

TSPOONS: TRACKING SALIENCE PROFILES OF ONLINE NEWS STORIES

A Thesis

presented to

the Faculty of California Polytechnic State University

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Kimberly Paterson

June 2014

© 2014

Kimberly Paterson

ALL RIGHTS RESERVED

COMMITTEE MEMBERSHIP

TITLE: TSPOONS: Tracking Salience Profiles Of
Online News Stories

AUTHOR: Kimberly Paterson

DATE SUBMITTED: June 2014

COMMITTEE CHAIR: Professor Alexander Dekhtyar, Ph.D.
Department of Computer Science

COMMITTEE MEMBER: Assistant Professor Foaad Khosmood, Ph.D.
Department of Computer Science

COMMITTEE MEMBER: Professor Franz Kurfess, Ph.D.
Department of Computer Science

ABSTRACT

TSPOONS: Tracking Salience Profiles Of Online News Stories

Kimberly Paterson

News space is a relatively nebulous term that describes the general discourse concerning events that affect the populace. Past research has focused on qualitatively analyzing news space in an attempt to answer big questions about how the populace relates to the news and how they respond to it. We want to ask when do stories begin? What stories stand out among the noise? In order to answer the big questions about news space, we need to track the course of individual stories in the news. By analyzing the specific articles that comprise stories, we can synthesize the information gained from several stories to see a more complete picture of the discourse. The individual articles, the groups of articles that become stories, and the overall themes that connect stories together all complete the narrative about what is happening in society.

TSPOONS provides a framework for analyzing news stories and answering two main questions: what were the important stories during some time frame and what were the important stories involving some topic. Drawing technical news stories from *Techmeme.com*, TSPOONS generates profiles of each news story, quantitatively measuring the importance, or salience, of news stories as well as quantifying the impact of these stories over time.

ACKNOWLEDGMENTS

Thanks to:

- Alex, because without you, I wouldn't be here.
- Dr. Gary Hughes for his help with statistics.
- Eriq Augustine for helping me multi-process, teaching me the #1 thing, and being the best roommate.
- Jeff McGovern for helping me figure out what was important.
- Andrew Guenther for being the Database hero I don't deserve.
- Michael Lady, Paul Biggins, Jeff McGovern, Eriq Augustine, Chris Hunt, Andrew Guenther for being the best guinea pigs.
- AlchemyAPI for granting me the academic tier of their platform.
- And Chris Hunt for the endless support and love.

TABLE OF CONTENTS

List of Tables	x
List of Figures	xii
Part 1 Introduction	1
1 Problem Overview	2
1.1 Impact	5
1.2 Problem Space	7
1.2.1 News Aggregator Selection	7
1.2.2 Techmeme Structure	8
1.3 News Story Life-Cycle	11
2 Background & Related Work	13
2.1 Background	13
2.1.1 Term Frequency * Inverse Document Frequency	13
2.1.2 Topic Modeling and Latent Semantic Analysis	15
2.1.3 MongoDB	16
2.1.4 Linear Regression	17
2.1.5 DBSCAN clustering	18
2.2 Related Work	21
2.2.1 TDT: Topic Detection and Tracking	21
2.2.2 News Story Salience	24

Part 2	TSPOONS Architecture	28
3	Architecture	29
3.1	Parser and Scrapers	31
3.2	Database	31
3.2.1	Cluster Snapshots	34
3.2.2	Links	35
3.2.3	Saliency Profiles	36
3.2.4	Story Clusters	36
3.3	Chainer	37
3.4	Saliency Profile Generator	37
3.5	Story Clustering	37
3.6	Query Engine	38
4	Analysis Pipeline	39
4.1	Dataset	41
4.2	Analysis Tools	41
4.2.1	AlchemyAPI	41
4.2.2	Gensim	42
4.2.3	NLTK	42
4.2.4	Scikit Learn	42
4.3	Linking	43
4.3.1	Snapshot Linking	43
4.3.2	Topical Linking	48
4.4	Entity Extraction	48
4.5	Saliency Profiles	51
4.6	Querying	53
5	Measuring Saliency	57
5.1	Supervised Importance Ranking	58
5.2	Weighted Feature Importance Ranking	62
Part 3	Experiment, Evaluation, and Results	65
6	Snapshot Cluster Chaining Evaluation	66

6.1	Design	66
6.2	Results	68
6.3	Discussion	68
7	Importance Modeling Validation	69
7.1	Design	69
7.1.1	Dataset	70
7.1.2	Method	70
7.2	Results	71
7.2.1	Assessing Judges' Agreement	72
7.2.2	Finding the Best Y-Values	73
7.2.3	Assessing the linear model	73
7.3	Qualitative Evaluative	77
7.4	Discussion	79
8	Weighted Feature Rank Evaluation	81
8.1	Design	81
8.2	Results	81
8.3	Discussion	83
9	Topical Link Evaluation	85
9.1	Design	85
9.2	Results	86
9.3	Discussion	87
10	Querying Evaluation	89
10.1	Design	89
10.1.1	Querying with Importance Ranking	89
10.1.2	Thematic vs. Event Querying	90
10.2	Results	91
10.2.1	Results of Querying with Importance Ranking	91
10.2.2	Thematic vs. Event Query Results	96
10.3	Discussion	96
10.3.1	Importance Ranking with Querying	96
10.3.2	Expanded and Non-Expanded Thematic and Event Queries	101

Part 4	Conclusions	102
11	Future Work & Conclusions	103
11.1	Threats to Validity	106
11.2	Final Thoughts	107
	Bibliography	108
A	Sample Cluster	113
B	Human-Rated Story Results	116
C	Query Results	120
D	Weighted-Feature Rank Full Results	138

LIST OF TABLES

4.1	Features in salience profiles	52
7.1	R^2 values for different Rank Aggregation types	73
7.2	P-value for each feature in the linear model	74
7.3	Stories with high leverage in the training set	75
7.4	Top 20 ranked stories using Linear Regression Rank	79
8.1	Bottom 5 Weighted-feature rank for human-rated dataset.	82
8.2	Top 5 Weighted-feature rank for human-rated dataset.	83
9.1	Results for single cluster where $minPts$ is 10 and ϵ is 0.75.	87
10.1	10 Thematic Queries used for evaluation	91
10.2	10 Event Queries used for evaluation	92
10.3	Top Ten Results from Regular Query (ranked by similarity)	93
10.4	Top 10 results for "Edward Snowden NSA" query ranked by the regression model.	94
10.5	Top 10 results for "Edward Snowden NSA" query ranked by weighted-feature rank.	95
10.6	Mean Average Precision of Thematic vs Event Queries with and Without Query Expansion	96
B.1	Training set for human judgment of importance	119
C.1	Results of the "Edward Snowden NSA" query	128
C.2	Top 50 Ranked query results for "Edward Snowden NSA" query ranked by weighted-feature rank.	132

C.3	Top 50 results for "Edward Snowden NSA" query ranked by the regression model.	137
D.1	Weighted-feature rank for human-rated dataset.	142

LIST OF FIGURES

1.1	Techmeme.com Screenshot on Dec. 16, 2013	10
2.1	Point p is directly density-reachable from point q , but q is not from p [26].	19
2.2	Point p is density-reachable to q , but the relation is still not symmetrical [26].	20
2.3	Points p and q are density-connected and are a part of the same cluster [26].	20
3.1	TSPOONS workflow diagram.	30
3.2	Snapshot cluster ordering	32
3.3	Techememe.com structure	33
3.4	<code>chain_link</code> document example	34
3.5	A link document example without entities (see Section 4.4 for details on entities)	36
4.1	TSPOONS analysis pipeline	40
4.2	Single Chain from dataset. Snapshot clusters split and then split again before merging later.	47
7.1	Statistical test for agreement among judges	72
7.2	Graphical summary and Anderson-Darling test for Normality of the residuals for the linear model.	76
7.3	Test for equal variances among levels of the fitted values.	77
7.4	Individual Measures Chart for Residuals	80
10.1	Precisions for thematic queries with expansion graphed against number of links.	97

10.2 Precisions for thematic queries without expansion graphed against number of links. 97

10.3 Precisions for event queries with expansion graphed against number of links. 98

10.4 Precisions for event queries without expansion graphed against number of links. 98

10.5 Precision comparison of event queries with and without expansion. . . 99

10.6 Precision comparison of thematic queries with and without expansion. 100

Part 1

Introduction

CHAPTER 1

Problem Overview

News space is a relatively nebulous term that describes the general discourse concerning events that affect the populace. The source of the news space comes from the media that produces content every minute of every day on a variety of topics. The content catalogs what is important to society in the form of individual articles published by multiple sources, the collection of which we call the *discourse*.

In the past, sociologists and linguists have focused on qualitatively analyzing news space in an attempt to answer big questions about how we relate to the news and how we respond to it; the field of discourse analysis relies heavily on qualitative analysis of language and culture to learn more about humanity through its media. For news space, we want to ask: when do stories begin? What stories stand out among the noise? In order to answer these big questions, we need to track the course of individual stories in the news. By analyzing the specific articles that comprise stories, we can synthesize the information gained from several stories to see a more complete picture of the discourse. The individual articles, the groups of articles that become stories, and the overall themes that connect stories together all complete the narrative about what is happening in society. However, this process need not remain

a manual task; computational methods can enhance qualitative analysis as well as provide the framework for quantitative analysis of large swaths of news space.

The goal of this work is to provide a framework for qualitatively and quantitatively profiling news stories and analyzing the breadth and the depth of news space: looking at stories over time and their impact, and looking at the discourse at a given time and seeing what stories dominate. By providing automated tools for gathering and performing this kind of news story analysis, we can easily analyze more data than if we performed the task manually. However, if we rely on an automated framework, we must develop a model for news stories that breaks down our human understanding of the news story life cycle into quantifiable features.

Stories are intangible re-tellings of events, either fictional or factual, that can be shared collectively among people. News stories are a subset of stories that cover (mostly factual) events that are happening, or have happened, and that are relevant to the target audience. As time passes, stories rise and fall, gathering information and interest as they develop, until people ultimately move on or there is no more information or content to be gleaned from the events that sparked the story. When news stories break, perhaps their most tangible artifacts are the written articles that describe the stories' different aspects, as told through the lens of each writer and the environment in which he or she wrote the article. When a story dies, the only remnants are its articles and our memory of it.

Each article encapsulates a piece of a story; taking the union of all articles about a given topic, or, more specifically, a news story, gives a complete-as-possible written report of the story. Reasonably, there is no way to capture all articles about a story, but if we have a significant portion of the articles, we can measure the story's impact and development over time.

If one looks at the discourse over time, one will find a variety of diverging narra-

tives; on any given day, the articles in the discourse will talk about several different events or topics. A group of articles that discusses the same topic can be combined together to form a cluster, with the cluster growing and changing over time. New story clusters may gain articles, merge with other clusters, diverge into multiple, related clusters, and eventually die out when the topic is no longer pertinent. Because of this, news story clusters have a temporal dimension; looking at the set of articles about a story at different times yields different pictures of the story. Opinion articles in the cluster may replace short, factual articles, or the diversity of information about a topic or story may develop and perhaps split the story into two or more stories. To measure this change, we can snapshot the news story cluster at given times and use the sequence of snapshots to look at the development of the stories.

Like an article, a cluster of articles represents a piece of a news story, but perhaps a more-complete view of the story. In order to understand the development of a news story, we need a reliable way to 1) identify news story clusters, 2) track the development of each cluster over time, and 3) determine when a story develops into multiple stories.

If we look at the overall discourse as being comprised of these evolving news story clusters, we can track the way the discourse evolves over time as well. We can measure this evolution by modeling news space as clusters of articles with a temporal dimension. We do this by collecting news story cluster snapshots at a predetermined time interval, tracking the development of clusters as time progresses, and analyzing the content of the clusters. With the analysis, we have a more complete picture of the life cycle of a news story. Linking these snapshots together produces a snapshot chain, which approximate news story clusters. Modeling stories as these chains enables us to measure the impact of a single story over time. Analyzing these snapshot chains also provides insight into detecting the genesis of important stories; with the hindsight of news story snapshots, we can examine and extract the features of breaking news

stories and can better detect the importance of a story early on.

Given a news story cluster, we can extract the themes and topics from the articles to expose some news story features. By understanding the topics of a news story, we can compare and contrast stories and measure the density of topics at any given snapshot. This is useful for determining what topics dominate public discourse, and we can measure the life cycle of a theme (or a recurring topic) as topics appear and disappear from discourse. Discovering news story topics enables this kind of meta analysis because we can aggregate the topics at a given time from all news stories.

The most useful result of tracking and analyzing news stories is being able to answer questions about the data. Once we have a structured view of the article clusters that comprise a story, we can begin mining the stories for information such as tracking salience of a news story, measuring the emotion surrounding a news story (i.e. does the story provoke emotion in the discussion) or looking for tipping points when articles develop rapidly. Using this information, we not only improve our understanding of stories, but we can improve the technologies that handle news stories such as recommendation systems, which can better recommend articles based on the meta information gleaned from our analysis.

1.1 Impact

This work introduces the Tracking Salience Profiles Of Online News Stories framework, or TSPOONS, a framework for analyzing news stories, which profiles news stories and answers two main questions: what were the important stories during some time frame and what were the important stories involving some topic. These questions span both the horizontal and vertical slices of news space. They span the horizontal, by looking at a story, topic or entity's impact over time, and spans the

vertical by looking at a specific time frame and determining the relative impact of all stories, topics or entities during that time. TSPOONS structures the results of analyzing individual news stories as salience profiles, detailing the importance of the story, sentiment, and impact over time.

The main contributions of this work are as follows:

1. TSPOONS generates a structured view of news space by developing a model for news stories.
2. TSPOONS generates a detailed profile of each news story, including relevant features like duration, impact, entities involved, and salience.
3. TSPOONS provides a query framework for retrieving stories based on topics.
4. TSPOONS provides a query framework for retrieving stories that began within a time period.
5. TSPOONS provides heuristics for deciding which stories were the most important.
6. TSPOONS attempts to group stories into topical clusters for identifying themes within news space.

