,:])	Arc
872	Reprinted Pieces	49	0.1103	
8673	A Columbus of Space	56	0.1077	
434	The Circular Staircase	189	0.0978	
1263	The Glimpses of the Moon	58	0.0864	
3075	The Return	73	0.0759	

Table B.33: Top 10 stories which are are represented by core emotional arc 16, sorted by downloads.

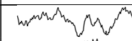
ID	Title	DL's	max(W[i,:])	Arc
434	The Circular Staircase	189	0.0978	
3075	The Return	73	0.0759	
1263	The Glimpses of the Moon	58	0.0864	
8673	A Columbus of Space	56	0.1077	
872	Reprinted Pieces	49	0.1103	

Table B.34: Stories which are are represented by core emotional arc 18, sorted by the variance explained in their emotional arc by core emotional arc 18.

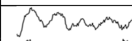
ID	Title	DL's	max(W[i,:])	Arc
5776	100%: the Story of a Patriot	64	0.1731	
3188	Mark Twain's Speeches	500	0.0778	

Table B.35: Top 10 stories which are are represented by core emotional arc 18, sorted by downloads.

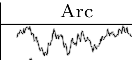
ID	Title	DL's	max(W[i,:])	Arc
3188	Mark Twain's Speeches	500	0.0778	
5776	100%: the Story of a Patriot	64	0.1731	

Table B.36: Stories which are are represented by core emotional arc 20, sorted by the variance explained in their emotional arc by core emotional arc 20.

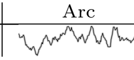
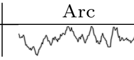
ID	Title	DL's	max(W[i,:])	Arc
15119	Handy Dictionary of Poetical Quo...	46	0.0728	

Table B.37: Top 10 stories which are are represented by core emotional arc 20, sorted by downloads.

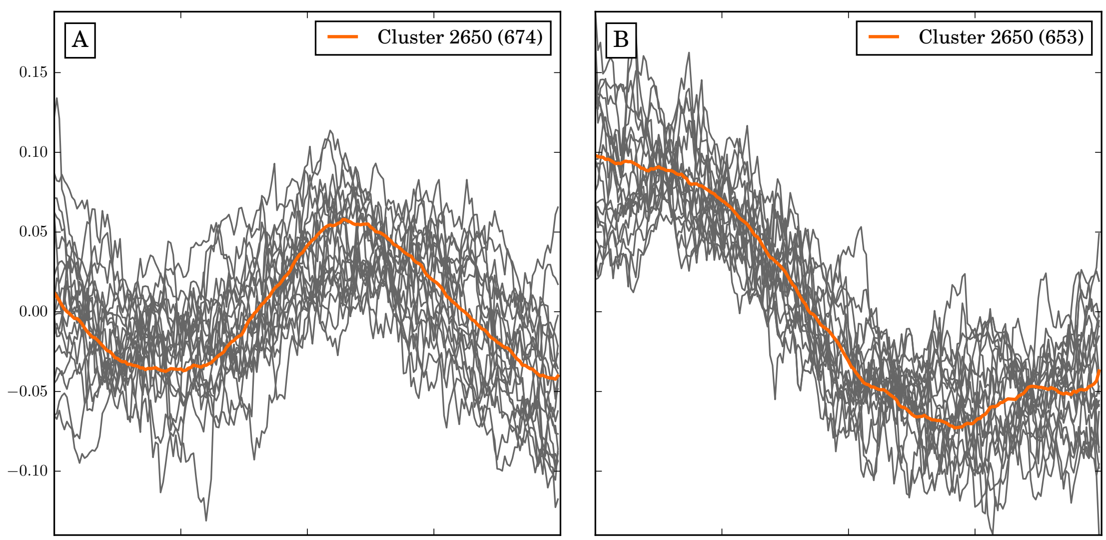
ID	Title	DL's	max(W[i,:])	Arc
15119	Handy Dictionary of Poetical Quo...	46	0.0728	

B.6 Additional Hierarchical Clustering Figures

In the section, we include additional results from the hierarchical clustering analysis. The distance function between clusters is defined in the `scipy` package using the incremental algorithm, starting with all arcs as separate clusters and iteratively merging them:

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

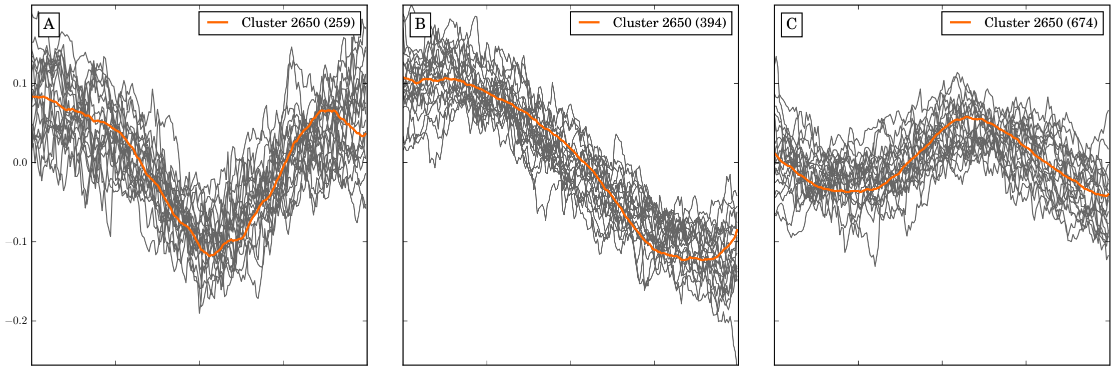
where $|v|$ denotes the cardinality of set v (single arcs have cardinality 1), u is the merged cluster of s,t, the denominator T is the sum of the sizes, and v is an unused cluster. Similar to the MATLAB implementation, this relies on a nearest-neighbor chain to be computed efficiently.



1. In a Glass Darkly, v. 2/3 (37173)
2. Daddy-Long-Legs (157)
3. The Contrast (554)
4. Arms and the Man (3618)
5. Black Amazon of Mars (32664)

1. The Taming of the Shrew (1107)
2. Those Extraordinary Twins (3185)
3. Heart of Darkness (219)
4. Our American Cousin (3158)
5. Brownies and Bogles (39782)

Figure B.9: The 2 clusters identified by Agglomerative Clustering using Ward's method.

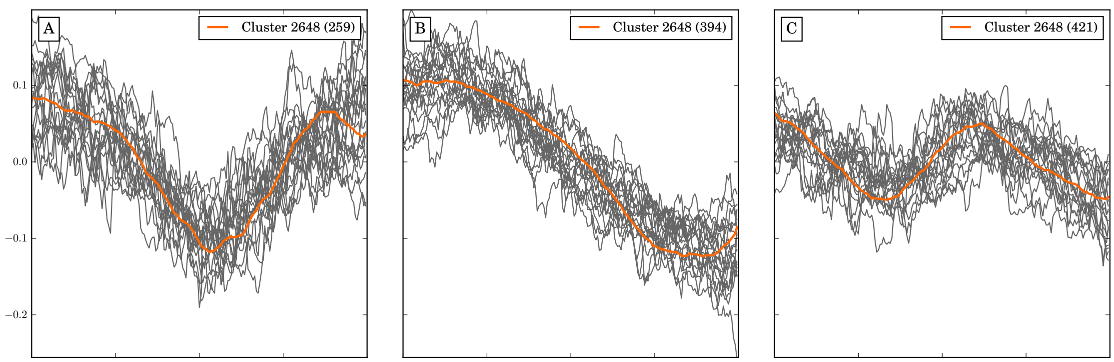


1. Anything You Can Do! (30742)
2. The Scarlet Plague (21970)
3. Measure for Measure (1126)
4. Eight Keys to Eden (27595)
5. Small Means and Great Ends (11435)

1. Lady Windermere's Fan (790)
2. Preface to Shakespeare (5429)
3. Brownies and Bogles (39782)
4. Anna Christie (4025)
5. Memoirs of the Author of a Vindication of the R... (16199)

1. In a Glass Darkly, v. 2/3 (37173)
2. Daddy-Long-Legs (157)
3. The Contrast (554)
4. Arms and the Man (3618)
5. Black Amazon of Mars (32664)

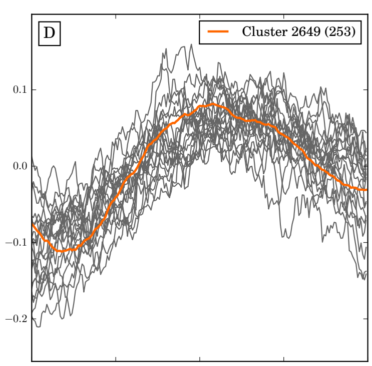
Figure B.10: The 3 clusters identified by Agglomerative Clustering using Ward's method.



1. Anything You Can Do! (30742)
2. The Scarlet Plague (21970)
3. Measure for Measure (1126)
4. Eight Keys to Eden (27595)
5. Small Means and Great Ends (11435)

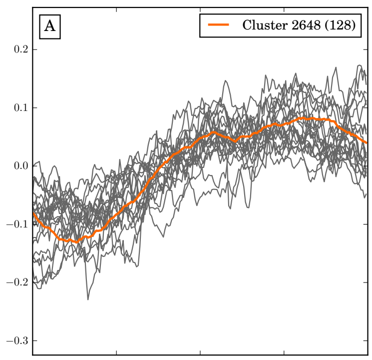
1. Lady Windermere's Fan (790)
2. Preface to Shakespeare (5429)
3. Brownies and Bogles (39782)
4. Anna Christie (4025)
5. Memoirs of the Author of a Vindication of the R... (16199)

1. The Silver Box: A Comedy in Three Acts (2906)
2. Tamburlaine the Great Part 2 (1589)
3. Twelve Types (12491)
4. A Parody Outline of History : Wherein May Be Fou... (1478)
5. She Stoops to Conquer; Or, The Mistakes of a Ni... (383)

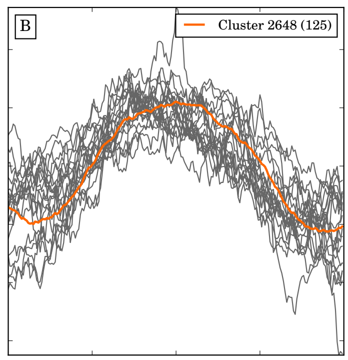


1. The Way of the World (1292)
2. The Royal Book of Oz : In which the Scarecrow go... (30537)
3. The First Part of King Henry the Fourth (1115)
4. The Sex Life of the Gods (40284)
5. Yesterdays (4006)

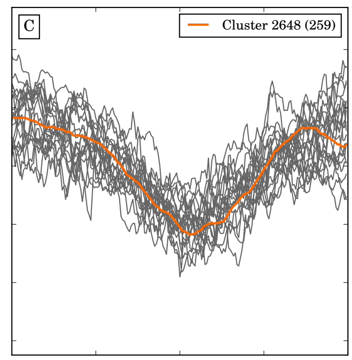
Figure B.11: The 4 clusters identified by Agglomerative Clustering using Ward's method.