The rest of the document is organized as follows. Chapter 2 covers background and related work. Part 2 discusses the design, implementation, and capabilities of TSPOONS. Part 3 outlines the experiment and evaluation of the TSPOONS framework. Part 4 explores the impact of TSPOONS and outlines future work.

1.2 Problem Space

In order to track the development of news stories and measure the salience of stories and entities within news events, TSPOONS collected articles from *Techmeme.com*, a news aggregator that focuses on technical news, involving tech companies, new tech, and all issues surrounding the current state of technology. *Techmeme.com* updates its content regularly, and clusters articles together by topic/story in ranked clusters. The dataset used for evaluating this work contains technical news stories from a period beginning October 1, 2013 at 12:00 AM and spanning to May 1, 2014 at 12:00 AM.

1.2.1 News Aggregator Selection

Although TSPOONS draws its content from *Techmeme.com*, TSPOONS could use many existing news aggregators to perform its data collection, or we could implement an aggregator that fits our criteria.

To work with TSPOONS data processing pipeline, the aggregator must do the following:

- regularly update the "front page" with new content, at least every hour,
- structure aggregated news articles into "story" groups,
- elect a single article as the headline article for the story,
- contain links to articles online
- provide a rank for news stories, where the most important are closer to the top of the page.

We chose to use *Techmeme.com* because it fit our criteria for a sufficient news aggregator. *Techmeme.com* updates the content regularly, and groups articles into story clusters, where one article is the headlining story. *Techmeme.com* also structures the stories by importance or prominence, placing the most interesting or impactful clusters near the top of the page.

TSPOONS relies on the aggregator structuring news content in this way in order for the content to be parsed and structured in the way TSPOONS is built to handle. Using a different aggregator would only require building a different scraper and parser that would process and structure the data (see Section 3.1). In the interest of time, we chose to use *Techmeme.com* solely rather than using multiple aggregators because it tends to perform well, capturing what we would consider major events in technical news. Although it is not as popular as other aggregators such as Google News [2], *Techmeme.com* still has many daily readers, about 10,000 [14], and has been called "the favorite news website of technology industry insiders" [8] and "one of the first Web sites loaded on Silicon Valleys laptops and iPhones each morning" [33]. In addition to being respected in the technical community, *Techmeme.com* provides a useful data collection feature—a time machine that allows TSPOONS to go back in time and scrape *Techmeme.com* pages from any time in the past. Although not necessary, this time machine feature expanded the data collection time period, allowing TSPOONS to collect historic data from before the TSPOONS data collection processes were written.

1.2.2 Techmeme Structure

Techmeme.com is a automated news aggregator, supported by human editing, that focuses on stories about technology, clustering related articles together and ranking the stories by importance. *Techmeme.com* was founded by Gabe Rivera in 2005 and

started as a fully-automated news curator, similar to Google News [1]. In 2008, *Techmeme.com* introduced human editing to the headlines featured on *Techmeme.com*; old stories disappeared faster after the change and breaking news appeared more quickly [37]. According to Rivera, "Interacting directly with an automated news engine makes it clear that the human+algorithm combo can curate news far more effectively than the individual human or algorithmic parts" [37]. *Techmeme.com* and its sister sites, *Memorandum.com* [4], *WeSmirch.com* [7], and *Mediagazer.com* [3], collect and cluster news stories continuously, as the clusters grow, shrink, and change when new stories arise and old ones become stale.

TSPOONS collects all article data from *Techmeme.com*'s Top News section and takes a snapshot of the clusters on within the Top News section, or front page, every hour. Sponsored posts are ignored and only organic clusters are collected. TSPOONS parses the page, storing the clusters, their positions on the page, and all of the articles in the cluster. TSPOONS then scrapes and stores each article's content, as described in Section 3.1. As the hours advance, the content of *Techmeme.com*'s front page changes; consequently, TSPOONS must keep track of stories from hour to hour, building snapshot chains, and determining when clusters appear and disappear, as described in Section 4.3. In the Top News section, *Techmeme.com* structures the clusters of news stories hierarchically, with the more important stories appearing at the top of the page and any sub-stories of a cluster placed directly below and indented inward, as shown in Figure 1.1. Each cluster has one headline story, where the content of the headline is human edited and may vary from the headline presented on the source article, and any number of "More" links, which represent different sources for the same story. In some cases, there may be relevant tweets associated with the story: these are linked in the "Tweets" section of the cluster, with the Twitter handles of the users listed with links to the tweets.


Mobile Mini Open Links In New Tab Archives Like Follow

Techmeme December 16, 2013, 6:10 PM Search

HOME RIVER LEADERBOARD ABOUT SITE NEWS SPONSOR MEDIAGAZER MEMORANDUM WESMIRCH

Top News

Andy Greenberg / Forbes:


An NSA Coworker Remembers The Real Edward Snowden: 'A Genius Among Geniuses' 

— Perhaps Edward Snowden's hoodie should have raised suspicions. — The black sweatshirt sold by the civil libertarian Electronic Frontier Foundation featured a parody of the National Security Agency's logo ...

More: CBS News, The Switch, The Register, The Verge, SiliconANGLE, Reuters, Business Insider, Mashable, Softpedia News, emptywheel, PC World, Pixel Envy and SecurityWeek

Tweets: @jesselynradack, @weldpond, @qthrul and @youranonnews

Spencer Ackerman / Guardian:

NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report 

— Our take on five things the spy agency would like the public to believe about its vast surveillance powers — The National Security Agency is telling its story like never before. Never mind whether that story is, well, true.

More: Slate and Yahoo! News. **Tweets:** @ggreenwald, @tonyromm, @schestowitz and @bartongellman. [See also Mediagazer](#)

Sara Morrison / The Wire:


'60 Minutes': NSA Good, Snowden Bad

More: Poynter, Errata Security, The Nation, CBS News, Gizmodo, emptywheel, Techdirt, Computerworld and Firedoglake

Tweets: @jbrodtkin, @jaycstanley, @qhardy, @csoghoian, @nickbilton, @ggreenwald, @benpopper, @ggreenwald, @tonyromm and @poumecoffee

[See also Mediagazer](#)

Josh Gerstein / Politico:


Judge: NSA phone program likely unconstitutional 

— A federal judge ruled Monday that the National Security Agency program which collects information on nearly all telephone calls made to, from or within the United States is likely unconstitutional. — U.S. District Court Judge Richard Leon found ...

More: ZDNet, Guardian, Ars Technica, Techdirt, The Verge, CNET, Daring Fireball, iMore, Gigaom, VentureBeat, ACLU, Motherboard, Yahoo! News, The


Sponsor Posts

Microsoft:

Queensland Government partners with Microsoft to bring Office 365 to 149,000 government employees 


— Queensland, the second-largest state in Australia, will partner with Microsoft to bring Office 365 ...

Atlassian:

Every team needs kick-ass code reviews 


— Code reviews help spread knowledge and coding best practices throughout a team. In this article we'll take a look at why code reviews are important, and how to optimize the practice.

Zoho Blogs:

2 Ways Live Chat Can Enhance Your Customer Service 

— Landing a new customer is 5x more expensive than keeping an existing customer. How do we keep our customers happy? By providing outstanding support.

Silicon Valley Bank:

Strengthening the U.S. Innovation Economy by Ending Abusive Patent Litigation 

— In Silicon Valley Bank's Innovation Economy Outlook 2014 survey, we asked executives what they see in today's patent system ...

Sponsor Techmeme

About This Page

This is a Techmeme archive page. It shows how the site appeared at 6:10 PM ET, December 16, 2013.

The most current version of the site as always is available at our [home page](#). To view an earlier snapshot click [here](#) and then modify the date indicated.

From Mediagazer

Ted Johnson / Variety:

Concept for Live TV On-the-Go Was Around Long Before Aereo

Rick Edmonds / Poynter:

Newspaper industry narrowed revenue loss in 2013 as paywall plans increased

Hussain Al-Qatari / Associated Press:

Kuwait court shuts 2 newspapers over coup articles

Figure 1.1: Techmeme.com Screenshot on Dec. 16, 2013

1.3 News Story Life-Cycle

From first-hand observance, news story snapshot clusters on *Techmeme.com* change in one of several ways as the hour advances: they either appear, continue, split, merge, die, or re-surge.

- **Genesis:** genesis happens when a new story appears and there is no snapshot cluster that discusses the same story in the hour before. This is the beginning of a story, which usually begins with one article link and grows over time.
- **Continuance:** continuance happens when a news story can be traced from the previous hour to the current hour. In other words, the story is developing or is still relevant enough to warrant front-page placement. The snapshot cluster may gain articles and the headline article may change as the snapshot cluster reshapes around other topics.
- **Split:** splitting is where the articles in one snapshot cluster are divided among two or more snapshot clusters on the next hour. The snapshot cluster may develop child snapshot clusters or may create two top-level snapshot clusters on the page, signifying two distinct stories.
- **Merge:** merging happens when the articles from two snapshot clusters are combined into one larger cluster. This happens when two snapshot clusters are discussing the same subject and converge, which means they become one snapshot cluster.
- **Death:** story death occurs when there is no snapshot cluster at the current hour that carries on the story from a previous hour. This happens when no new articles are being produced about the story, or that the story is no longer important enough or developing to warrant a front-page spot.

- **Resurgence:** resurgence happens when a snapshot cluster discusses the same topic/event as a cluster chain that had already died.

Part of this work involves determining when these processes occur in order to track the development of a single snapshot cluster and news story over time.

CHAPTER 2

Background & Related Work

2.1 Background

TSPOONS uses several well-known algorithms to complete its analysis which we outline in this section for reference.

2.1.1 Term Frequency * Inverse Document Frequency

Term Frequency * Inverse Document Frequency [40], or TF*IDF, was developed as a tool for document indexing, where each document is represented as a vector of keyword weights over a vocabulary of every word in the entire corpus. The vocabulary V typically excludes common stopwords and is normalized by stemming each word, or removing suffixes and simplifying a word down to its root. This creates less variability in the corpus under the assumption that words like "dog" and "dogs" are inherently the same keyword, and should not be treated as different words. When modeling documents with TF*IDF vectors, a document $d_j \in D$, where D is the set of documents in a corpus, becomes a vector over V that has TF*IDF weights for each word that appears in d_j .

TF*IDF relies on two main intuitions about what defines a keyword:

1. The more a word appears in a document, the more important it is to the document.
2. The less frequently the word appears in other documents, the more important each occurrence of the word is.

As a result, TF*IDF wants to reward terms in a document that appear frequently in the given document, but less frequently in other documents in the collection. The term frequency of a word captures the first rule and the inverse document frequency captures the second. To calculate the TF*IDF for every word in all documents $d_j \in D$, we simply multiply the word's term frequency (tf_{ij}) by its inverse document frequency (idf_i).

The term frequency tf_{ij} of a word in a given document $d_j \in D$ is calculated by counting the number of times w_i appears in the document. However, since the length of documents vary, we want to normalize the term frequency to prevent longer documents having significantly larger term frequency values than shorter documents. We get the normalized term frequency of a word by taking the frequency of the word and dividing it by the maximum frequency of all words in d_j , as shown in Equation 2.1.

$$tf_{ij} = \frac{f_{ij}}{\max(f_{1j}, f_{2j}, f_{3j}, \dots, f_{Mj})} \quad (2.1)$$

The second piece of the TF*IDF computation is the inverse document frequency of all words $w_i \in V$. The document frequency is the number of documents in which the word w_i appears at least once. The equation for document frequency is given in

Equation 2.2.

$$df_i = |d_j \in D | f_{ij} > 0| \quad (2.2)$$

The document frequency captures the notion of keyword popularity, where a high document frequency means a keyword is common to many documents. However, a keyword that appears in every document is less meaningful according to the second rule, therefore we want to take the inverse of the document frequency, as shown in Equation 2.3.

$$idf_i = \log_2 \frac{n}{df_i} \quad (2.3)$$

We first normalize the document frequency like we normalized the term frequency by taking the df_i over the total number of documents n in the collection. By taking the inverse and the log of the document frequency, we calculate higher values for more infrequent words and lower values for highly frequent words, thus fulfilling the second rule.

If we take the TF*IDF together, we get the following equation, where we calculate the weight of every word w_{ij} in every document d_j :

$$w_{ij} = tf_i \cdot idf_i = \frac{f_{ij}}{\max(f_{1j}, f_{2j}, f_{3j}, \dots, f_{Mj})} \cdot \log_2 \frac{n}{df_i} \quad (2.4)$$

2.1.2 Topic Modeling and Latent Semantic Analysis

TSPOONS uses two mathematical processes for extracting semantically interesting features from news stories; each story is treated as a document, and TSPOONS extracts the topics of the story by using Latent Dirichlet Allocation (LDA) and rep-

resents the stories as vectors using Latent Semantic Analysis(LSA) for the topical clustering task.

LDA is a process that extracts the latent, or hidden, topics from a set of documents, producing a probability distribution of words, where the higher the probability, the more likely the document belongs to that word, or topic. LDA requires a set number of topics as a parameter to the algorithm, which may require tuning to determine which number produces the most intuitively accurate summarization of what the documents are "about." TSPOONS uses a simple implementation of LDA from the Gensim topic modeling package to generate a small number of topics for each news story (see Section 4.2.2).

LSA tries to identify patterns across documents, making the assumption that words which appear in similar contexts have similar meanings. LSA starts by creating a matrix of word counts per document, where each column is a unique word in the document and each row represents a document. Through the mathematical process of Singular Value Decomposition (SVD) [40], the size of the matrix is reduced to only the most salient features, or words, and the matrix is transformed into semantic space. The result of the process creates vectors for each document with the most meaningful features as the values within the vectors. The document vectors can be compared and can be grouped by similarity using standard clustering algorithms, or indexed in an information retrieval context, which is typically referred to as Latent Semantic Indexing, or LSI. TSPOONS uses LSA for clustering, as described later in Section 4.3.2.

2.1.3 MongoDB

MongoDB [5] is an open-source, NoSQL document database that provides rich querying and flexible schemas for JSON documents. MongoDB has the concept of

collections, which are comparable to groups of similarly-structured documents. Each document is a JSON object, with only one required field, the `object_id`, which is automatically generated as the key for each document. Documents can have multiple fields, but documents in a collection are not required to have the same fields (this is the idea of a flexible schema).

To retrieve documents, MongoDB has a simple querying language that allows users to retrieve documents based on the fields within the documents. However, because MongoDB does not support joining documents, querying retrieves a list of single documents.

Document store databases are useful for applications with large amounts of data and when joining is not required, but flexible schemas is required. TSPOONS uses MongoDB for these reasons.

2.1.4 Linear Regression

Linear regression is a statistical approach for modeling the relationship between a dependant variable Y and one or more regressor variables X . Linear regression attempts to fit a line over observed data to find a line of best fit. Given a set of values $\{y_i, x_{1i}, x_{2i}, \dots, x_{pi}\}_{i=1}^n$, linear regression assumes that the relationship between dependent variable y_i and the parameter vector (p-vector or β) of the regressors x_i is linear and attempts to find a line of best fit for the data points. The p-vector can be thought of as weights to the individual X values that, when applied, generate the line of best fit. The epsilon value represents the noise or error term that accounts for the variability in Y . The line of fit can be written as:

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \epsilon_i \quad i = 1, \dots, n \tag{2.5}$$

or simplified to be:

$$y = X\beta + \epsilon \tag{2.6}$$

Ordinary Least Squares linear regression minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation. Simple least squares linear regression attempts to use all X variables in the equation for the line, even if the values are not correlated with the y value. Often, we can achieve more meaningful results if we determine which variables are statistically correlated with the resulting Y values and only use these X values in the computation. Least Angle regression [25] (or LARS) provides an algorithm for determining which X values are meaningful by systematically removing values that negatively affect the fit of the line. The Least Absolute Shrinkage and Selection Operator (LASSO) is an optimization that penalizes the model for choosing large β values, meaning the coefficients for single X values can drop to 0, effectively removing the X variable from the equation [42].

Once we generate a model for the observed data, given the known Y values, we can use the model to predict new Y values on unseen X data. Therefore, linear regression is a simple and useful tool for predicting values based on a trained model.

2.1.5 DBSCAN clustering

Density Based Spatial Clustering of Applications with Noise, or DBSCAN [26], is a density-based clustering algorithm that attempts to separate dense clusters from noise by taking into account a neighborhood distance (ϵ) and a minimum number of points (*MinPts*) that must be within a radius of every single point in the cluster.

DBSCAN relies on the idea of *density-reachability* to form clusters. A point q in

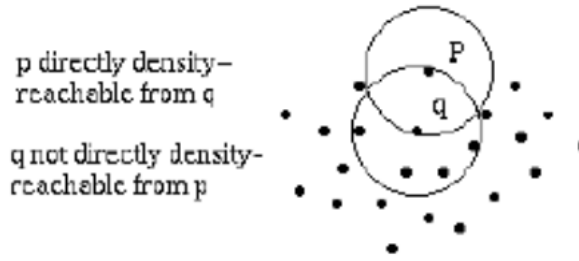


Figure 2.1: Point p is directly density-reachable from point q , but q is not from p [26].

directly density-reachable from another point q if p is not farther away than a distance ϵ (or within its ϵ -neighborhood) and p is surrounded by enough points to say that p and q are in a cluster.

We can define p 's ϵ -neighborhood, denoted by $N_{Eps}(p)$, as $N_{Eps}(p) = \{q \in D \mid dist(p, q) \leq \epsilon\}$. We know q is *directly-density-reachable* from p if $p \in N_{Eps}(q)$ and $|N_{Eps}(p)| \geq MinPts$.

The *directly density-reachable* relationship between p and q is not symmetrical, since as Figure 2.1 shows, the point q is not *directly density-reachable* from p although p is from q .

We can say that two points are also *density-reachable*, but not directly. They can be just *density-reachable* through a chain of points rather than a single point. A point p is *density-reachable* from a point q with respect to ϵ and $MinPts$ if there is a chain of points p_1, \dots, p_n , where $p_n = p$ such that p_{i+1} is directly density reachable from p_i , as shown in Figure 2.2.

To fully form the clusters, DBSCAN applies the notion of *density-reachability* from all points in a cluster to a single point o . If points p and q are both *density-reachable* from o , then they can be said to be *density-connected*, as shown in Figure 2.3. Points p and q can be said to be in the same cluster if p and q are *density-connected* with respect to $MinPts$ and ϵ .

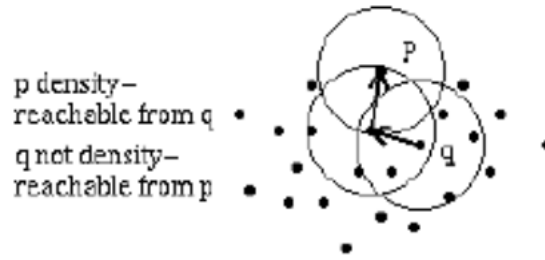


Figure 2.2: Point p is density-reachable to q , but the relation is still not symmetrical [26].

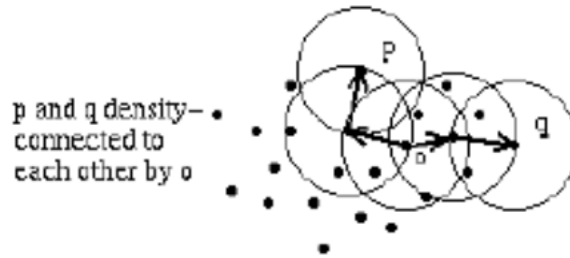


Figure 2.3: Points p and q are density-connected and are a part of the same cluster [26].

Any point that is not *density-connected* to another point can be considered noise, and is not a part of any cluster.

The procedure for DBSCAN starts with an arbitrary point p in the dataset. DBSCAN finds all of the points that are *density-reachable* from p , forming a cluster. If there are no points that are *density-reachable* from p , then DBSCAN moves onto the next point. DBSCAN will merge clusters together if they are close enough together, eventually determining when all points are clustered or are considered noise.

DBSCAN is a useful algorithm for datasets that may contain a lot of outliers or noise and that contain uneven cluster sizes or unusually shaped clusters.

2.2 Related Work

In accomplishing its goal, this work draws from two major bodies of work involving news story analysis. The first is in the computational task of Topic Detection and Tracking (TDT), outlined in Section 2.2.1 and the second is the sociological approach to measuring salience, outlined in Section 2.2.2.

2.2.1 TDT: Topic Detection and Tracking

Topic Detection and Tracking (TDT) is a body of research focused on event-based news organization that also provides an evaluation framework for each of the subtasks; TDT began as a DARPA-sponsored research program that opened up for competitive evaluations in the early 2000's, with James Allen as one of the pioneers [21, 23]. TSPOONS tackles three of the main tasks associated with TDT: topic tracking; topic detection; and link detection. The topic tracking task involves detecting stories that discuss a known topic where topic detection involves determining when stories discuss the same topic, which leads into link detection; link detection involves determining when two stories are linked because they discuss the same topic.

Topic Clustering

Topic detection is a clustering task [22] that involves placing a story into a cluster based on it's topic similarity to other stories in the cluster. Redefining this task using our definition of news story, this task involves clustering articles into topical clusters which discuss a single story. We can conceptualize topics generally as in "wearable tech", or "online privacy" topics, or we can consider topics more specifically as tied with events, such as "Facebook's acquisition of WhatsApp" or "Edward Snowden's NSA leaks."

Techmeme.com does the initial clustering for us by grouping related articles together — where each "cluster" discusses a single event or current state of a story. At each hour, we can determine what happens to the story by measuring how articles persist over time, without needing to delve into the content of the articles. However, once a thread of a cluster ends, as in, it no longer shows up on the page, the story is considered dead, unless a new development happens later on. We can detect the resurgence of a story by analyzing the content of articles and detecting when two cluster chains are about the same topic. By using clustering, we can compare snapshot chains to later chains to see if a story resurges.

Event Link Detection

The link detection task involves determining when two events are linked. Redefining the task in our terms, this means detecting when one cluster chain represents a development of a story discussed by a previous cluster chain. Typically, this task is achieved by analyzing the textual content of a story, usually by looking at keyword matching or finding a correlation between keyword occurrences in two stories. Researchers usually take the approach of measuring the presence of the same keywords across documents, as Zhai and Shah did [46], or by measuring the similarity/correlation of the word distributions using metrics like Term Frequency * Inverse Document Frequency (TF*IDF) as Hsu and Chang demonstrated [29]. These approaches are fairly simple to implement and are relatively naïve because it assumes that if the documents have similar distributions of keywords, they are "about" the same topics.

Feng and Allen [27] made an early attempt at link detection task called event threading, in which they tracked a single event across multiple stories and built a dependency tree of documents; however, event threading provides no information about what the dependency or relationship is, rather it provides a binary, "Are these

two events related, yes or no?”, result. Ideally, we could learn from the field of discourse analysis, which attempts to use human judgement to arrive at a label for the dependency between document: e.g. this document also discusses privacy and the NSA. The problem with discourse analysis is that the links are inherently subjective and one comparison with multiple human judges produces varying answers. Discourse analysis is also done with humans not computers. Ideally, there exists a middle ground between these two.

Feng and Allen [27] believed event threading was on the right track and implemented a more refined method called incident tracking. Their baseline was comparing the TF*IDF of two documents and creating a link if the TF*IDF similarity was above a given threshold. They proposed another algorithm where the TF*IDF was used with other features like main characters, locations, time stamps, key verbs, etc. that groups the articles into the same incident, then they use rules and classification to determine if two incidents are related. By incorporating more factors into the analysis, the researchers found a more intelligent method for capturing the nuances of the relations between articles and subsequently, the relation between two different events.

Another similar approach proposed by Paul Waring is relating events based on the "who", "what", "where", "when" features of a story [44]. These attributes are strong measures that augment the textual similarity of two event clusters by placing importance on keywords that represent the answers to the four "w" questions. Waring defines a set of rules for determining when events are linked:

- If the events have all these attributes in common, they are identical and should be placed in the same cluster.
- If the events share some, but not all, attributes, then they are related and their event clusters should be linked.

- If the events have no attributes in common, they are unrelated and they should not be linked.

The research surrounding the task of TDT shows that a composite approach to linking events or stories is better than a naïve approach of keyword matching.

2.2.2 News Story Saliency

Saliency, in the domain of mass media, can be thought of as a measure of importance/prominence for a given story, entity, issue, etc. Historically, tracking the saliency of an issue involves extreme manual effort whereby humans must diligently collect news sources, process the content, and manually extract the topics and entities contained within the text. There are some attempts at automated analysis, but much of the current research is still manually-bound.

The goal of automatically measuring news story saliency would be to avoid manual effort by creating an automated framework that can handle higher-volume datasets and can be tuned to answer questions rapidly.

Measuring Saliency

The idea of measuring saliency is prominent in research surrounding agenda-setting in the media. Saliency of issues, events, stories, and entities has been measured in order to determine how they influence public knowledge and opinion, with news story/issue volume and coverage as the primary metric of saliency.

Typically, researchers aggregate the number of times certain keywords are mentioned in articles. In some regard, an article that contains certain keywords is naïvely about a topic associated with these keyword. Keyword mentions are a simple way to represent topics and collect counts for coverage of a topic in the news, although the

approach may not capture the nuances of coverage.

Coverage alone may not be the best metric for measuring salience, as Kiouisis explains [31]. Coverage, or attention, as Kiouisis calls it, is useful under the intuitive assumption that the more we talk about a topic, the more important it is or seems to us. However, this misses two other aspects of importance: the way in which the material was presented to us and the controversy surrounding the topic. Kiouisis outlines these two other features as prominence, or the position of the text within a medium, and valence, which either captures a dimension of sentiment polarization or conflict (variability in opinion). Kiouisis studied a combination of these three over a manually-collected dataset of *New York Times* articles about the 2000 presidential election. Kiouisis analyzed hand-chosen topics in the dataset and counted the number of stories about these topics using keywords. To analyze the data, Kiouisis used principle component factor analysis with varimax rotation to see if his conceptual model corresponded with empirical findings. He found that their results were promising, but not conclusive.

Although Kiouisis was unable to show solid proof that a multi-dimensional approach to measuring salience was better, his approach matches our human intuition that the salience of a topic *depends* on many things. His work also shows that coverage, although uncomprehensive, is a good starting point for measuring importance of a topic in the media.

Automation

In all aspects of studying news space and measuring salience, gathering and preparing the data takes the majority of the work over performing the analysis. Ideally, not only the analysis would be performed by computers, but the gathering and preparation stages would be automated as well. There has been some research into

using computational methods for measuring salience; for example Scharl and Weichselbraun implemented a system for investigating media coverage of US presidential elections that used frequency of candidate references and opinions about candidates to measure sentiment to summarize the most important topics associated with the candidates [41]. Scharl and Weichselbraun used fairly simplistic methods for analyzing their data, namely keyword presence and co-occurrence. However, similarly to the TDT approaches, more nuanced approaches to analyzing salience may provide better results. The goal of TSPOONS is, in part, to enhance the ability of sociological researchers to examine news content, allowing them to use more sophisticated methods for measuring salience, with as little burden on the collection and processing stages of the analysis.

The Usefulness of Salience

Salience is inherently an interest of social science, but more and more as humans rely on automated sources for reading the news, it becomes an issue of technological interests. In some regard, measuring salience is a scientific endeavor, to examine salience for the sake of understanding how humans prioritize their attention to issues and understand the world; it is also important for researchers to understand salience so that the media can be aware of how their coverage of current events affect people's opinions. However, being able to analyze news content computationally and answer fundamental questions about the importance of topics, issues, events, and entities enables social scientists to answer questions more rapidly and efficiently, and provides the groundwork for media aggregators and generators who are using technological means to distribute, rank, and display news content.