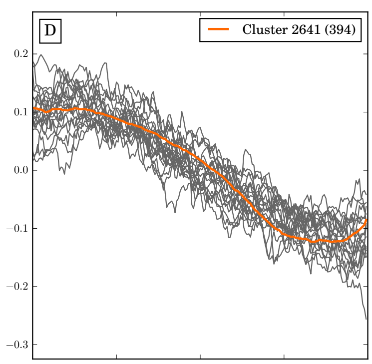
1. A Slave is a Slave (20726)
2. The Aesop for Children : With pictures by Milo W... (19994)
3. Book of Wise Sayings: Selected Largely from East... (21130)
4. No Hero (11153)
5. At the Back of the North Wind (18614)



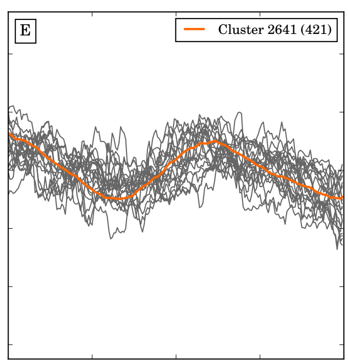
1. Caesar and Cleopatra (3329)
2. A Christmas Carol: The original manuscript (30368)
3. A Christmas Carol (19337)
4. The Land That Time Forgot (551)
5. The Rome Express (11451)



1. Anything You Can Do! (30742)
2. The Scarlet Plague (21970)
3. Measure for Measure (1126)
4. Eight Keys to Eden (27595)
5. Small Means and Great Ends (11435)



1. Lady Windermere's Fan (790)
2. Preface to Shakespeare (5429)
3. Brownies and Bogles (39782)
4. Anna Christie (4025)
5. Memoirs of the Author of a Vindication of the R... (16199)



1. The Silver Box: A Comedy in Three Acts (2906)
2. Tamburlaine the Great Part 2 (1589)
3. Twelve Types (12491)
4. A Parody Outline of History : Wherein May Be Fou... (1478)
5. She Stoops to Conquer; Or, The Mistakes of a Ni... (383)

Figure B.12: The 5 clusters identified by Agglomerative Clustering using Ward's method.

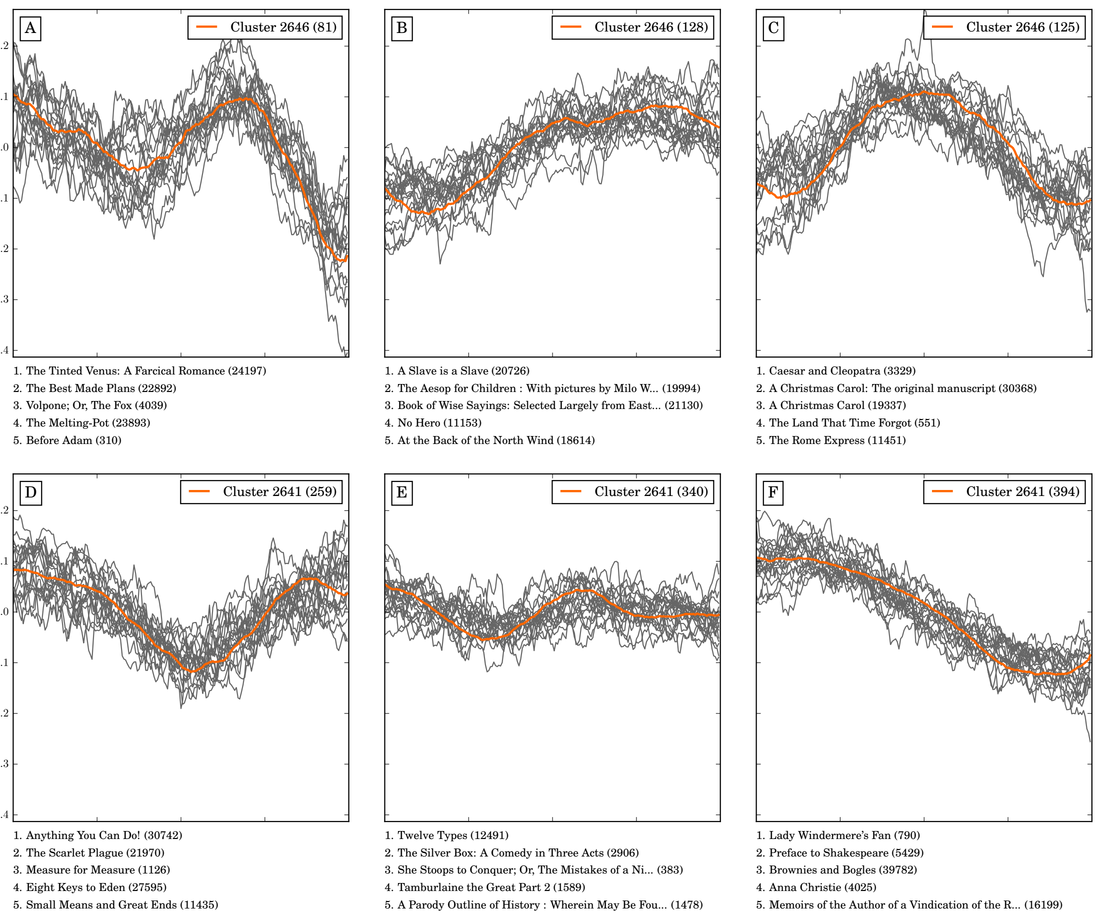
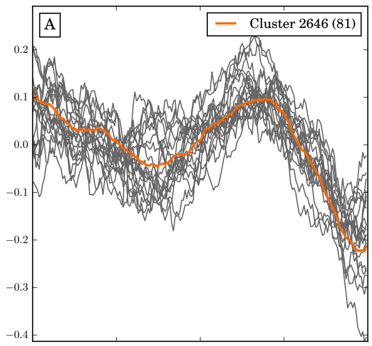
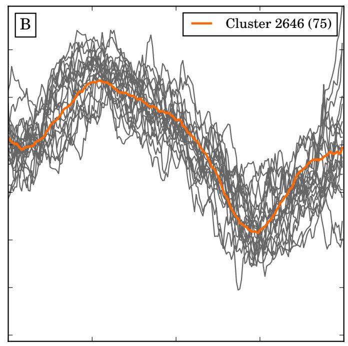


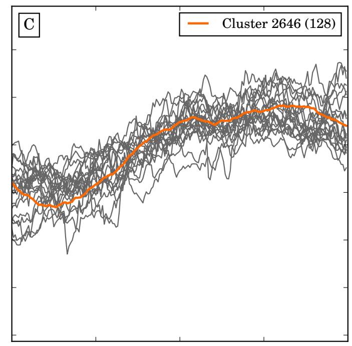
Figure B.13: The 6 clusters identified by Agglomerative Clustering using Ward's method.



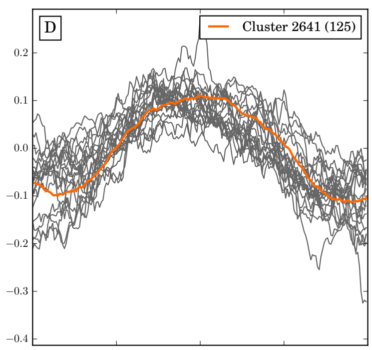
1. The Tinted Venus: A Farcical Romance (24197)
2. The Best Made Plans (22892)
3. Volpone; Or, The Fox (4039)
4. The Melting-Pot (23893)
5. Before Adam (310)



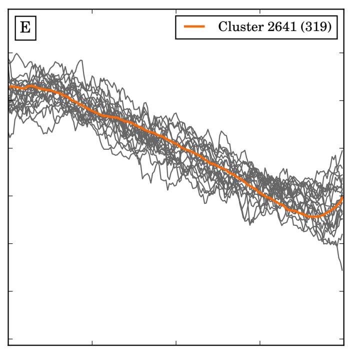
1. Tom Swift and His Motor-Cycle; Or, Fun and Adve... (4230)
2. Waifs and Strays: Part 1 (2295)
3. The Celtic Twilight (10459)
4. The Europeans (179)
5. The Exploits of Brigadier Gerard (11247)



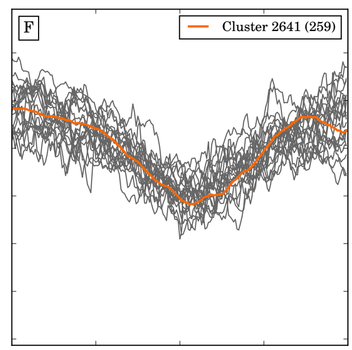
1. A Slave is a Slave (20726)
2. The Aesop for Children : With pictures by Milo W... (19994)
3. Book of Wise Sayings: Selected Largely from East... (21130)
4. No Hero (11153)
5. At the Back of the North Wind (18614)



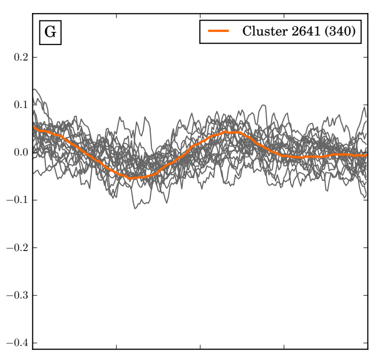
1. Caesar and Cleopatra (3329)
2. A Christmas Carol: The original manuscript (30368)
3. A Christmas Carol (19337)
4. The Land That Time Forgot (551)
5. The Rome Express (11451)



1. Lady Windermere's Fan (790)
2. Anna Christie (4025)
3. Preface to Shakespeare (5429)
4. Brownies and Bogles (39782)
5. Memoirs of the Author of a Vindication of the R... (16199)



1. Anything You Can Do! (30742)
2. The Scarlet Plague (21970)
3. Measure for Measure (1126)
4. Eight Keys to Eden (27595)
5. Small Means and Great Ends (11435)



1. Twelve Types (12491)
2. The Silver Box: A Comedy in Three Acts (2906)
3. She Stoops to Conquer; Or, The Mistakes of a Ni... (383)
4. Tamburlaine the Great Part 2 (1589)
5. A Parody Outline of History : Wherein May Be Fou... (1478)

Figure B.14: The 7 clusters identified by Agglomerative Clustering using Ward's method.

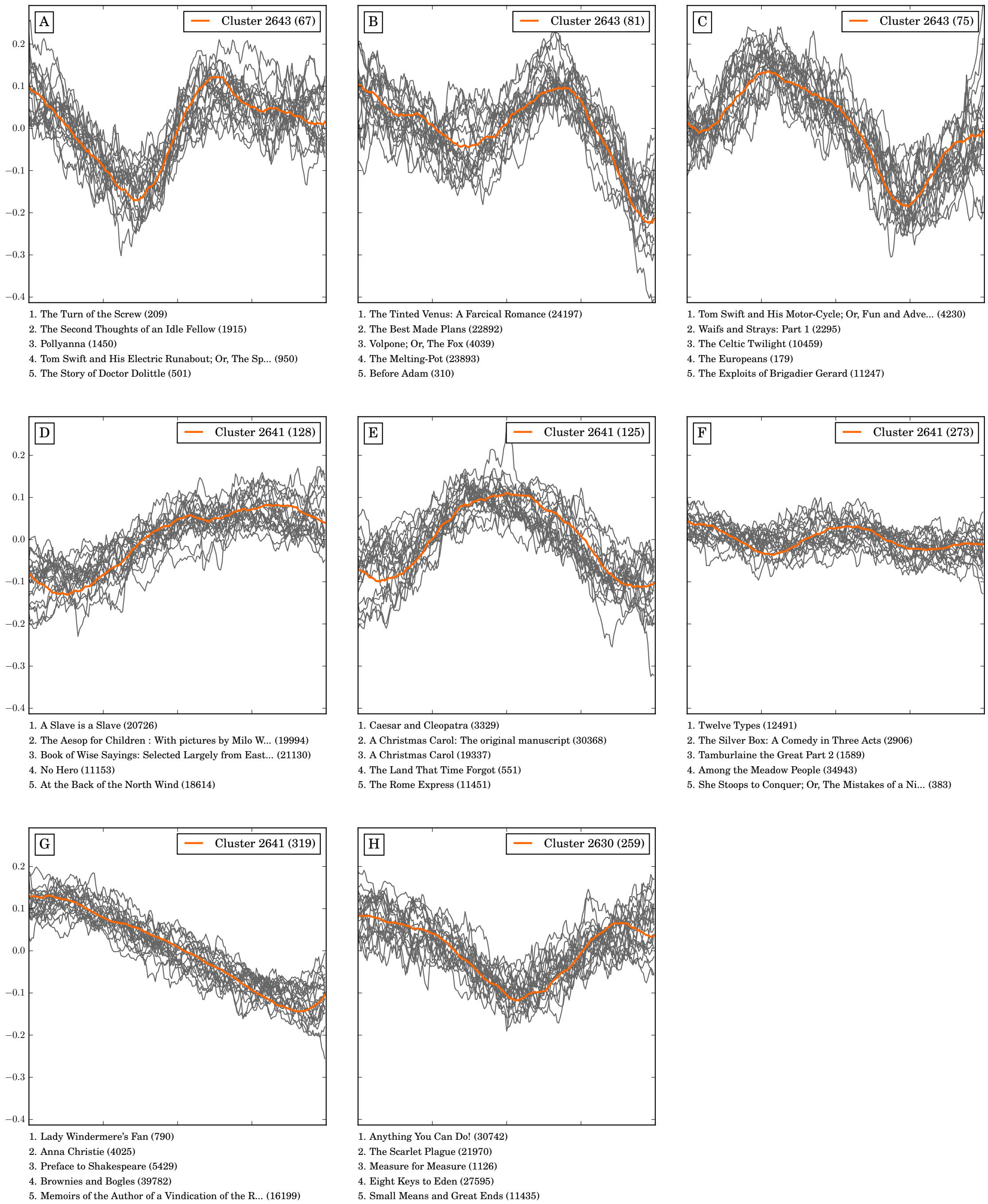


Figure B.15: The 8 clusters identified by Agglomerative Clustering using Ward's method.

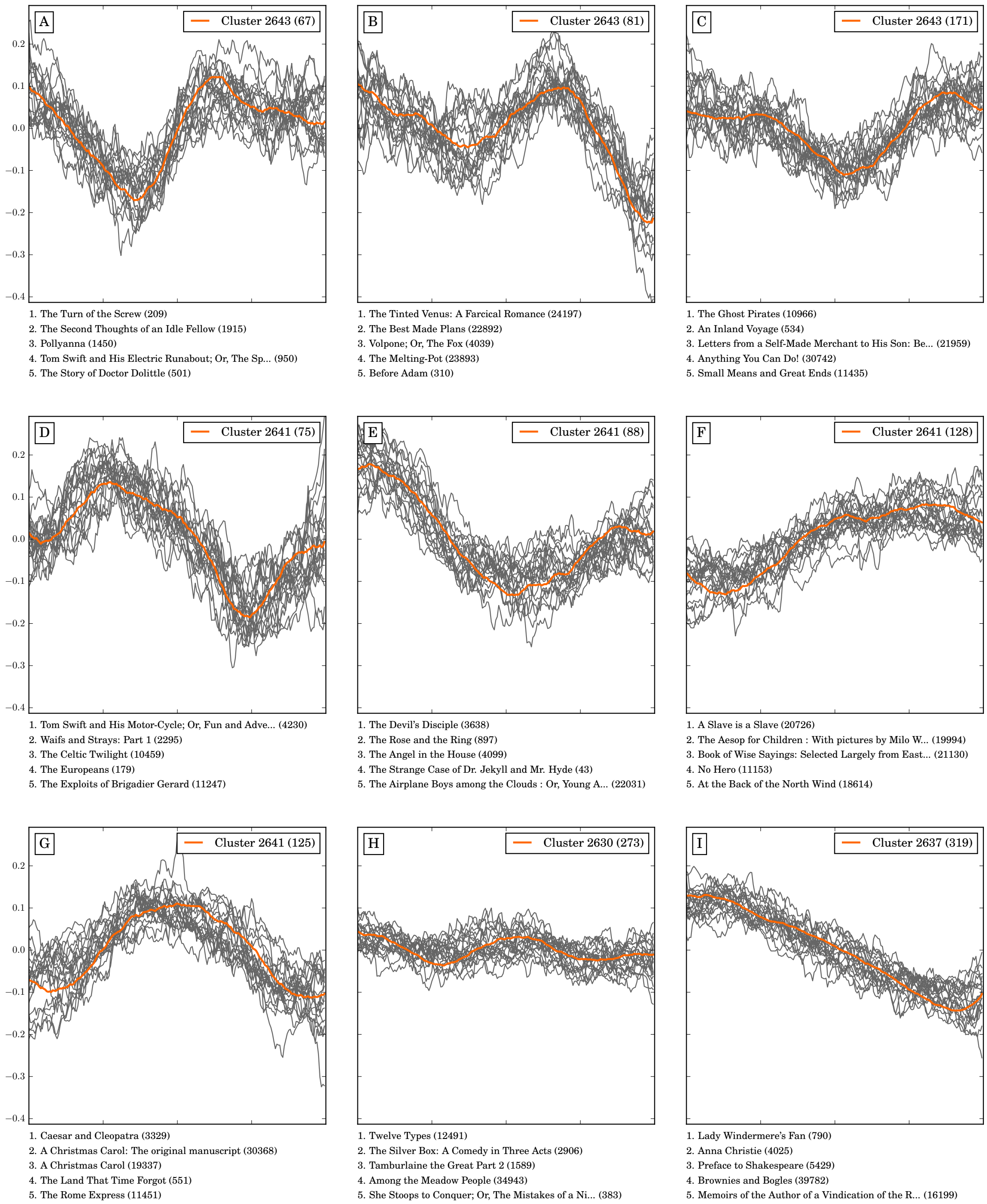


Figure B.16: The 9 clusters identified by Agglomerative Clustering using Ward's method.

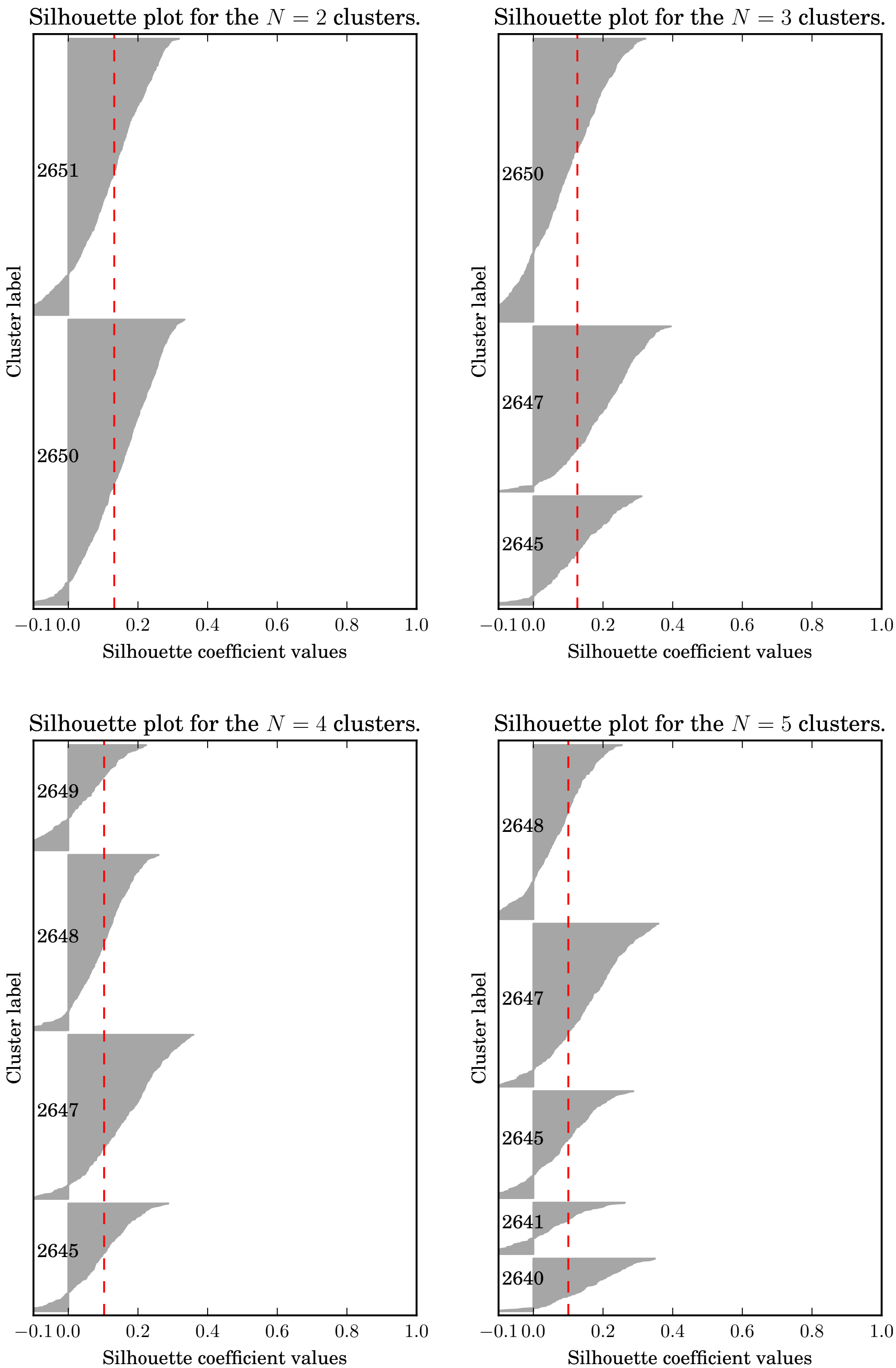


Figure B.17: The silhouette plots for 2–5 clusters identified by Agglomerative Clustering using Ward's method.