The technological uses for automated news story tracking and salience measuring come in two main forms: stream ranking and retrospective analysis. Stream analysis is

heavily researched with applications like recommender systems and news aggregators, where you have a set of articles generated every minute that need to be grouped, ranked, and recommended to users; the system must present the articles in a *processed* manner, rather than an unintelligently ranked list of documents and, therefore, must make decisions about the documents based on what it can glean from the article content. The system must answer the question: given a set of articles generated from the current time window, which articles take priority over others and how is a system supposed to rank and present the articles without a human understanding of what makes some stories more important or useful than others?

Search engines encounter this problem as well: ranking documents is a fundamental problem for delivering content to users on the web. However, the window of analysis is wider, encompassing not only what is being generated at the current moment, but what has been generated in the past. When we have knowledge of past articles and their importance, studying the importance or prestige of a topic, article, entity or event in the past becomes useful in understanding what features make some stories fundamentally more important than others. Retroactive importance is not directly useful in the case of a streaming system, but what we learn from retroactive analysis can help inform feature selection for systems that need to analyze streaming documents in order to make decisions about how to present information to users.

Part 2

TSPOONS Architecture

CHAPTER 3

Architecture

TSPOONS is made up several processing components, beginning with scraping the content and ending with the query engine. Each of the processes shown in the workflow diagram in Figure 3.1 is outlined in this chapter; the work of each process is discussed in detail in Chapter 4 and 5.

The workflow of TSPOONS begins with scraping content from *Techmeme.com* at every hour. These snapshots are parsed and stored into the TSPOONS database in the `clusters` collection and the article links are stored into the `links` collection (see Section 3.2 for a more detailed description of the collections). Once the snapshot clusters and links are stored, TSPOONS begins its analysis work with the Chainer (Section 3.3), which connects single snapshot clusters into snapshot chains. For the resulting chains, TSPOONS extracts features of the news story chains, generates a profile for the story (Section 3.4), then uses heuristics to measure the salience of each story (Chapter 5). These profiles are stored back into the database, and retrieved by the Query Engine (Section 4.6), ranked by the importance of the story chains.

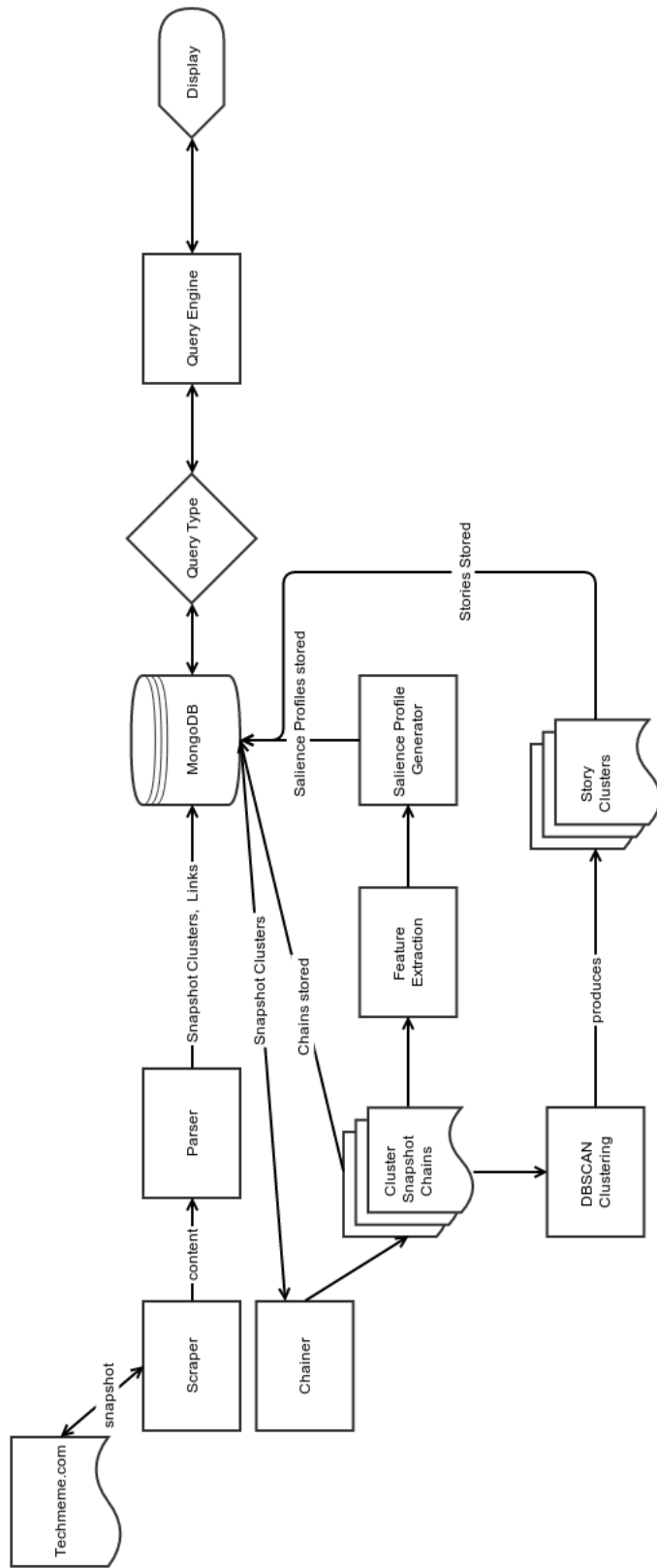


Figure 3.1: TSPOONS workflow diagram.

3.1 Parser and Scrapers

The first stage of the analysis pipeline involves gathering clusters and articles from *Techmeme.com*. Since *Techmeme.com*'s content can change minute to minute, we chose to take a snapshot of the data at a regular time interval to catalog the progression of news story cluster snapshots. Every hour, the scraper runs, gathering the content for the hour. The scraper stores the snapshot on disk with the time stamp in the following format: `tm_YY-MM-DD_HH-MM`, where `tm` stands for the source, *Techmeme.com*, followed by the year, month, and day of the snapshot, and the hour it was collected (in military time).

Once TSPOONS scrapes the page, the parser reads from the HTML dump and extracts all of the snapshot clusters on the page under the Top News section. The parser labels the clusters from 0 to $N - 1$, where N is the number of clusters on the page, resulting in the numbering scheme pictured in Figure 3.2. Links within each cluster are given a `link_id` that corresponds to their position in the cluster, starting at 0 and ending at $M - 1$, where M is the number of the links in the cluster. Additionally links are stored with a type, either `headline`, `more` or `tweet`, depending on what section of the cluster they fall under. Subclusters are listed as children of their parents in the database, and contain the `cluster_id` of their parent.

3.2 Database

TSPOONS uses MongoDB as its database back end because the data naturally fall into the document paradigm, which MongoDB is well-suited for, and because storing information about news story cluster snapshots requires a flexible schema. Not all articles have information about entities, since some are too short to use (i.e. tweets). Additionally, there is little reason to "join" collections.

Techmeme
HOME RIVER LEADERBOARD ABOUT SITE NEWS SPONSOR

Cluster 0 → **Top News**
Andy Greenberg / Forbes:
An NSA Coworker Remembers The Real Edward Snowden: 'A Genius Among Geniuses'
— Perhaps Edward Snowden's hoodie should have raised suspicions. — The black sweatshirt sold by the civil libertarian Electronic Frontier Foundation featured a parody of the National Security Agency's logo ...
More: CBS News, The Switch, The Register, The Verge, SiliconANGLE, Reuters, Business Insider, Mashable, Softpedia News, emptywheel, PC World, Pixel Envy and SecurityWeek
Tweets: @jesselynradack, @weldpond, @qthrul and @youranonnews

Cluster 1 → Spencer Ackerman / Guardian:
NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report — Our take on five things the spy agency would like the public to believe about its vast surveillance powers — The National Security Agency is telling its story like never before. Never mind whether that story is, well, true.
More: Slate and Yahoo! News. Tweets: @ggreenwald, @tonyromm, @schestowitz and @bartongellman. See also Mediagazer

Cluster 2 → Sara Morrison / The Wire:
'60 Minutes': NSA Good, Snowden Bad
More: Poynter, Errata Security, The Nation, CBS News, Gizmodo, emptywheel, Techdirt, Computerworld and Firedoglake
Tweets: @jbrodtkin, @jaycstanley, @qhardy, @csoghoian, @nickbilton, @ggreenwald, @benpopper, @ggreenwald, @tonyromm and @pourmecoffee
See also Mediagazer

Cluster 3 → Josh Gerstein / Politico:
Judge: NSA phone program likely unconstitutional — A federal judge ruled Monday that the National Security Agency program which collects information on nearly all telephone calls made to, from or within the United States is likely unconstitutional. — U.S. District Court Judge Richard Leon found ...
More: ZDNet, Guardian, Ars Technica, Techdirt, The Verge, CNET, Daring Fireball, iMore, Gigaom, VentureBeat, ACLU, Motherboard, Yahoo! News, The

Figure 3.2: Snapshot cluster ordering



Figure 3.3: Techememe.com structure

There are four main collections in the database. Each one is described below.

3.2.1 Cluster Snapshots

The `clusters` collection contains the information about each cluster snapshot with the following sub-document schema:

- `chain_links`: a list of one or more `chain_link` documents, containing the `cluster_ids` of the next snapshot cluster in the chain with the distance from the current snapshot cluster to the next snapshot cluster, the distance metric used to calculate the distance, and the inclusion and exclusion scores. The parameters specified in Figure 3.4 are discussed in detail in Section 4.3.1.

```
chain_links" : [  
  {  
    "distance" : 0.8888888888888888,  
    "exclusion" : 0.8888888888888888,  
    "inclusion" : 1,  
    "dist_type" : "jaccard",  
    "next_cluster" : "tm_14-02-23_01-00_5",  
  }  
]
```

Figure 3.4: `chain_link` document example

- `children`: a list of `unique_ids` of any cluster snapshots that are children of the current cluster snapshot. Children usually focus on different aspects of a story.
- `id`: an integer that represents the position of the cluster snapshot on the page. The first cluster snapshot has an `id` of 0, and every subsequent snapshot is counted from top to bottom and assigned an `id`.

- **key**: a string in the format "tm_YY-MM-DD_HH-MM" that represents the source ("tm" for *Techmeme.com*) and the date and time the snapshot was taken.
- **links**: a list of all the article links inside the cluster snapshot. The content of the link documents is described in Section 3.2.2.
- **parent**: an integer representing the position id of the current cluster snapshot's parent cluster, if the current cluster snapshot is a child of another cluster. Otherwise this value is null, which is true for most cluster snapshots.
- **unique_id**: a combination of the **key** and the **id** fields that creates a unique id for the cluster snapshots.

A sample record is in Appendix A.

3.2.2 Links

All of the articles within clusters are stored in the **links** collection. Each article has a link document which contains the following meta data:

- **link_url**: The URL for the article from the original website.
- **link_type**: A category, either **More**, **Tweet**, or **Headline**. A **Headline** link is the selected headlining article from the snapshot cluster that is representative of the snapshot cluster. **More** links are ones that appear as part of a list below the headline and summary on *Techmeme.com*; they represent other sources for the same story as the headline. **Tweet** links are sources for the story that come from Twitter, which may be tweets about the story or acknowledgement for bringing a story to *Techmeme.com*.
- **key**: A MD5 hash of the URL which provides a unique key for the article link.

- `link_id`: The position of the link within the snapshot cluster; headlines are 0, and all links after that increase the count.
- `entities`: A JSON dump from AlchemyAPI's Entity Extraction API outlined in Section 4.4.

An example link document is shown in Figure 3.5.

```
{
  "_id" : ObjectId("5309277081176c20f7e5d098"),
  "link_url" : "http://blogs.wsj.com/digits/2014/02/21/
               qaline-executive-on-whatsapppower-of
               -messaging/",
  "link_type" : "headline",
  "key" : "fde7be89a5657dc59429996c2e936dc4",
  "link_id" : 0
  "entities": [ ... ]
}
```

Figure 3.5: A link document example without entities (see Section 4.4 for details on entities)

3.2.3 Salience Profiles

TSPOONS stores the aggregated results of all analysis in the database in the `salience_profiles` collection. The content of salience profiles is described in Section 4.5.

3.2.4 Story Clusters

After TSPOONS produces chains, a second process runs on the data to produce topically-grouped stories. TSPOONS uses the DBSCAN algorithm to cluster chains and stores the results into the `story_clusters` collection. For each chain, the resulting `story_id`, or cluster number, is stored.

3.3 Chainer

The Chainer process performs a vital function of connecting cluster snapshots from one hour to the next hour through the process of snapshot linking (see Section 4.3.1). When the Chainer determines that two clusters are similar enough to each other, it forms a chain link between them. If no cluster is similar enough to the current cluster, the Chainer does not make a link, representing the end of a chain, and the death of a story.

The Chainer takes in as input a time range to chain cluster snapshots, considering only the clusters that fall within the range inclusively. If no starting timestamp and ending timestamp, TSPOONS defaults to the beginning of time and end of time as TSPOONS understands it.

3.4 Saliency Profile Generator

The aggregated data derived from the chain analysis are called saliency profiles. They contain metadata and the output of various analysis tasks TSPOONS performs. Before a profile is generated for each chain, TSPOONS passes the chains through a feature extraction process, which recursively iterates over the chains, keeping track of statistics described in Section 4. The saliency profile generator aggregates all information about a chain into one document, which is stored in the database, for all chains in the dataset.

3.5 Story Clustering

TSPOONS uses scikit-learn's implementation of DBSCAN [17], taking as input the chain's associated articles as HTML-stripped documents and outputting a list of

cluster numbers and document indexes. The results are stored and later retrieved by the query engine.