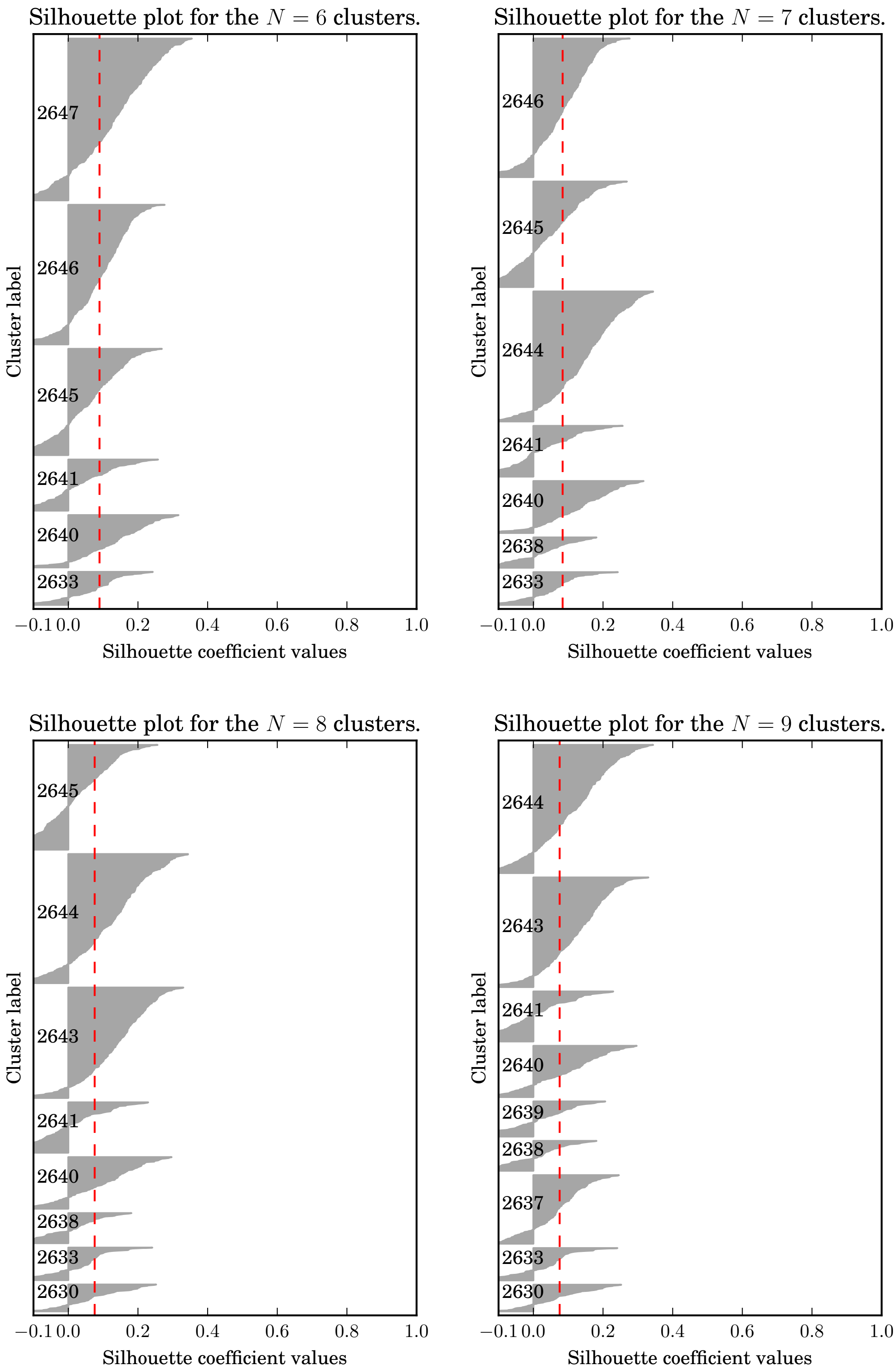


Figure B.18: The silhouette plots for 6–9 clusters identified by Agglomerative Clustering using Ward's method.

B.7 Additional SOM Figures

In Fig. B.19 we show the emotional arcs that are closest to each of 9 most frequently winning nodes in the winner-take-all implementation the Self Organizing Map.

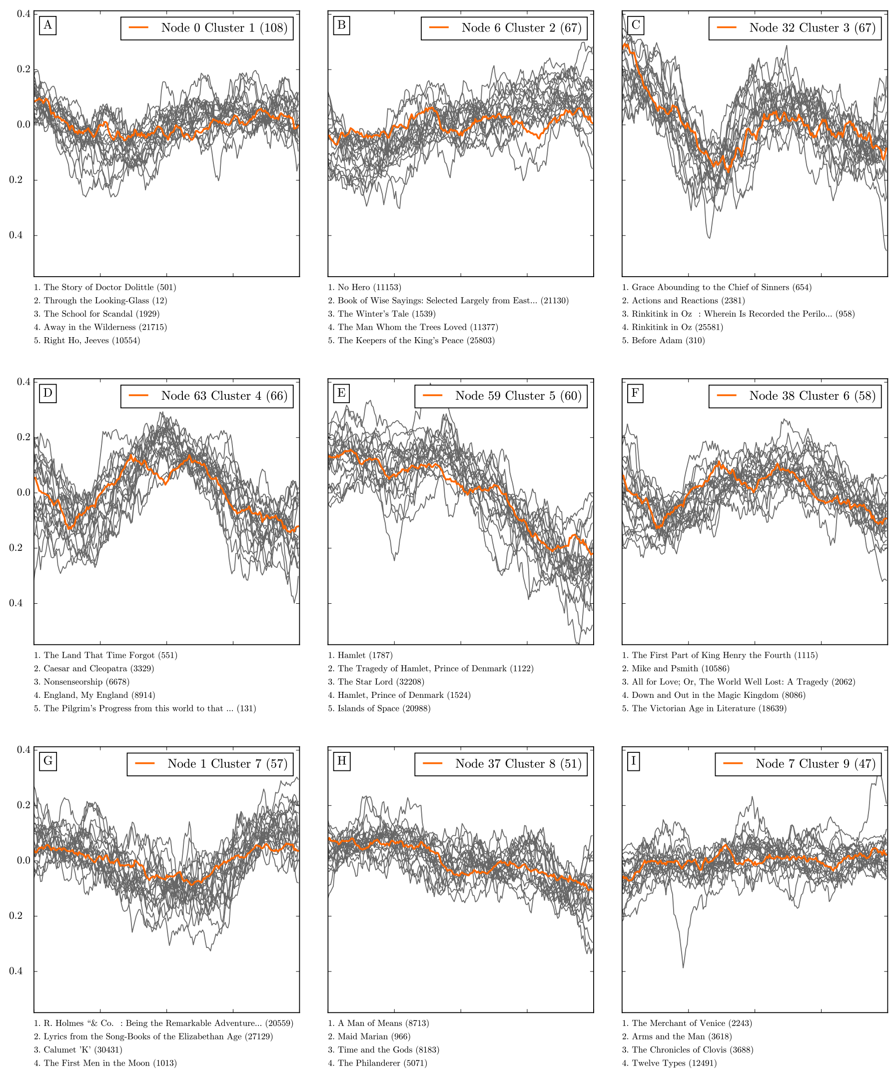


Figure B.19: The vector for each of the top 9 SOM nodes, accompanied with those sentiment time series which are closest to that node. The core stories which we have found with other methods are readily visible.

B.8 Null comparison details

An example of the "nonsense" and "word salad" text is presented first in Appendix B.3. First, we examine the resulting timeseries for an example book in Figs. B.20 and B.21. We then go on to present the full result of the SVD, agglomerative clustering, and SOM to "nonsense" English fiction books with more than 40 downloads.

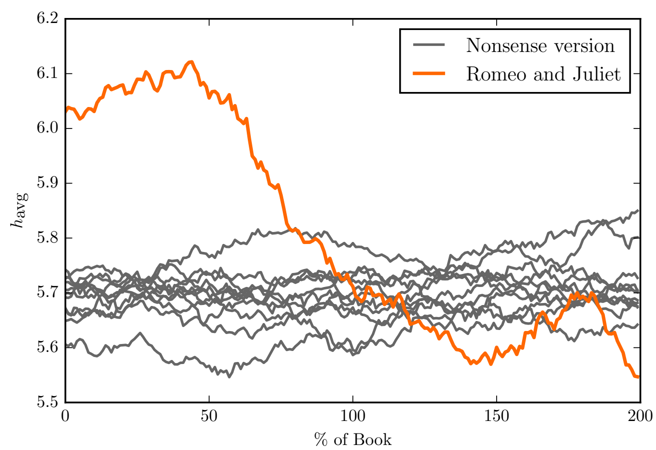


Figure B.20: The emotional arc of Romeo And Juliet by William Shakespeare (Gutenberg ID 1777), along with 11 "nonsense" versions, as produced by a 2-gram Markov model. We see that the emotional arc from the true version has more structure than the nonsense versions.

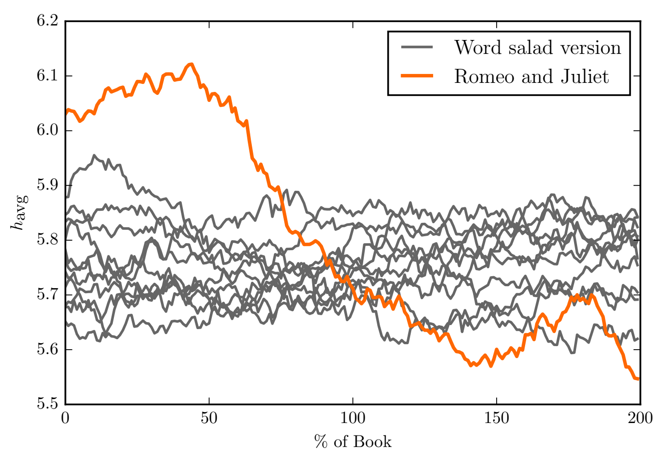


Figure B.21: The emotional arc of Romeo And Juliet by William Shakespeare, along with 11 "word salad" versions, as produced by randomly shuffling the words in the book. We see that the emotional arc from the true version has more structure than the word salad versions as well.

B.8.1 Null SVD

SVD modes from the emotional arcs of word salad books. We observe higher frequency modes appearing more quickly, and a more even spread of mode coefficients.