DBSCAN is an appropriate algorithm for clustering news stories because it matches human intuition about news stories: that stories that are topically related have high similarity (which means they are close together) and have robust boundaries between related news stories and other events. Within the dataset, there is also an unknown number of clusters, meaning algorithms that rely on a known number of clusters, such as K-means, are inappropriate for the data. Ideally, we also would like the clusters to not have simple subdivisions, which makes hierarchical clustering less appropriate because it focuses on combining similar stories until a desired threshold is reached.

3.6 Query Engine

How TSPOONS processes a query depends on the intent of the query, since TSPOONS offers 3 main types of queries: topical, range, and entity. To perform topical queries, TSPOONS relies on standard information retrieval techniques of a TF*IDF-transformed, inverted document index, and a simple index searcher that retrieves documents based on their similarity to the given query. For topical queries, TSPOONS uses the Python wrapper for the Apache open-source search engine, Lucene, called PyLucene [12]. PyLucene offers a simple interface for performing document similarity queries.

Range queries and entity queries are fulfilled by simple database look-ups, for those chains matching a time range, and those containing relevant entities, respectively. Query processing on range and entity queries is very simple since it uses the results of precomputed analytic tasks.

CHAPTER 4

Analysis Pipeline

TSPOONS's news story analysis starts by linking snapshot clusters, then analyzes the resulting chains and finds related stories. After the data has been processed, the query engine takes over to retrieve information about the news stories. This section focuses on the analysis portion of TSPOONS architecture, as shown in Figure 4.1.

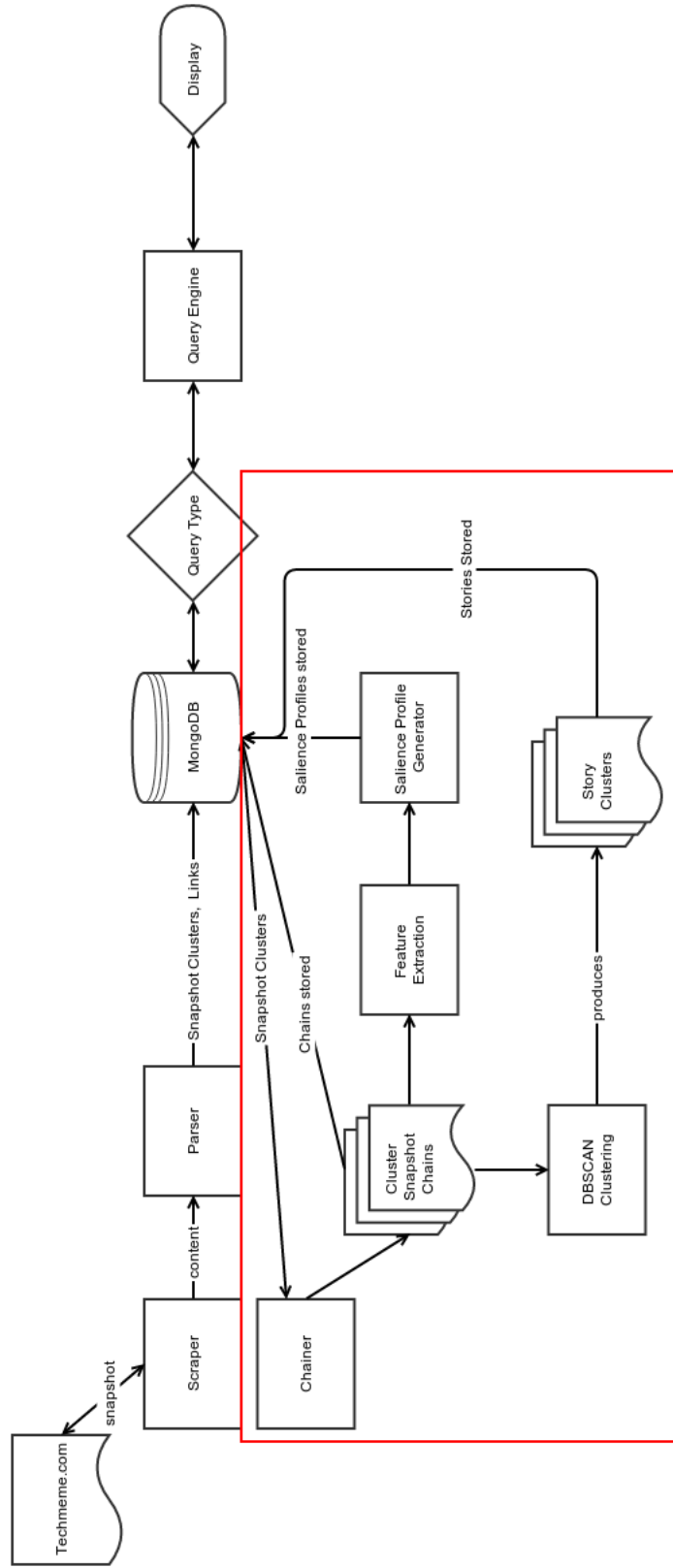


Figure 4.1: TSPOONS analysis pipeline

4.1 Dataset

TSPOONS uses a sample of the data collected, spanning a 6-month period beginning October 1, 2013 at 12:00AM and ending May 1, 2014 at 12:00 AM. TSPOONS collected 96,653 snapshot clusters (which were not necessarily unique in content), 98,911 unique articles, and produced 8,262 snapshot chains during this time.

4.2 Analysis Tools

Many of the algorithms and techniques for text mining are available to researchers through software packages and APIs that streamline the process of writing complex algorithms. To accomplish the goal of tracking and analyzing news stories, TSPOONS uses the following 3rd party APIs and software packages to assist of some of the simpler, less novel pieces of TSPOONS' analysis framework.

4.2.1 AlchemyAPI

AlchemyAPI [9] is a text mining platform that puts NLP techniques in the hands of researchers and companies seeking to do sentiment analysis, named entity extraction, keyword extraction, relation extraction, and concept tagging. The platform is backed by a "very large neural network" that is trained to make decisions and answer questions in the same way humans do [9]. AlchemyAPI pulls from several knowledge bases to perform its tasks, among them Google's Freebase knowledge base that powers the Google Knowledge Graph [10], in order to recognize entities in a text.

The platform offers a Python API that allows users to scrape content from URLs or give the API raw text for processing. TSPOONS uses the entity extraction API to identify entities within news stories and track which stories involve certain entities.

4.2.2 Gensim

Gensim [36] is a platform that provides "topic modeling for humans." Gensim has simple interfaces for performing two important text analysis tasks: TF*IDF (see Section 2.1.1) and Latent Dirichlet Allocation (LDA).

TSPOONS uses the LDA interface for performing TSPOONS' topic modeling, and although TSPOONS performs a TFIDF transformation for articles, it does not use Gensim's interface because it is much faster to handle the calculation outside of the toolkit.

4.2.3 NLTK

The Natural Language Toolkit (NLTK) is a Python toolkit that provides simple implementations of almost every type of NLP task [35]. NLTK also provides a list of stopwords and tools for fundamental NLP tasks such as tokenization, parsing, classification, and more.

TSPOONS uses several of the preprocessing NLTK libraries for tokenization, HTML-stripping, and stopword removal.

4.2.4 Scikit Learn

Scikit-learn [34] is an open source data mining library in Python that offers many different Machine Learning algorithms, including regression, clustering, and classification. David Cournapeau started `scikit-learn` as a Google Summer of Code project in 2012 and it has since gained many contributors in addition to becoming a popular tool for data science.

TSPOONS uses the Linear Regression module from `scikit-learn` in the importance

rank calculation in Section 5 and uses the DBSCAN clustering implementation for clustering the news story chains.

4.3 Linking

The first process run on the collected snapshot clusters is the Chainer (see Section 3.3). The chainer’s primary purpose is to link snapshot clusters from one hour to snapshot clusters in the next hour. In the field of TDT, link detection, or determining when two stories are topically linked, is one of the primary tasks; linking distinct documents allows for effective clustering of news articles, producing a narrative of the event or topic associated with the articles. TSPOONS links stories together in two ways: first, by detecting when two snapshot clusters represent the same news story at different times, and second, by grouping snapshot cluster chains by topic. The first step tracks a single news story over the period of consecutive hours and the second step groups snapshot cluster chains by similarity to detect chains that are generally ”about the same thing.” Initially, the task of grouping articles together is achieved by *Techmeme.com*, which aggregates articles into topical snapshot clusters, which TSPOONS collects. TSPOONS uses this as a starting point, tracking a single story over time, then finding how it fits into the grander narrative about the topics it discusses.

4.3.1 Snapshot Linking

Although snapshot clusters on *Techmeme.com* change every minute, some percentage of the cluster’s articles remains constant. As a result, we can determine when two subsequent snapshot clusters are ”the same” by measuring the overlap between the articles in each cluster. If there is sufficient overlap, then the snapshot clusters

are said to be linked. A link, or chain link, is made up of three components: distance, inclusion, and exclusion. Distance represents the overlap between clusters' articles, which TSPOONS computes using two main similarity metrics: Jaccard Index and the Sørensen-Dice Coefficient.

Jaccard Index measures the similarity between finite sample sets by taking the intersection of the sets over the union of the sets, as shown in Equation 4.1[32]. The sets, in terms of the clusters, are the articles within each cluster.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (4.1)$$

The Sørensen-Dice Coefficient also measures similarity between two sets by taking 2 times the intersection of two sets over the size of the two sets combined, as shown in Equation 4.2 [32].

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|} \quad (4.2)$$

TSPOONS compares every cluster at a given snapshot time against every cluster in the snapshot for the next hour. If the distance from one cluster at the first hour to another cluster at the second hour is greater than 0, then the two clusters are linked. For every snapshot in the dataset, all of the clusters are linked to clusters in the next hour if there exists a cluster in the next hour whose distance to the first cluster is greater than 0.

In some cases, a cluster snapshot may have a low distance to another cluster in the next hour. This is not because articles appear in multiple clusters, but because clusters may split, merge or grow over time.

However, since both Jaccard and Dice are symmetrical measures, meaning that the distance between two cluster snapshots is the same from the first hour to the

second hour as it is from the second hour to the first hour, the distance alone does not let us know whether a snapshot cluster grows or shrinks from hour to hour. In order to illustrate the change in the cluster, TSPOONS also measures two more features about the cluster snapshots from hour to our, the inclusion and exclusion. Inclusion measures the number of links in the first hour’s cluster snapshot that are included in the second hour’s cluster snapshot (see Equation 4.3). Inclusion will be low if a cluster splits into two or more clusters, since a smaller percent of the article links in the first hour’s cluster snapshot will be present in the second hour’s cluster snapshot.

$$I(A, B) = \frac{|A \cap B|}{|A|} \tag{4.3}$$

Exclusion measures the number of links that are in the second hour’s cluster snapshot which do not appear in the first. Exclusion is high if a cluster grows significantly in the next hour as new article links are added to the cluster or if two clusters merge together. The equation for exclusion is shown in Equation 4.4.

$$E(A, B) = \frac{|A \cap B|}{|B|} \tag{4.4}$$

If both inclusion, exclusion and the distance are high, then the cluster did not change very much in the last hour. Incorporating all of these metrics rationalizes why a cluster distance might be low, since we can determine what a low distance means in context of news story development.

An example chain is shown in Figure 4.2. The nodes represent the cluster snapshots with the edge labels as the Jaccard index, inclusion value, and exclusion value. The headline for this chain was "A Stream of Music, Not Revenue", which discussed how music streaming is not profitable. The story remained constant in the number

of links, with the exclusion and inclusion remaining both at 1. This indicates that the story did not change, since all the articles in the first snapshot cluster were in the second. At 06-00, the snapshot cluster splits into two snapshot clusters, with the left subchain as the parent snapshot cluster and the right subchain as a child of the left chain. The left chain continued the headlining story, while the child cluster discussed a different aspect of the story: music downloading hitting "middle age." Before the snapshot clusters merge again at time 16-00, the right subchain splits again, with some of the main articles from the right chain continuing in the subcluster, and the rest of the articles returning to the parent (or left) snapshot cluster. This chain a very simple example of the parent-child snapshot cluster relationship.

The algorithm for chaining is as follows:

<p>Algorithm 1: Chaining procedure</p> <p>Data: All snapshot clusters separated by hour collected</p> <p>Result: All Chains in the dataset</p> <p>for each c_i in all clusters in the current hour do</p> <p> for each c_j in all clusters in the next hour do</p> <p> if $\text{similarity}(c_i, c_j) > 0$ then</p> <p> inclusion = $I(c_i, c_j)$;</p> <p> exclusion = $E(c_i, c_j)$;</p> <p> store chain link with similarity, inclusion, and exclusion;</p> <p> end</p> <p> end</p> <p>end</p>
--

The similarity computation is either Jaccard or Dice for the chaining procedure and any value above 0 will result in a chain link being formed. Inclusion and exclusion are calculated using the Equations 4.3 and 4.4.