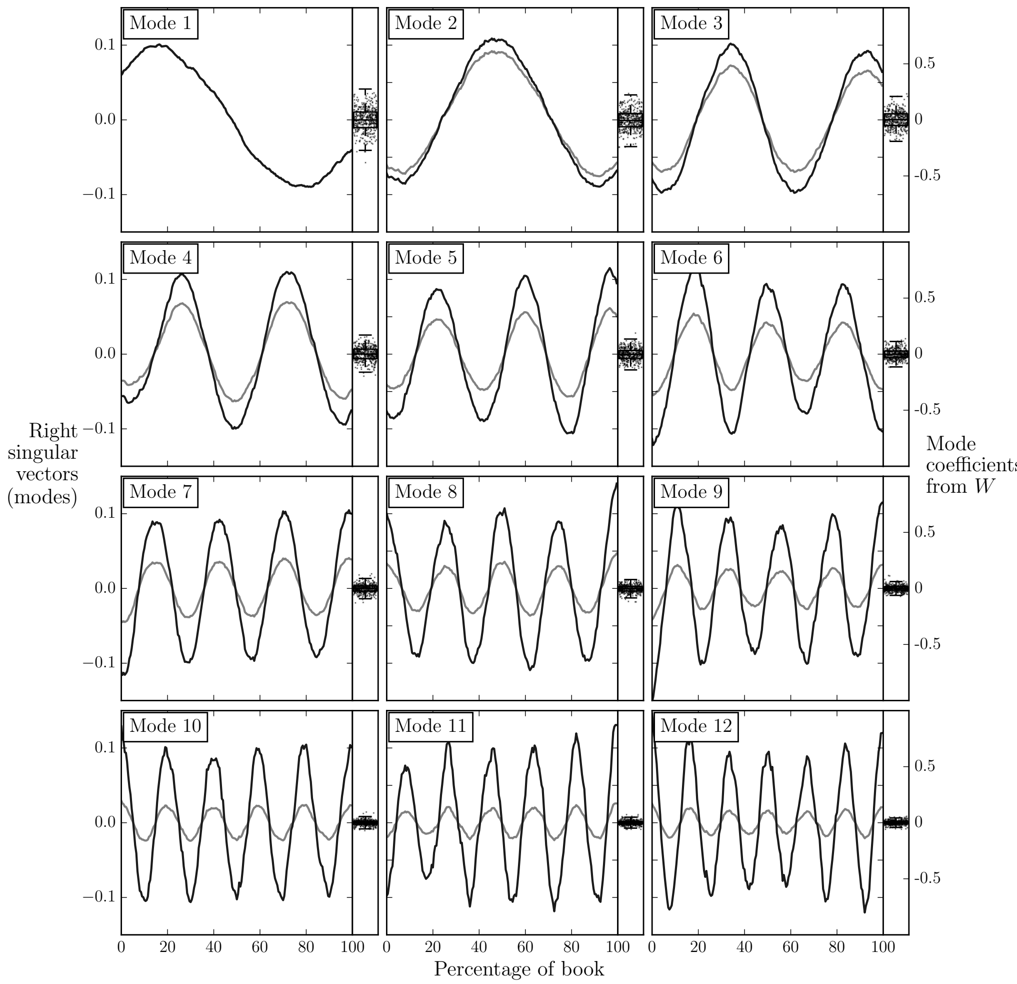


Figure B.22: Top 12 modes from the Singular Value Decomposition of 1,327 nonsense Project Gutenberg books. We show in a lighter color modes weighted by their corresponding singular value, where we have scaled the matrix Σ such that the first entry is 1 for comparison. The mode coefficients normalized for each book are shown in the right panel accompanying each mode, in the range -1 to 1, with the "Tukey" box plot.

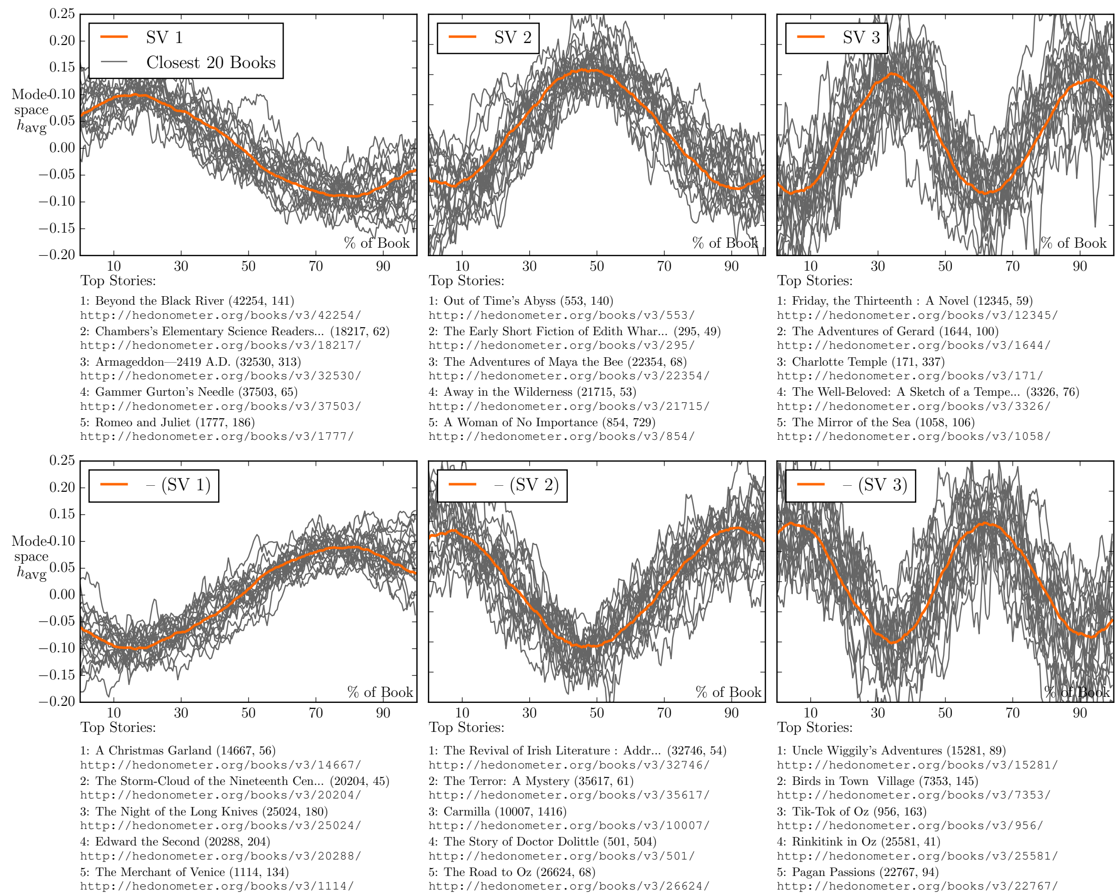
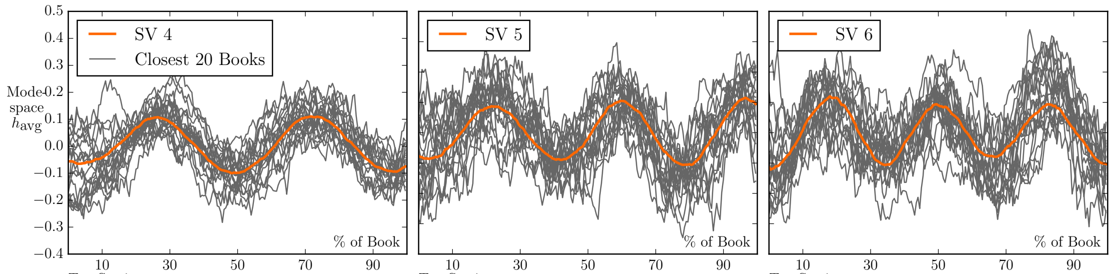


Figure B.23: First 3 SVD modes from nonsense books and their negation with the closest stories to each. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.



Top Stories:

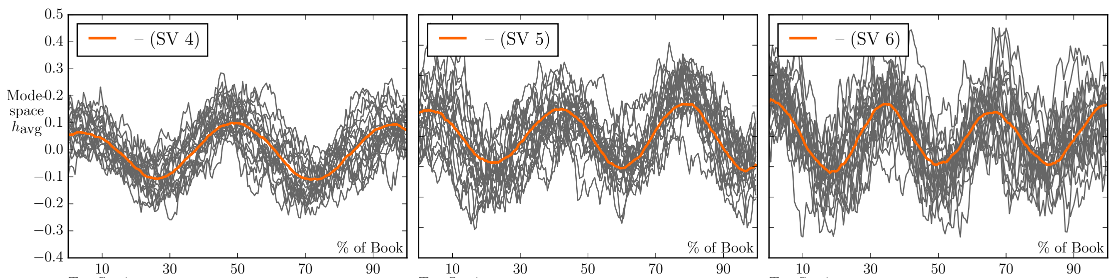
1: That Affair at Elizabeth (35247, 62)
http://hedonometer.org/books/v3/35247/
2: Empire (28215, 214)
http://hedonometer.org/books/v3/28215/
3: Flappers and Philosophers (4368, 287)
http://hedonometer.org/books/v3/4368/
4: The Gentle Grafter (1805, 84)
http://hedonometer.org/books/v3/1805/
5: Just Patty (21048, 47)
http://hedonometer.org/books/v3/21048/

Top Stories:

1: The Search (25866, 56)
http://hedonometer.org/books/v3/25866/
2: The Norwegian Fairy Book (38070, 149)
http://hedonometer.org/books/v3/38070/
3: The Insidious Dr. Fu Manchu (173, 245)
http://hedonometer.org/books/v3/173/
4: A Damsel in Distress (2233, 201)
http://hedonometer.org/books/v3/2233/
5: The Angel in the House (4099, 175)
http://hedonometer.org/books/v3/4099/

Top Stories:

1: White Fang (910, 1178)
http://hedonometer.org/books/v3/910/
2: Puck of Pook's Hill (26027, 78)
http://hedonometer.org/books/v3/26027/
3: Do and Dare — a Brave Boy's Fight for... (5747, 61)
http://hedonometer.org/books/v3/5747/
4: Familiar Quotations (16732, 86)
http://hedonometer.org/books/v3/16732/
5: A Daughter of the Sioux: A Tale of th... (19023, 61)
http://hedonometer.org/books/v3/19023/



Top Stories:

1: My Lady Nicotine: A Study in Smoke (18934, 60)
http://hedonometer.org/books/v3/18934/
2: The Young Pitcher (19246, 49)
http://hedonometer.org/books/v3/19246/
3: Betty Gordon at Boarding School; Or, ... (10317, 42)
http://hedonometer.org/books/v3/10317/
4: The Haunted Bookshop (172, 132)
http://hedonometer.org/books/v3/172/
5: The Mysterious Key and What It Opened (8188, 83)
http://hedonometer.org/books/v3/8188/

Top Stories:

1: Humorous Masterpieces from American L... (15585, 53)
http://hedonometer.org/books/v3/15585/
2: The Second Jungle Book (1937, 304)
http://hedonometer.org/books/v3/1937/
3: The Untamed (10886, 109)
http://hedonometer.org/books/v3/10886/
4: Henry IV, Part 1 (2251, 76)
http://hedonometer.org/books/v3/2251/
5: The Black Arrow: A Tale of the Two Roses (32954, 115)
http://hedonometer.org/books/v3/32954/

Top Stories:

1: The Longest Journey (2604, 88)
http://hedonometer.org/books/v3/2604/
2: The Little Nugget (6683, 95)
http://hedonometer.org/books/v3/6683/
3: Those Extraordinary Twins (3185, 42)
http://hedonometer.org/books/v3/3185/
4: The Enchanted Castle (3536, 220)
http://hedonometer.org/books/v3/3536/
5: Humorous Readings and Recitations, in... (36775, 136)
http://hedonometer.org/books/v3/36775/

Figure B.24: Modes 4–6 from the SVD analysis of nonsense books and their negation with the closest stories to each. Links below each story point to an interactive visualization on http://hedonometer.org which enables detailed exploration of the emotional arc for the story.

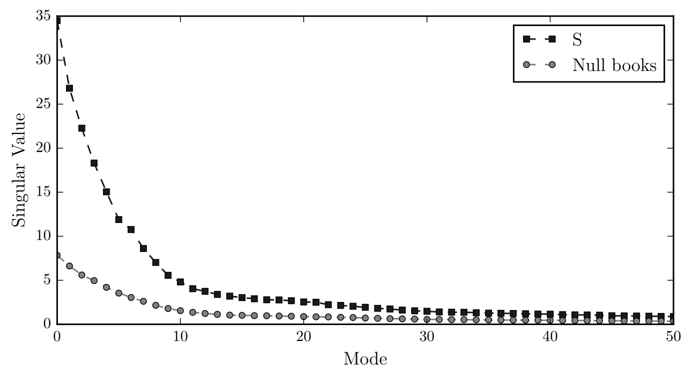


Figure B.25: Comparison of the singular value spectra from the emotional arcs of nonsense books and the emotional arcs of individual Project Gutenberg books. The spectra from the nonsense books is muted, indicating both lower total variance explained and less important ordering of the singular vectors.

B.8.2 Null Hierarchical Clustering

Dendrogram of clustering using Ward's method on the emotional arcs of word salad books. We observe comparatively low linkage cost for these emotional arcs, indicating the absence of distinct clusters.

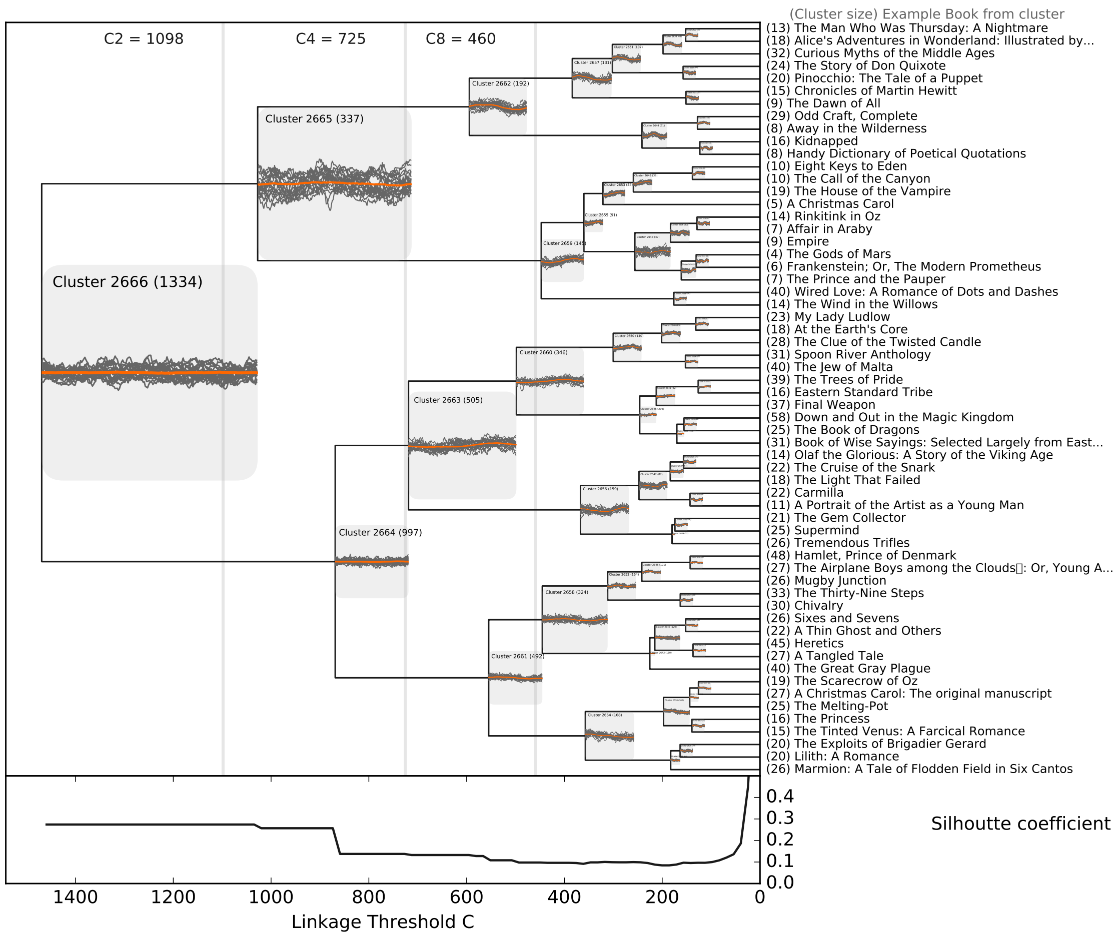


Figure B.26: Dendrogram from the agglomerative clustering procedure using Ward's minimum variance method on nonsense books. For each cluster, a selection of the 20 most central books to a fully-connected network of books are shown along with the average of the emotional arc for all books in the cluster, along with the cluster ID and number of books in each cluster (shown in parenthesis). At the bottom, we show the average Silhouette value for all books, with higher value representing a more appropriate number of clusters. For each of the 60 leaf nodes (right side) we show the number of books within the cluster and the most central book to that cluster's book network.

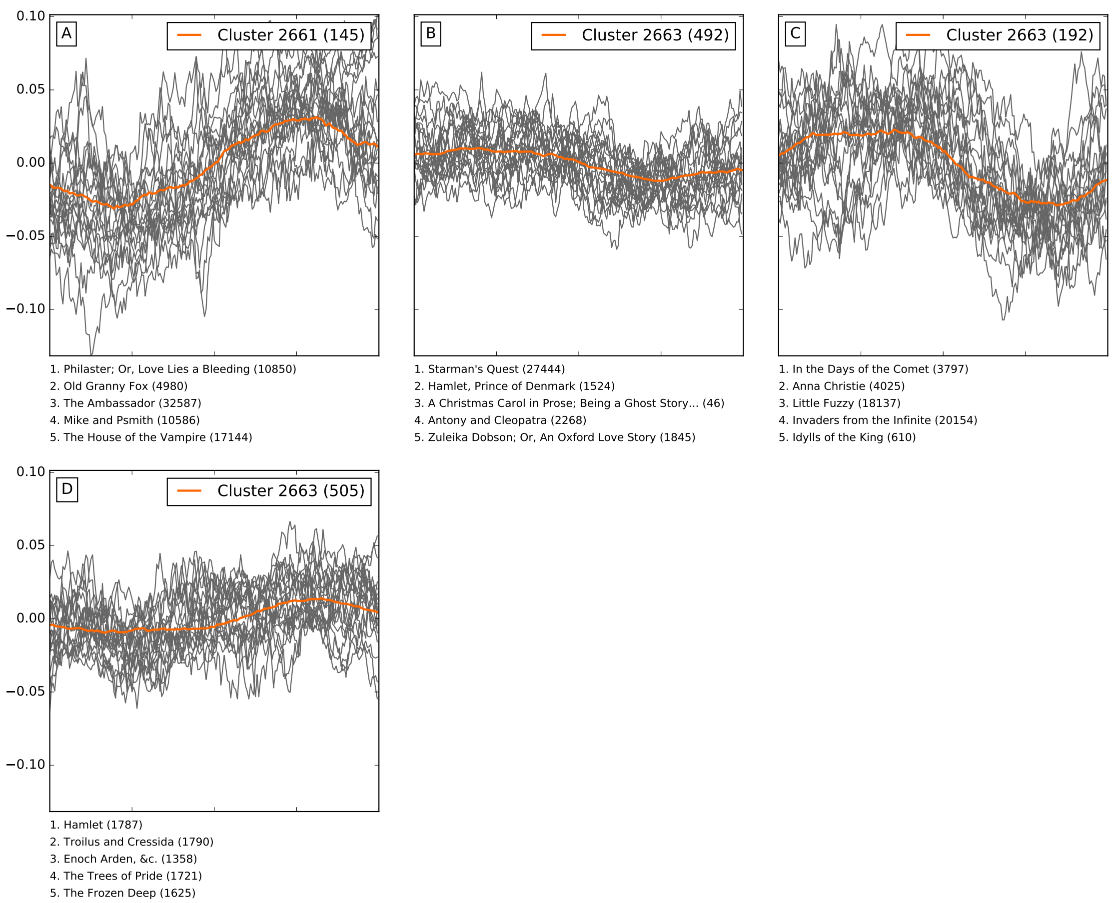


Figure B.27: Four clusters (linkage threshold 850) from the hierarchical clustering of word salad books. We observe that the cluster mean emotional arc and the most central emotional arcs have high variance, without a visible signal.

B.8.3 Null Self Organizing Map (SOM)

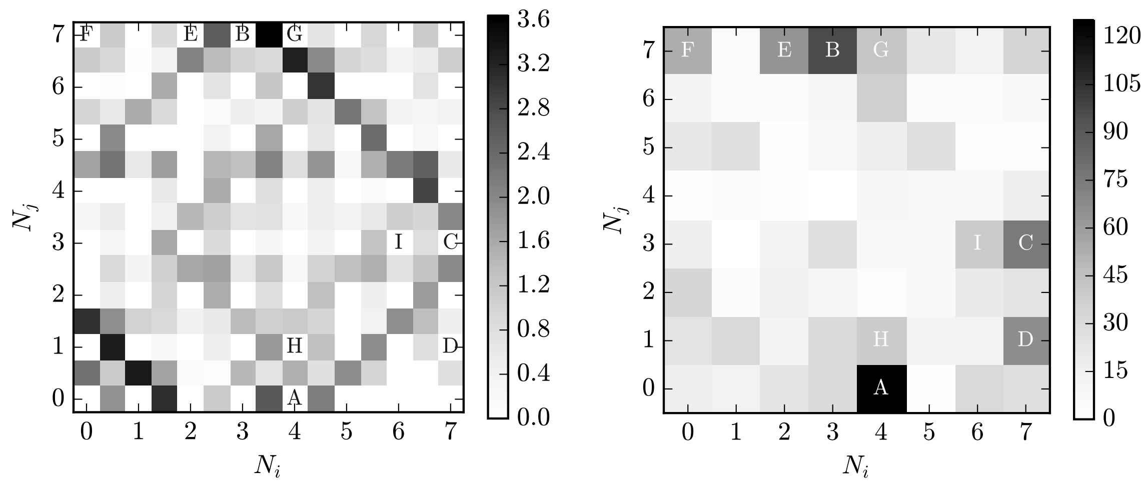


Figure B.28: Results of the SOM applied to nonsense versions of Project Gutenberg books. Left panel: Nodes on the 2D SOM grid are shaded by the number of stories for which they are the winner. Right panel: The B-Matrix shows that there are clear clusters of stories in the 2D space imposed by the SOM network.