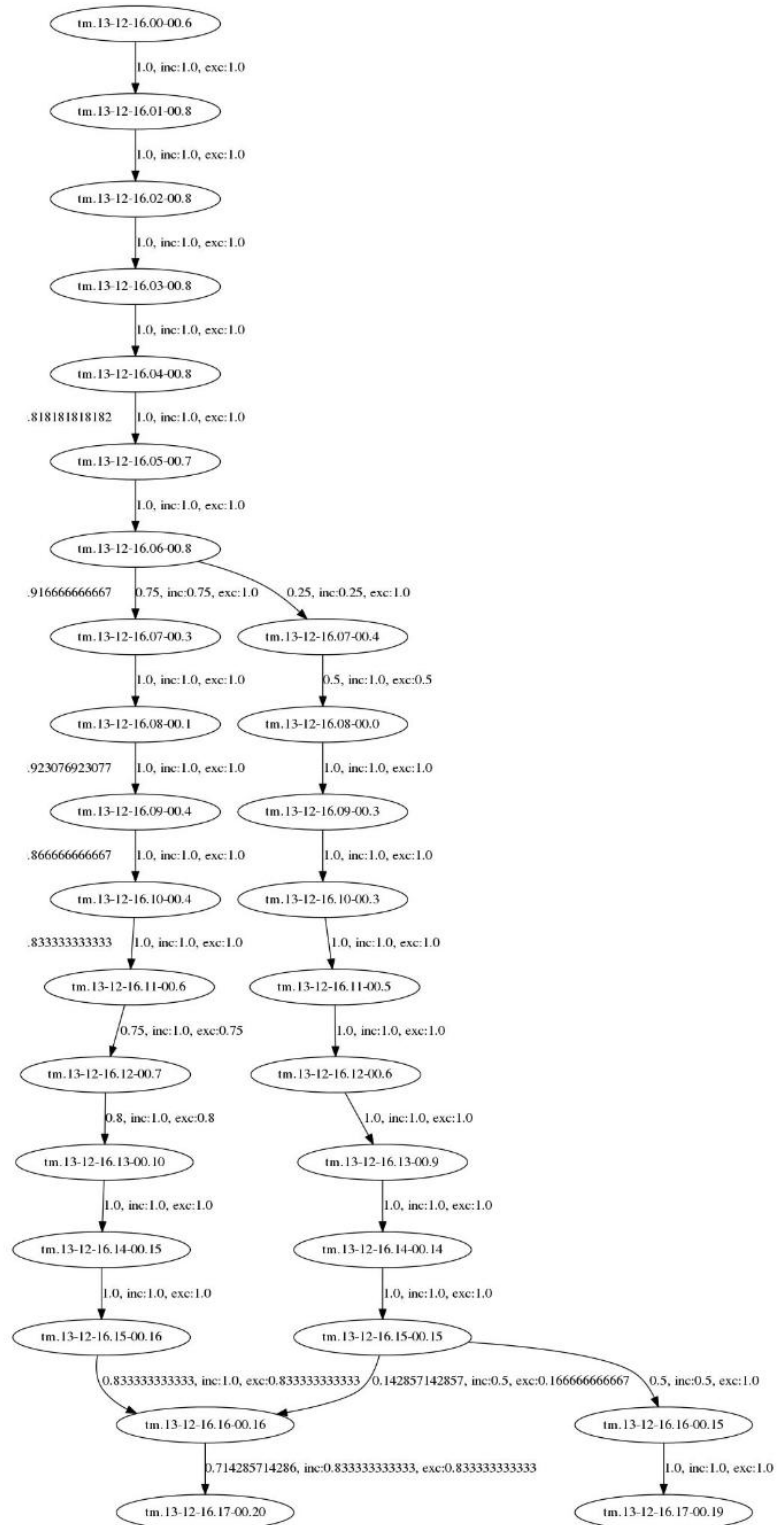


Figure 4.2: Single Chain from dataset. Snapshot clusters split and then split again before merging later.

4.3.2 Topical Linking

To track themes and find similar story chains over time, TSPOONS performs density-based clustering on the aggregated chain documents. For every unique article in the chain, TSPOONS aggregates the content of the articles into one representative "document", which is then transformed into feature vectors. TSPOONS uses scikit-learn's `TFIDFVectorizer` class [11] to transform each document into the TF*IDF Vector space with a dimension of N chains by 10,000 features. To reduce the size of the vectors and extract the most salient features, TSPOONS performs dimensionality reduction through truncated singular value decomposition (SVD) on the TF*IDF vectors, or LSA, Latent Semantic Analysis, in this context. Truncated Singular Value Decomposition [20, 38] is widely used to estimate the structure of a document by its word usage, and find the word usage patterns that are most indicative of the document's nature. The mathematics behind LSI is similar to Principle Component Analysis (PCA), which produces a matrix of latent semantic features that has smaller dimensions than the original document space.

Once the vectors have been reduced, TSPOONS performs density-based clustering using the DBSCAN algorithm (see Section 2.1.5), using cosine similarity as the distance metric. Scikit-learn includes a DBSCAN algorithm, which TSPOONS employs, taking in the parameters `epsilon` and `MinPts`. The resulting clusters are groups of similar articles which share salient features. Each cluster can be considered a theme or a group of themes relating to similar events or entities.

4.4 Entity Extraction

In order to determine which entities are present in news articles, TSPOONS uses AlchemyAPI's entity extraction interface for all articles in the dataset. AlchemyAPI

has a large knowledge base with 7.4 billion RDF¹ triples, interlinked by 142+ million RDF links, that pulls from several well-known and comprehensive knowledge bases, including:

- DBpedia [39]: a crowd-sourced effort to extract structured Wikipedia information.
- Freebase [10]: a Google-owned, but open database that contains semantic tags for entities.
- US Census [6]: a good resource for facts about people, business and geography.
- GeoNames [28]: a geographical database that contains 8 million placenames.
- UMBEL [30]: "Upper Mapping and Binding Exchange Layer," an online ontology of concepts and vocabulary.
- OpenCyc [16]: an open gateway into the world's largest and most-complete knowledge base and commonsense engine.
- YAGO [19]: a very large knowledge base that contains more than 120 million facts and more than 10 million entities.
- MusicBrainz [15]: a music encyclopedia containing music metadata that is open to the public.
- CIA Factbook [18]: information about 267 world entities, including their history, people, economy, geography, communications, military and transportation.
- CrunchBase [13]: a free database of technology companies, people, and investors.

¹RDF stands for Resource Description Framework, which is a W3 Consortium standard for metadata that is widely used in Semantic Web technologies and knowledge bases.

AlchemyAPI uses all of these sources to identify and extract entities, while tagging them with appropriate semantic information. An example result from an AlchemyAPI entity extraction call contains the following information:

```
{
  "type": "Company",
  "relevance": "0.712822",
  "sentiment": {
    "type": "positive",
    "score": "0.884934"
  },
  "count": "1",
  "text": "Twitter",
  "disambiguated": {
    "subType": [
      "Website",
      "VentureFundedCompany"
    ],
    "name": "Twitter",
    "website": "http://twitter.com/",
    "dbpedia": "http://dbpedia.org/resource/Twitter",
    "freebase": "http://rdf.freebase.com/ns/m.0289n8t",
    "crunchbase": "http://www.crunchbase.com/company/twitter"
  }
}
```

In this example, Twitter was mentioned in the article text in the sentence,

”Box is betting that a juicy NASDAQ and investors still high on **Twitter’s** offering will be receptive to its shares, even as its losses appear to be budging. Its strong revenue acceleration could be its ticket to a stable offering, however” (emphasis added)[45].

AlchemyAPI correctly identifies that Twitter is a company, disambiguating whether the phrase ”Twitter” was referring to the website or the company, ultimately choosing that the company for this context. Additionally, AlchemyAPI measures the sentiment of the company in the context of the article and it’s relevance to the article (on a scale of 0 to 1 for each). After every entry, AlchemyAPI cites its sources, so developers can trace AlchemyAPI’s decision.

AlchemyAPI's algorithms are proprietary, but the tool is useful for fast and easy entity extraction and provided a simple way of integrating multiple knowledge bases into TSPOONS.

4.5 Saliency Profiles

After the Chainer produces chains, TSPOONS performs feature extraction looking for all of the metrics used for the importance calculation as well as a few other, such as entities, the average number of articles in a cluster, both calculated importance ranks, and latent topics derived from Latent Dirichlet Allocation (LDA). TSPOONS aggregates all of these features into a saliency profile for each chain in the dataset.

Saliency profiles provide an overview of information gained through analysis in TSPOONS, and are the content returned from querying TSPOONS. The goal is to provide useful information to the user that details the features of news story chains, as shown in Table 4.1. See Section 5.1 for full descriptions of each feature.

Relevant saliency profiles are returned to the user as JSON objects, providing a simple API for users to get information out of TSPOONS.

Feature Name	Description of Feature
num_articles	Total number of unique articles in the chain
rank	Highest rank achieved by the clusters in the chain
headline	The headlining story of the chain
duration	The duration of the chain, or the number of hours it was present in the Top News section of <i>Techmeme.com</i>
num_subclusters	The average number of subclusters present in the chain
cardinality	The cardinality of the chain (total number of clusters)
chain_id	The {chain_id} of the chain
cluster_coehsiveness	The average cluster cohesiveness of the chain
avg_num_articles	The average number of articles per cluster in the chain
features	The top 10 latent features of the articles in the chain
regression_importance	The predicted importance rank from the linear regression model
entities	Any and all entities mentioned in the articles of the chain.
weighted_importance	The weighted importance rank generated from the weighted rank heuristic

Table 4.1: Features in salience profiles

4.6 Querying

TSPOONS can perform 3 main types of queries on the data: topical queries; range queries; and entity queries. Each of the query types are described below.

Topical queries answer the question, "What were the stories about _____?", where the blank can be a person (such as Larry Page), topic (such as mobile computing), or any type of information that can be extracted by a simple comparison of the query to the articles in the chains. There are two main types of topical queries that TSPOONS distinguishes against: event and thematic. Event queries are meant to return results related to a specific event, such as "Facebook buys Oculus Rift" or "Amazon Launches Fire TV." Thematic queries are meant to return stories that are generally related to the query rather than a specific event, such as "Government Surveillance" or "startup acquisitions."

Whether the topical query is an event or topical query, topical querying can work in one of two ways: first, by expanding the query using relevant Wikipedia articles based on Wikipedia's search suggest feature, which suggests relevant pages based on a query. For example, a query "Edward Snowden" returns the pages for: *PRISM (surveillance program)*; *National Security Agency*; *The Guardian*; *Glenn Greenwald*; *Global surveillance*; *Global surveillance disclosures (2013—present)*; and *Government Communications Headquarters*. TSPOONS aggregates the resulting Wikipedia pages, creating an expanded query which it then uses as a query against the chain documents. The second approach does not use query expansion and just uses the raw query from the user as input. It makes sense to expand thematic queries because we want to capture things that are generally related to the query, which Wikipedia pages often provide. Additionally, it makes sense *not* to expand event queries because there are rarely Wikipedia pages about technical events, and adding the documents to the query may muddle the results. We test these assumptions in Section 10. For now,

TSPOONS requires the user select which type of topical query they would like to perform; however, we would like to make this an automated process in the future.

TSPOONS uses PyLucene to index the chain documents, and returns the top 200 chain salience profiles to the user.

Topical queries (expanded or non-expanded) are executed using the following procedure:

Algorithm 2: Topical query execution procedure
<p>Data: Simple Query, q, and all aggregated chain documents D_c</p> <p>Result: The ranked top N documents d_n, where $d_n \in D_c$</p> <p>$q_{expanded} = \text{expansion}(q)$;</p> <p>$mostSimilarDocuments = \emptyset$;</p> <p>for $i = 0 \rightarrow D_c$ do</p> <p style="padding-left: 2em;">sim = similarity($q_{expanded}$, d_i);</p> <p style="padding-left: 2em;">$j = N$;</p> <p style="padding-left: 2em;">while $sim > mostSimilarDocuments_{j_{sim}}$ do</p> <p style="padding-left: 4em;">decrement j;</p> <p style="padding-left: 2em;">end</p> <p style="padding-left: 2em;">if $j \neq N$ or $mostSimilarDocuments == 0$ then</p> <p style="padding-left: 4em;">$mostSimilarDocuments_j = d_i$;</p> <p style="padding-left: 4em;">$mostSimilarDocuments_{j_{sim}} = sim$;</p> <p style="padding-left: 2em;">end</p> <p>end</p> <p>rank($mostSimilarDocuments$);</p>

The algorithm will find N documents with the highest cosine similarity to the query. Before returning the results, the query engine will rank the stories by importance and retrieve the salience profiles for each result.

Algorithm 3: Range query execution procedure

```
Data: A start time and end time,  $s$  and  $e$ , and all chains  $C$   
Result: The salience profiles for the chains that begin between  $s$  and  $t$   
 $D_t = \emptyset$ ;  
for  $i = 0 \rightarrow |C|$  do  
  | if  $s < c_{i,time} < e$  then  
  | | add  $c_{i,time}$  to  $D_t$ ;  
  | end  
end  
rank( $D_t$ );
```

Range queries represent a simpler form of queries, where the user specifies a time range to observe and all chains within that time are returned. Range queries answer the question, "What was happening at this time?" TSPOONS ranks the stories by importance, since the number of chains during a time period might be large. The procedure for range query execution is as follows:

Entity queries are more specific than topical queries because they use simple keyword matching to find the chains that mention the entities within the query. Entity queries are useful for answering the question, "What stories are there about _____?" TSPOONS performs an entity query by retrieving the stories that mention the entities in the query and then ranks the results by importance. The procedure for entity queries is as follows:

Algorithm 4: Range query execution procedure

Data: An entity e , and all story chains C

Result: The salience profiles for the chains that contain the entity

$D = \emptyset$;

for $i = 0 \rightarrow |C|$ **do**

if $e \in c_i$ **then**

 add c_i to D ;

end

end

rank(D);

CHAPTER 5

Measuring Saliency

The goal of measuring the importance of a news story cluster chain is to provide a model that orders chains by an intuitive definition of what humans would consider important. Importance, or saliency, has typically been measured by coverage; however, as Kiousis explains [31], coverage may not be a sufficient metric for determining saliency. Although Kiousis was unable to show that a multidimensional approach to measuring saliency was better, many people would agree that other features of a story, such as the entities involved and the cultural impact of an event, affect the saliency of a story. *There is no definitive way to measure saliency because importance is inherently subjective and how important a story is relies heavily on the context in which story appears.* As a result, TSPOONS implements two heuristics for measuring saliency that use features about a news story which we can measure.