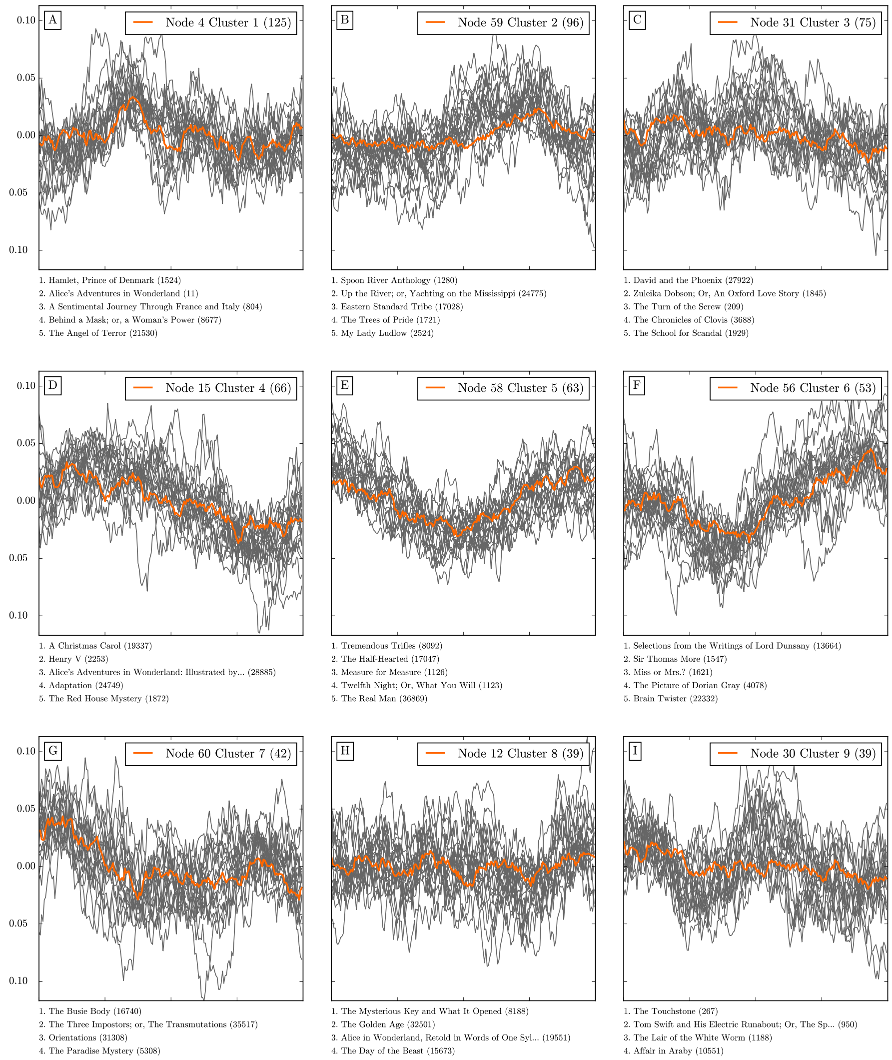


Figure B.29: The vector for each of the top 9 SOM nodes for null emotional arcs, accompanied with those sentiment time series which are closest to that node. Panels D and E show what appear to be similar arcs to the six we identified in real books, but overall see that the emotional arcs from null arcs show little coherent structure, especially considering the y-range here being 0.1 compared to the 0.4 of the real books (had we used the same y-range, very little of the variation would be visible at all).

Appendix C: labMTsimple: A Python Library for Sentiment Analysis

C.1 Getting Started

In this chapter, we provide details for a Python package called `labMTsimple`. The package exposes a simple, but quickly growing, labMT usage library.

C.1.1 Usage

This package uses the language assessment by Mechanical Turk (labMT) word list to score the happiness of a corpus. The labMT word list was created by combining the 5000 words most frequently appearing in four sources: Twitter, the New York Times, Google Books, and music lyrics, and then scoring the words for sentiment on Amazon's Mechanical Turk. The list is described in detail in the publication Dodds' et al. 2011, PLOS ONE, "Temporal Patterns of Happiness and Information in a Global-Scale Social Network: Hedonometrics and Twitter."

Given two corpora, the script "storylab.py" creates a word-shift graph illustrating the words most responsible for the difference in happiness between the two corpora. The corpora should be large (e.g. at least 10,000 words) in order for the difference to be meaningful, as this is a bag-of-words approach. As an example, a random collection of English tweets from both Saturday January 18

2014 and Tuesday January 21 2014 are included in the “example” directory. They can be compared by moving to the test directory, using the command

```
python example.py example-shift.html
```

and opening the file `example-shift.html` in a web browser. For an explanation of the resulting plot, please visit

http://www.hedonometer.org/shifts.html

C.1.2 Installation

Cloning the github directly is recommended, and then installing locally:

```
git clone https://github.com/andyreagan/labMT-simple.git
cd labMT-simple
python setup.py install
```

This repository can also be installed using pip

```
pip install labMTsimple
```

in which case you can download the tests from github and run them, if desired.

C.1.3 Running tests

Tests are based on nose2, `pipinstallnose2`, and can be run inside the by executing

```
nose2
```

in the root directory of this repository.

This will compare the two days in test/data and print test.html which shifts them, allowing for a changable lens.

C.1.4 Developing with labMT-simple locally

It is often useful to reload the library when testing it interactively:

```
try:
    reload
except NameError:
    \# Python 3
    from importlib import reload
```

C.1.5 Building these docs

Go into the docs directory (activate local `virtualenv` first), and do the following:

```
\rm -rf _build/*
make html
make latexpdf
git add -f *
git commit -am ``new docs, probably should just add a pre-commit hook''
```

Note that these docs will build locally in python 2 because the dependencies exist. With python 3 available, these dependencies will be mocked (and this is set for the online readthedocs site).

(`sphinx-apidoc-o.../labMTsimple` was run once.)

C.2 Detailed Examples

C.2.1 Preparing texts

This is simple really: just load the text to be scored into python. This is using a subset of a couple days of public tweets to text, and they have already put the tweet text into `.txt` files that are loaded into strings:

```
f = codecs.open(``data/18.01.14.txt'',''r'',''utf8'')
saturday = f.read()
f.close()

f = codecs.open(``data/21.01.14.txt'',''r'',''utf8'')
tuesday = f.read()
f.close()
```

C.2.2 Loading dictionaries

Again this is really simple, just use the `emotionFileReader` function:

```
lang = `english'
labMT,labMTvector,labMTwordList = emotionFileReader(stopval=0.0,lang=lang,returnVector=True)
```

Then we can score the text and get the word vector at the same time:

```
saturdayValence,saturdayFvec = emotion(saturday,labMT,shift=True,happsList=labMTvector)
tuesdayValence,tuesdayFvec = emotion(tuesday,labMT,shift=True,happsList=labMTvector)
```

But we don't want to use these happiness scores yet, because they included all words (including neutral words). So, set all of the neutral words to 0, and generate the scores:

```
tuesdayStoppedVec = stopper(tuesdayFvec,labMTvector,labMTwordList,stopVal=1.0)
saturdayStoppedVec = stopper(saturdayFvec,labMTvector,labMTwordList,stopVal=1.0)

saturdayValence = emotionV(saturdayStoppedVec,labMTvector)
tuesdayValence = emotionV(tuesdayStoppedVec,labMTvector)
```

C.3 Making Wordshifts

With merged updates to the d3 wordshift plotting in labMTsimple, and combined with phantom crowbar (see previous post), it's easier than ever to use the labMT data set to compare texts.

To make an html page with the shift, you'll just need to have labMT-simple installed. To automate the process into generating svg files, you'll need the phantom crowbar, which depends on phantomjs. To go all the way to pdf, you'll also need inkscape for making vectorized pdfs, or rsvg for making better formatted, but rasterized, versions.

Let's get set up to make shifts automatically. Since they're aren't many dependencies all the way down, start by getting phantomjs installed, then the phantom-crowbar.

C.3.1 Installing phantom-crowbar

For the phantomjs, use homebrew:

```
brew update
brew upgrade
brew install phantomjs
```

Then to get the crowbar, clone the git repository.

```
cd \textasciitilde{}
git clone https://github.com/andyreagan/phantom-crowbar
```

To use it system-wide, use the bash alias:

```
alias phantom-crowbar=''/usr/local/bin/phantomjs \textasciitilde{}/phantom-crowbar/phantom-crowbar.js''
```

Without too much detail, add this to your `~/.bash_profile` so that it's loaded every time you start a terminal session.

C.3.2 Installing inkscape

You only need inkscape if you want to go from svg to pdf (and there are other ways too), but this one is easy with, again, homebrew.

```
brew install inkscape
```

C.3.3 Installing rsvg

You only need inkscape if you want to go from svg to pdf (and there are other ways too), but this one is easy with, again, homebrew.

```
brew install librsvg
```

C.3.4 Installing labMTsimple

There are two ways to get it: using pip of cloning the git repo. If you're not sure, use pip. Pip makes it easier to keep it up to date, etc.

```
pip install labMTsimple
```

C.3.5 Making your first shift

If you cloned the git repository, install the thing and then you can check out the example in `examples/example.py`. If you went with pip, see that file on github.

Go ahead and run that script!

```
python example-002.py
```

You can open the html file to see the shift in any browser, with your choice of local webserver. Python's SimpleHTTPServer works fine, and generally the node based http-server is a bit more stable.

To take out the svg, go ahead and use the `phantom-crowbar.js` file copied to the `example/static` directory. Running it looks like this, for me:

```
/usr/local/bin/phantomjs js/shift-crowbar.js example-002.html shiftsvg wordshift.svg
```

Using inkscape or librsvg on my computer look like this:

```
  /Applications/Inkscape.app/Contents/Resources/bin/inkscape \
  -f \$(pwd)/wordshift.svg \
  -A \$(pwd)/wordshift-inkscape.pdf

rsvg-convert --format=eps worshift.svg \textgreater{} wordshift-rsvg.eps
epstopdf wordshift-rsvg.eps
```

And again, feel free to tweet suggestions at @andyreagan, and submit pull requests to the source code!

C.3.6 Full Automation

This procedure wraps up what is potentially the most backwards way to generate figure imaginable. SThe `shiftPDF()` function operates the same way as the `shiftHTML()`, but uses the headless web server to render the d3 graphic, then exectues a piece of injected JS to save a local SVG, and uses command line image manipulation libraries to massage it into a PDF.

On my macbook, this works, but your mileage will most certainly vary.

C.4 Advanced Usage

C.4.1 About Tries

For dictionary lookup of scores from phrases, the fastest benchmarks available and that were reasonable stable were from the libraries `datrie` and `marisatrie` which both have python bindings.

They're used in the `speedy` module in an attempt to both speed things up, and match against word stems.

C.4.2 Advanced Parsing

Some dictionaries use word stems to cover the multiple uses of a single word, with a single score. We can very quickly match these word stems using a prefix match on a trie. This is much better than using many compiled RE matches, which in my testing took a very long time.

Appendix D: Code for VACC Twitter Database Keyword Searches

In this Appendix we describe a strategy for utilizing the computational resources available at the University of Vermont's supercomputing center for searching through Twitter data. A schematic of the general approach is provided below in Figure D.1. We provide scripts for a minimum working example of this approach online at https://github.com/andyreagan/VACC-keyword-search.

The basic approach is to use the cron scheduler to make sure that around 100–150 jobs are running all the time. Each job is short, processing at most an hour of Tweets, so that each job takes less time to run and can utilize the shortq, which has a limit of 200 jobs that run immediately under most circumstances.

Cron calls the shell script `cron.sh` directly, and that shell script invokes the Python script `qsub.py` to handle the more complex logic of dates and job submission (which is just easier in Python). The work of looking through Tweets happens in `processTweets.py`. By utilizing pipes and unzipping from disk directly into Python, no unzipped files are written to disk (this would be prohibitively slow and use too much storage). Instead, just the minimum necessary output from the search is written to disk, by Python.

To make this run for new keywords, do the following:

1. Edit the keywords in `processTweets.py`.

2. Set the start date in `currdate.txt`.

3. Create folders for the keywords (manually, or using `makeFolders` function from the `processTweets.py` script). Including the base `raw-tweets` folder.