The first heuristic tries to approximate human intuition about what stories are the most important by using human judges to rank a set of stories, then building a predictive model which we can use to estimate the importance of other stories. This approach is extensible when new stories are evaluated, but may be biased depending on the dataset and the humans. As an alternative, the second heuristic attempts to

weight the features that are most associated with importance; features like duration, number of clusters, and size of clusters all contribute to the idea of "coverage," which has been used as an importance metric in the past [31]. By weighting the features of a news story cluster chain and taking a linear combination of these features, we can produce any number of possible equations to describe importance. The second heuristic develops two of the many possible equations as a proof of concept for modeling importance in this way.

5.1 Supervised Importance Ranking

The first approach tries to approximate human judgment on importance by using supervised machine learning to assign ranks to news stories. Using a training set of manually ranked stories, we can develop a model that fits the judgement of the persons who evaluated the training set. We can also extract features from the stories that may correlate to the judges' importance ranks. If we assume the relationship between the features and the ranks given by the judges is linear, then we can use a model like multiple linear regression (see Section 2.1.4) to derive a line of best fit that approximates the ranking scheme used by the judges. TSPOONS uses *scikit-learn*'s toolkit for LassoLARS linear regression [34].

Linear regression takes in as parameters a set of ranks Y and a feature vector X that contains numerical values. The X values are derived from analyzing the articles in snapshot chains and extracting numerical values that can be measured or generated from cluster chains. The features include:

Average Size/Number of Articles: As an analog to coverage, the number of articles in a snapshot chain provides a simple, measurable feature. We take the average number of articles per cluster snapshot, since the cluster size may change

over time, and the average gives a picture of the whole chain over time.

Average Percent of Discourse: At a given hour, one snapshot chain's clusters may be dominating the content of the front page. If a story is large enough, it may have several subclusters and a large number of links associated with the story. Average percent of discourse measures the domination of a news story over its lifetime and is calculated in Equation 5.1:

$$AveragePercentDiscourse = \sum_{i=0}^n \frac{|c_i|}{N_i} \quad (5.1)$$

Here, c_i is a cluster in the snapshot chain, n is the number of clusters in the snapshot chain, and N_i is the total number of articles on the front page at the time of the snapshot i .

Story Cardinality: Cardinality is another aspect of coverage, but measures the number of clusters in a snapshot chain instead of the number of articles. Since the number of articles in a cluster snapshot varies greatly, measuring the number of clusters gives us more information about the dominance of a story.

Average Number of Subclusters: Some stories may never have subclusters, but the presence of a subcluster means that not only is the media covering the story, but they are focusing on covering more than one aspects of the story. Subclusters provide nuance to a story, and from a qualitative perspective, indicate that a story has impact since most major stories gain subclusters over time. The average number of subclusters is the count of subclusters over the number of cluster snapshots in a chain.

Average Cluster Cohesiveness: Cluster cohesiveness measures another type of coverage: the uniqueness of the article content over all of the articles in a cluster snapshot. Techmeme.com elects a single headlining article as representative of the snapshot cluster, and this article either provides the best overview of the story or is from the most important or seminal news source. The other articles in the cluster may or may not provide new or different information from the headline link, since many news sources tend to rehash content from other sources. What cluster cohesiveness tries to do is measure the level of uniqueness for each snapshot cluster, rewarding cluster snapshots for having "More" links that are dissimilar to the headline story. TSPOONS measures the cluster cohesiveness for each cluster in the snapshot chain, and takes the inverse of the average cohesiveness across the chain. Thus, if "More" links are too similar to the headline story, then the cluster's overall similarity score is inverted, and the cluster is penalized. The equation for cohesiveness is outlined in Equation 5.2.

For each snapshot cluster in a chain, and for each article in the cluster, TSPOONS cleans the data, removing stopwords from NLTK's `corpus.stopwords`, stems the words using NTLK's version of the Porter Stemmer [43] and then transforms the articles into TF*IDF vectors. TSPOONS calculates the similarity between the headline and each article in the "More" links using the cosine similarity, which produces a similarity score from 0 to 1. The results for each cluster are averaged, and the results for all clusters in the chain are averaged over the number of clusters in the chain, as shown in the following equation:

$$ClusterCohesivness = \frac{\sum_{i=0}^N \frac{\sum_{k=0}^{n_i} similarity(a_{i_k}, a_{i_h})}{n_i}}{N} \quad (5.2)$$

Here, a_{i_k} is a "More" article in a cluster snapshot i and a_{i_h} is the "Headline"

article of the cluster, n_i is the number of more links in a cluster snapshot, i and N is the number of clusters in the snapshot chain.

The intuition behind cluster cohesiveness is that although there may be a lot of articles about a story (which contribute to its importance), the articles may be re-hashings of one single news source, and may not provide more content, even though they provide more words. Cluster cohesiveness tries to measure the nuance between "more content" from "more information" in a chain.

Highest Page Position: Although the position of a snapshot cluster on the front page of *Techmeme.com* does not necessarily mean it is less or more important than the clusters below or above it, the highest positions on the page tend to contain the most important stories for the hour. Stories at the top of the page also tend to be seen more, which means that they tend to stick in public consciousness. Therefore, the rank provides a useful feature for measuring importance. It's tempting to consider the position as a metric for importance; however, the position on the page is relative, and its influence depends on the number of stories shown on the page, making the pure position on the page a poor metric. For example, a story that has the 8th position on the page may appear to be high-ranked in an hour where there are 22 clusters on the page, and low-ranked when there are only 12. To counteract this while still extracting a useful feature, the position of the cluster is normalized over the number of clusters on a page, producing a value between 0 and 1 that we can use as an informative feature. The highest page position is calculated in Equation ??.

$$MaxPagePosition = \max(p_{c_i}/n_p) \text{ for all clusters } c \text{ in chain} \quad (5.3)$$

Here p_{c_i} is the position of the cluster c at snapshot i , and n_p is the number of clusters on the page at the time of snapshot i .

Duration: Duration is another measure of coverage which looks at how long a snapshot chain was on the front page of *Techmeme.com*. Duration is the length of the chain, where each snapshot cluster counts as length of 1. Since snapshot chains split and merge, TSPOONS measures the length by the longest subchain.

These seven parameters of a snapshot chain become the X values in the linear regression computation. Once the model is trained, these features can be extracted for any snapshot chain and used to provide an importance ranking for the chains.

5.2 Weighted Feature Importance Ranking

The second heuristic for measuring salience takes a different approach to ranking stories; it approximates salience by evaluating the contributions each story makes to the discourse, applying generalized weights to the size and content of each cluster in a chain. We take a combination of three features of clusters in a chain, the size, cohesiveness, and maximum rank on the page, which are strong indicators of snapshot cluster dominance on the page, and when aggregated for every cluster in a chain, a good approximation of chain dominance during a time range.

The score is first calculated by assigning a weight to every article in the cluster; we then scale the original score by several factors to take into account other features, such as cohesiveness and duration. In the initial score calculation, headline articles receive a weight greater than the weight of the **More** articles¹ so that clusters that may only contain a headline still have weight, and each subsequent article has a weight less than the headline². Once the total is taken for the cluster, we scale the score by

¹We set headline weight to 4.

²We set the weight of **More** links to 1.

taking into account whether the cluster is larger than average or smaller than average; we scale the size by calculating the count of articles in a cluster divided by the average cluster size and then we multiply the total cluster score by the normalized value. This weight will favor clusters that are larger than average and penalize clusters that are smaller than average, thereby normalizing the cluster size.

We apply two more scaling factors to the score: the average cohesiveness of a cluster; and the highest position of the cluster on the page over the chain. Cohesiveness is measured the same as with the first approach, but the score is inverted before we add it to the computation. Inverting the similarity means that clusters that have high cohesiveness will be penalized with a lower resulting score and those with low cluster cohesiveness will receive a higher overall score; for example, a cluster with cohesiveness of 0.33 will have an inverted score of $(1-0.33)$ or 0.67, which rewards the cluster for having low cohesiveness with a larger score when it is combined with the other factors. The cluster page position is simply the normalized position of the cluster over the number of clusters on the the page at the time. The overall score for the cluster then becomes the product of the original weight of the cluster and the three features, as shown in the Equation 5.5.

$$importance = \sum_{i=0}^k \left(\left(\sum_{j=0}^n a_j \right) \cdot \frac{n_i}{n_{avg}} \cdot (1 - CC_i) \cdot \frac{p_{c_i}}{n_p} \right) \quad (5.4)$$

Here k is the number of clusters in a chain, a_j is the weight of article j in the cluster, CC_i is the *ClusterCohesiveness* of cluster i , p_{c_i} is the position of the cluster i on the page, and n_p is the total number of clusters on the page.

The second approach makes one strong assumption about cluster cohesiveness that low cohesiveness is ideal; however, low cohesiveness can mean that a cluster is ill-formed, making the content too diverse, which means it may not accurately

represent a single story but bits and pieces of many stories. Ideally, we want to find an optimal cohesiveness that captures the level of cohesiveness that makes a cluster "good" but not redundant. If we know the optimal level of cohesiveness, we can scale clusters by how close they are to achieving optimal cohesiveness, thus the rank equation becomes the following:

$$importance = \sum_{i=0}^k \left(\left(\sum_{j=0}^n a_j \right) \cdot \frac{n_i}{n_{avg}} \cdot (1 - distance(CC_i, CC_{optimal})) \cdot \frac{p_{c_i}}{n_p} \right) \quad (5.5)$$

Here the middle term now becomes the distance between the cluster's cohesiveness, CC_i , and the optimal cohesiveness, $CC_{optimal}$. The term is inverted in order to reward low distance with a higher score, and higher distance with a lower overall score. The method for generating the optimal cohesiveness is outlined in Chapter 8.

These approaches do not capture all of the possible equations we could generate to derive a numeric salience value, but they represent good heuristics for capturing human intuition and story impact, respectively. In Part 3, we evaluate the effectiveness of these approaches.

Part 3

Experiment, Evaluation, and Results

CHAPTER 6

Snapshot Cluster Chaining Evaluation

The Chainer occupies a pivotal role in the analysis pipeline, connecting snapshot clusters over time and creating the snapshot cluster chain, or story. To evaluate the Chainer, we measured the precision and recall of the Chainer, given a sample set of chains that began in a 24-hour window.

6.1 Design

We extracted chains from a 24-hour window of time, beginning December 16th, 2013 at 12 AM and ending December 16th at 11:59 PM. We then measured the precision and recall of the resulting chains, by counting the number of correct cluster links and incorrect cluster links. Precision and recall are calculated as shown in Equation 6.1 and Equation 6.2, respectively. Precision and recall measure the rate of errors, which include:

- True Positive (tp): a chain linking is correct and was outputted by the Chainer.

- False Positive (fp): a chain linking is incorrect and was outputted by the Chainer.
- False Negative (fn): a chain linking is correct but was not outputted by the Chainer.
- True Negative (tn): a chain linking is incorrect and was not outputted by the Chainer.

Precision measures how well the Chainer correctly identifies a linking; a precision of 0 means the Chainer is always wrong when making a chainlink, a precision of 0.5 means the Chainer is right half of the time when making a chainlink, and a precision of 1 means the Chainer is always correct.

$$Precision = \frac{tp}{tp + fp} \quad (6.1)$$

Recall measures how well the Chainer captures all of the links that should be made; for example, if a snapshot cluster splits and becomes two clusters, the Chainer should link the original cluster to the two cluster in the subsequent hour. If the Chainer does not make any chainlinks, essentially signalling the story is dead, then the recall would be 0. If the Chainer correctly links the original snapshot cluster to one snapshot cluster, but not the other, then the recall is 0.5. If the Chainer correctly makes both links, then the recall would be 1.

$$Recall = \frac{tp}{tp + fn} \quad (6.2)$$

6.2 Results

For one day's worth of data, there were no errors. As a result, precision and recall were both 1, and the accuracy of the Chainer is 100%.

6.3 Discussion

The Chainer is highly accurate at finding the next cluster on the page because it determines that if a snapshot cluster at time i has any overlap in articles with a snapshot cluster at time j , then there is a chain link between them. However, since the Chainer only looks ahead one hour, it does not try to detect story resurgence, so it will not detect a snapshot cluster that has some of the same articles as one from more than one hour before. Typically, it is unusual for a story to appear on the front page, disappear, and come back with the same links, however it does happen. In these cases, the topical linking should be able to find the story resurgence, since the snapshot clusters will have high similarity, and will be clustered together.

CHAPTER 7

Importance Modeling Validation

One goal of TSPOONS is to provide an analytic framework for measuring the importance or impact of news stories about a given topic or time period. Ideally, the news stories TSPOONS returns should be ranked according to a human definition of what is important. However, in order to rank stories as humans would require a quantitative approach to what is essentially a qualitative assessment. This chapter describes a quantitative method for estimating the importance of news stories using a regression model.

7.1 Design

We performed an experiment where seven human judges were asked to rank a series of news stories off of the front page of *Techmeme.com* based on their perception of which stories were more important than others. The judges were volunteers and all members of technical fields or enrolled at Cal Poly in an engineering discipline. They were asked to rank the stories on the page based on what they, as members of the technical community, felt were important to them or to the technical community in

general.

They ranked the importance of a story on a scale of 1 to N , where 1 is the most important and N was the lowest rank. The value of N was also the number of news story cluster snapshots visible on the page. *Techmeme.com* subclusters were considered to be a part of the parent cluster, and a rank was only given to a parent cluster.

We assigned the median rank given by the judges for each story as the true rank of story importance. Since the number of news story cluster snapshots changes from hour to hour, the median ranks, R_m were normalized over the number of stories on a page N to produce a normalized rank, R_N , as shown in Equation 7.1.

$$R_N = \frac{R_m}{N} \tag{7.1}$$

7.1.1 Dataset

The experiment dataset consisted of 51 unique stories drawn from December 16, 2013 at 5:00, 6:00, 8:00 PM and March 25th, 2014 at 12:00 PM and 6:00 PM. Some of the stories contained subclusters, which were included as part of the parent news story cluster’s ranking. See full dataset in Appendix A.