4. Instantiate a virtual environment. It is called set it up in the job script submitted by `qsub.py`, and in the `cron.sh` to call `qsub.py`. (Or don't use one, at your own peril!).

5. Edit the folders in `cron.sh` and `qsub.py` to where this is running.

6. Add the script `cron.sh` to your crontab.

7. Profit.

Using the commands `mmlsquota` and `showq`, you can see your file system usage and track the individual jobs (look at `currdate.txt` for submitted progress).

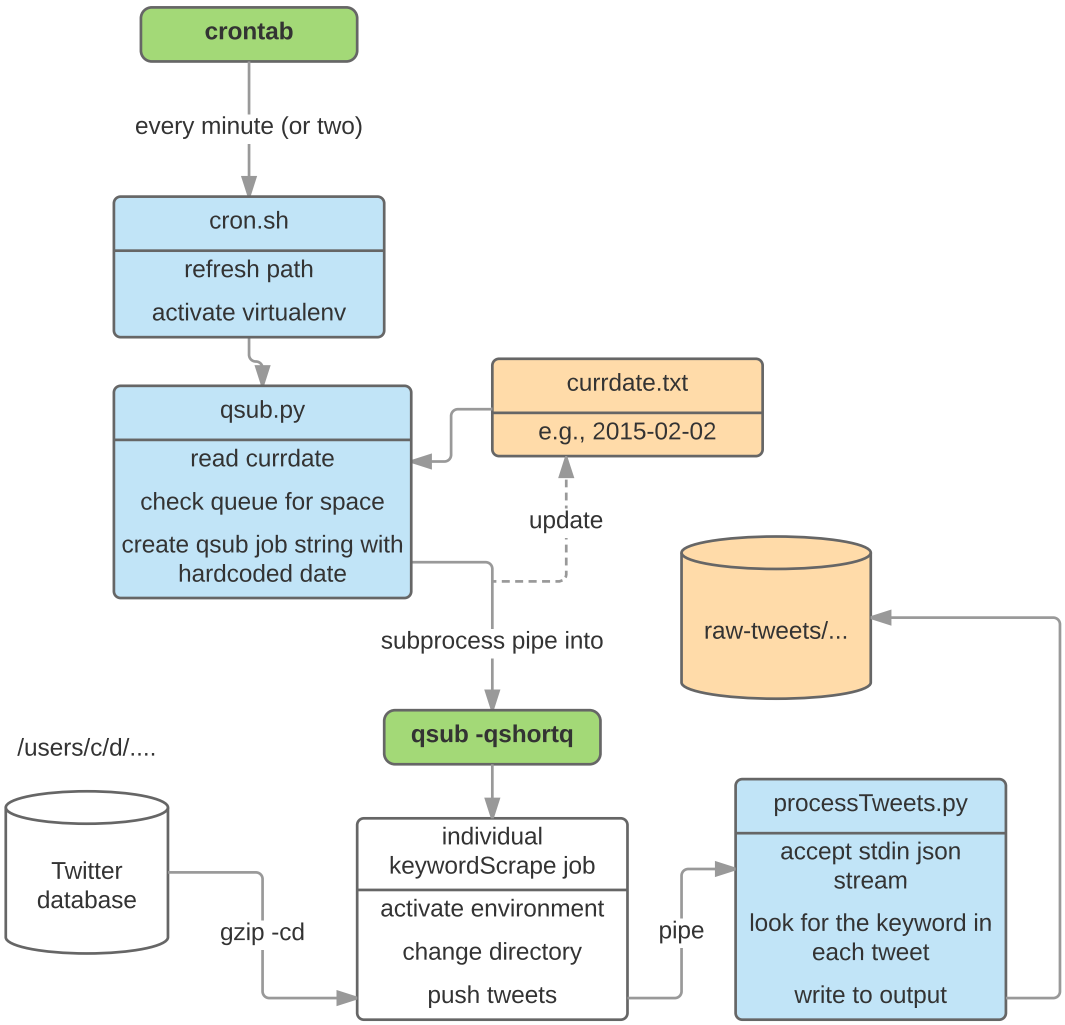


Figure D.1: Schematic of the keyword search framework on the Vermont Advanced Compute Cluster. The three core Python files are invoked by the cron scheduler, and the computation is handed off to compute nodes through the PBS schedulers qsub command. An example code base is provided at https://github.com/andyreagan/VACC-keyword-search.

Appendix E: Infrastructure of Hedonometer.org

The Hedonometer website at http://hedonometer.org is comprised of three main parts: (1) the web server processing including the base code in Python Django, (2) the data processing on the server, and (3) data processing on the VACC. The deployment of the webserver is done using templates and the Ansible tool. Settings and detailed instructions for deploying development and production servers are at https://github.com/andyreagan/hedonometer-vagrant-ansible-deployment. In Figure E.1 we diagram the web server side of the server, included the deployment settings mentioned above and the Django server linked in the caption.

The data side of the server is run separately from the web server side. Nginx serves all files in the `/data` URL ending at Hedonometer, and the files can be browsed at http://hedonometer.org/data/. The files here are used in the front end visualizations across the site, and represent files that loaded for the details-on-demand, as well as the overview files. The structure is optimized for front end performance. The code base is on GitHub at https://github.com/andyreagan/hedonometer-data-munging. As seen in the overall schematic of the server, these files are all inside of `/usr/share/nginx` and they are managed by the `root` user.

Every hour on the hour, these files are updated by a cascade of processes through the cron scheduler. The process is simple enough to do without a diagram: `cron` calls `cronregions.sh` every hour, which simply calls `regions.py` with Python. The `regions.py` loops over dates, looks for files in the `word-vectors` folders for each region, and uses `rsync` to copy over the missing files. Once whole days are downloaded in the 15-minute pieces, it creates the daily sum-

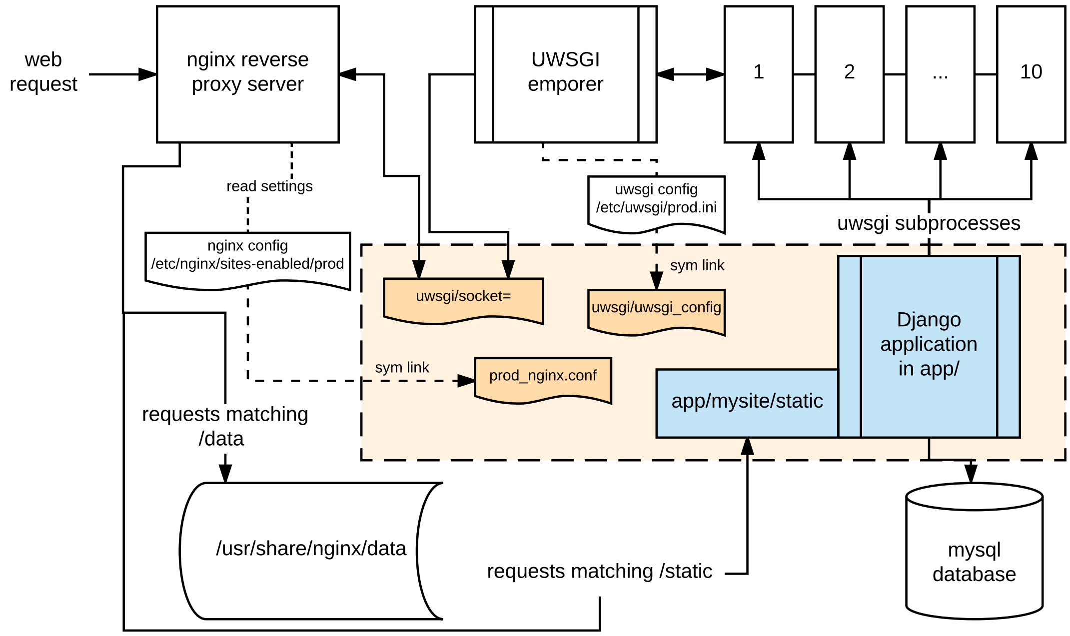


Figure E.1: Schematic of the Hedonometer server architecture. The section in orange is contained in the prod user account, and includes the code stored on GitHub at https://github.com/andyreagan/hedonometer. The settings files for UWSGI and Nginx are written by an Ansible playbook based on the user account under which the code is distributed.

mary files (e.g., `word-vectors/vacc/2017-04-07-sum.csv`) and updates the overall summary at http://hedonometer.org/data/word-vectors/vacc/sumhapps.csv (being wary of duplicates, and keeping the most recent).

The copy of files from the VACC uses `rsync`, which operates of `ssh` and relies on the public key of the server being present on the VACC for seamless access. The files on the VACC are created by a separate process, which is managed in much the same way as the keyword searches in the previous appendix. The cron scheduler runs every hour on the hour, and submits jobs to the PBS queue that turn 15-minute zipped JSON Tweet files into length 10,222 word vectors. The full code for this process is available at https://github.com/andyreagan/hedonometer-VACC-processing.

Appendix F: Online Code Repositories

In this Appendix we collect the repositories that store the code used across all of the projects in this dissertation.

From Chapters 2 and 3, we make the code for the papers publicly available in following two repositories : (1) https://github.com/andyreagan/sentiment-analysis-comparison and (2) https://github.com/andyreagan/core-stories. The code for the online Appendix for the sentiment comparison paper is available at https://github.com/andyreagan/sentiment-analysis-comparison-online-appendix, and for the emotional arcs paper at https://github.com/andyreagan/core-stories-online-appendices. In addition, the code for the Hedonometer website that hosts the interactive emotional arc visualizations is at https://github.com/andyreagan/hedonometer.

From Chapter 4, we provide a link to the repository for each project, in the respective order.

- Collective Philanthropy: Describing and Modeling the Ecology of Giving — code at https://github.com/andyreagan/philanthropy-distributions-code and online appendices at https://github.com/andyreagan/philanthropy-distributions-online-appendices.

- Shadow networks: Discovering hidden nodes with models of information flow — https://github.com/andyreagan/twitter-reply-networks.

- Human language reveals a universal positivity bias — https://github.com/andyreagan/many-happy-languages-appendix.

- Climate change sentiment on Twitter: An unsolicited public opinion poll — code for generating figures is at https://github.com/andyreagan/climate-change-twitter and for performing the keyword search is at https://github.com/andyreagan/climate-change-twitter-keyword-search.

- Reply to Garcia et al.: Common mistakes in measuring frequency dependent word characteristics — the code for this project is contained in the repository for a previous project, at https://github.com/andyreagan/sentiment-analysis-comparison.

- The game story space of professional sports: Australian Rules Football — https://github.com/andyreagan/game-stories-code.

- The Lexicocalorimeter: Gauging public health through caloric input and output on social media — the code for the online Appendix for this paper is at https://github.com/andyreagan/lexicocalorimeter-appendix, code for creating wordshift graphs from the website is at https://github.com/andyreagan/lexicocalorimeter-shifts, and Panometer website is at https://github.com/andyreagan/panometer.org.

- Tracking the Teletherms: The spatiotemporal dynamics of the hottest and coldest days of the year — the website code is at https://github.com/andyreagan/teletherm.org and the online appendix code is at https://github.com/andyreagan/teletherms-online-appendices.

- Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter — https://github.com/andyreagan/livesmatter-keyword-search.