The story features described in Section 5 were extracted for each news story cluster and served as the features for the analysis.

7.1.2 Method

Using the story features, we built a model that estimates the human importance ranking using stepwise multiple linear regression (see Section 2.1.4) from `scikit-learn`

(see Section 4.2.4). The R_N value for each story was used as the Y value, or the dependant variable, for the regression and the feature vectors were the X values, or independent variables. Using the model, we are able to make predictions on unknown news stories based on the coefficients from the model and the features TSPOONS generates for each story.

To assess the effectiveness of the linear model used to estimate the importance ranking of a story, we need to (1) evaluate the judges' agreement, (2) choose the best aggregation method for the judges' Y rank values, and (3) assess the linear model through statistical testing. Low variability in the judges' rankings should indicate that the judges generally agree about how to rank news stories. Low variability indicates that their rankings are grouped closely together, or that the data is tightly clustered around the mean or median ranking, which makes the mean or median good measures of overall rank. We can build two models, one using the mean and the other using the median and compare the two to see which produces a better fit model for the features we extracted from the cluster chains. The last evaluation of the model is to test whether the features are statistically correlated with the ranks, and to select only the most predictive features in the linear regression model.

7.2 Results

The results for (1) evaluating the judges' agreement are outlined in Section 7.2.1, (2) choosing the best aggregation method for the judges' Y rank values are in Section 7.2.2, and (3) assessing the linear model through statistical testing are in Section 7.2.3.

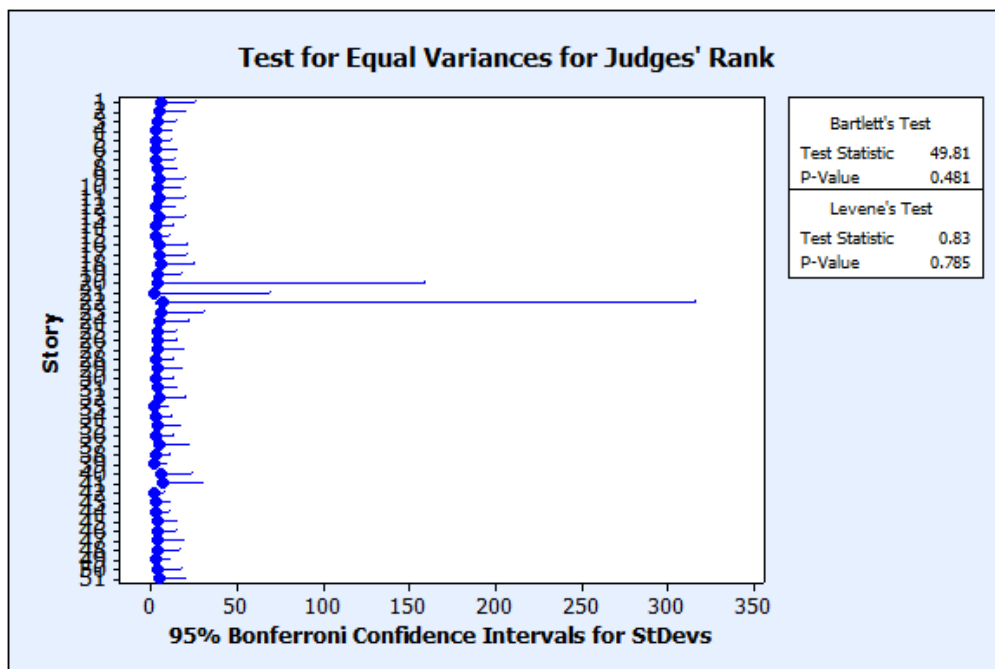


Figure 7.1: Statistical test for agreement among judges

7.2.1 Assessing Judges' Agreement

In order to test the agreement among judges, we calculated the variance of the judges' ratings for each story. If the variances are similar among all judges' ratings for each story and the variances are small compared to the number of stories on the page, then we consider the judges to be in agreement.

We performed a hypothesis test with the null hypothesis that there is no statistically significant difference in the judges' ratings for each story with a confidence level of 95%. The results of the test are shown in Figure 7.1. For both the Bartlett's and Levene's normality tests [24], the p-value was greater than 0.05; as a result, we failed to reject the null hypothesis that there is no significant difference in the values, meaning the judges ranks for each story are not significantly different with 95% confidence.

Three of the stories in the dataset had unusually high variances, this was due to

missing data in from some of the judges. However, even with the missing data, the variances are still considered statistically equal.

7.2.2 Finding the Best Y-Values

Since we have seven possible ranks for each story, we want to aggregate the judge’s responses into one single Y -value for each story. In order to determine which aggregation produced the best model, we generated models using the normalized median rank, the normalized mean rank, and the normalized 5% trimmed mean rank¹. We then calculated the R^2 values for each model. The results show that normalized median rank produced the best R^2 coefficient of determination at 41.15%, as shown in Table 7.1. As a result, the median rank was chosen as the best Y -value for the regression model among the responses that were tested.

Rank Type	R^2 Percent
Normalized Median Rank	41.15
Normalized Mean Rank	36.34
Normalized 5% Trimmed Mean Rank	36.34

Table 7.1: R^2 values for different Rank Aggregation types

7.2.3 Assessing the linear model

To determine whether or not the features we extracted from each snapshot cluster chain were good predictors of the judges’ ranks, we performed an ANOVA for regression F-Test to measure whether there was a statistically significant relationship between each independent variables he dependent rank variables.

¹The trimmed rank removes the top and bottom 5% of data values from consideration when computing the mean. It’s useful for removing noisy outliers.

Feature	p-value
Cohesiveness	0.755154
Average Number of Articles	0.382456
Average Percent of Discourse	0.299817
Number of subclusters	0.119084
Total Number of Articles	0.068430
Maximum Rank	0.018952
Duration	0.013894
Cardinality	0.006805

Table 7.2: P-value for each feature in the linear model

The null hypothesis states that for each X , there is no statistically significant relationship between the X and the Y value. We tested each X , the results of which are in Table 7.2.

Features Number of Articles, Average Percent of Discourse, Number of Subclusters, Cohesiveness, and Total Number of Articles had p-values greater than $\alpha = 0.05$, meaning they failed to reject the null hypothesis and therefore have no statistically significant relationship to the human-generated rank with 95% confidence. Since Number of Articles was still within at least a 90% confidence, we consider this feature to be sufficiently correlated. Features Maximum Rank, Duration, and Cardinality had p-values lower than $\alpha = 0.05$, meaning they have a statistically significant relationship with Y , therefore they are useful features for approximating human importance.

The leverage values of certain stories indicate whether an observation is unusual compared to the rest of the data. Leverage measures the distance between the X -values of the observation and the mean X -value for all observations. When leverage is large, the X -values are far from the mean, which indicate that these observations have large influence on the regression line. The stories which had the highest leverage

Chain Id	Story Headline	Normalized Median Rank
13-12-16_14-00_3	Judge: NSA phone program likely unconstitutional	0.1333333333
14-03-25_12-00_2	HTC announces the new One with depth-sensing camera and larger screen	0.1764705882
14-03-25_01-00_18	Mobile apps eclipse file-sharing services, digital lockers as most widely used source for pirated music	0.5294117647
14-03-24_18-00_2	Box files IPO: Big losses mask bigger ambitions	0.8235294118
13-12-16_14-00_12	Avago to Buy LSI for \$6.6 Billion	0.8666666667

Table 7.3: Stories with high leverage in the training set

for the model are shown in Table 7.3. The stories that are ranked highest and lowest among the training data appear to have the highest leverage. Without including these data points in the regression, the R^2 value drops to 0.3612 from 0.4115.

Overall, the model produced an R^2 of 0.4115, which means that 41.15% of the variability in rank is explained by the regression model with the chosen X -values.

The assumptions of our ANOVA F-test for regression are (1) that the residuals are normally distributed, (2) that the Y -values have similar variances for each X -value, (3) that the residuals are independent.

(1) The results shown in Figure 7.2, describe the distribution of the residuals and the results of an Anderson-Darling test for normality. With a p-value of 0.515 and

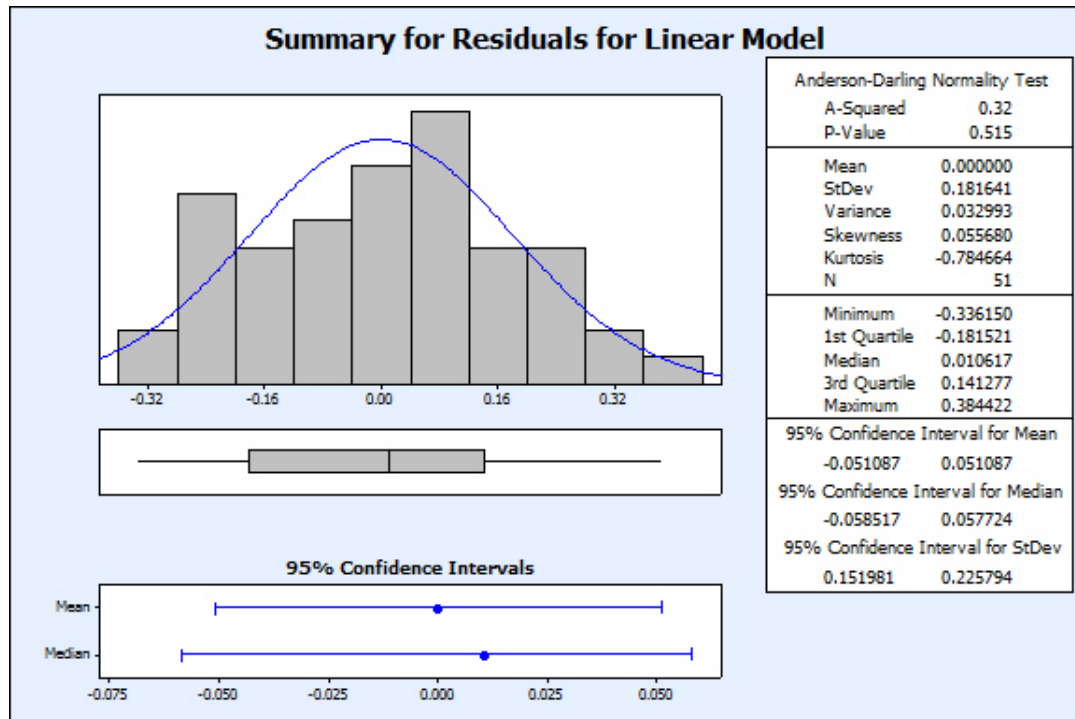


Figure 7.2: Graphical summary and Anderson-Darling test for Normality of the residuals for the linear model.

95% confidence, the test shows the distribution of the residuals is normal, validating our assumptions for the ANOVA F-test for regression.

(2) The second assumption assumes that there are equal variances among the model's residuals, which were sorted by the predicted value and binned into 5 groups. The variance of each group was measured and a hypothesis test for equal variances was performed. We only measured the variance of the residuals for the most predictive features. The results are shown in Figure 7.3. Since the p-value for both the Levine's and Bartlett's normality test were larger than 0.05, we fail to reject the null hypothesis, meaning the variances are equal.

(3) In order to evaluate the independence of the residuals, we plotted the residuals on an individual measures control chart, which tests for systematic patterns in the data. None of the rules were broken for the control chart. The results of the plot (as seen in Figure 7.4) show that there are no discernible patterns, which means the

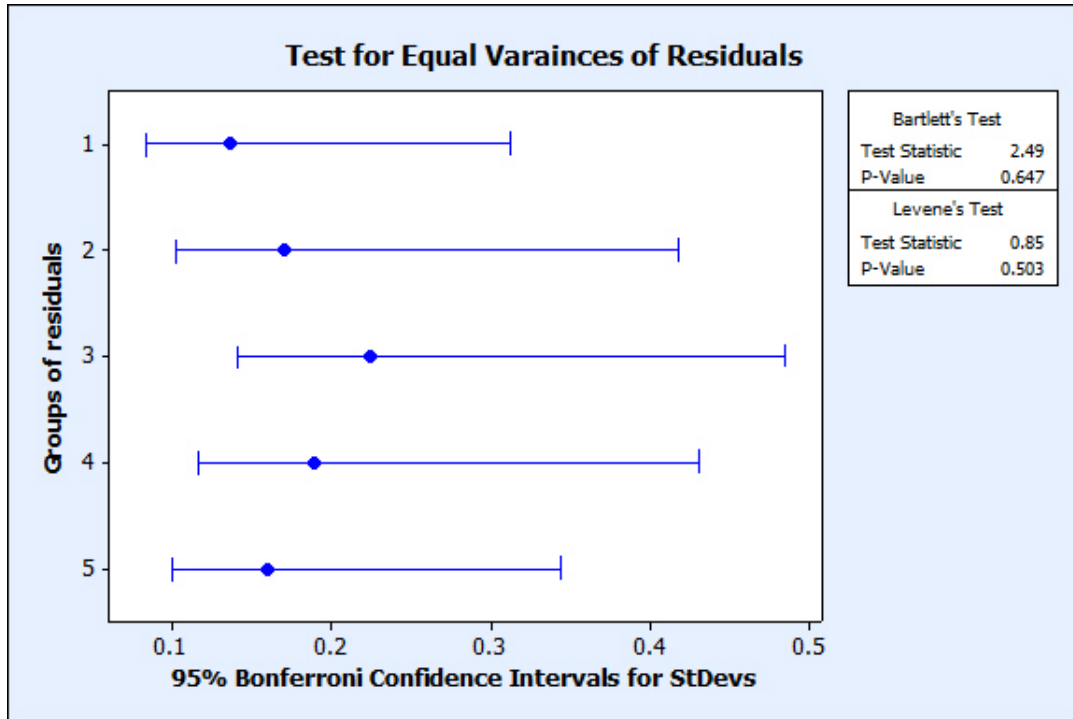


Figure 7.3: Test for equal variances among levels of the fitted values.

residuals are independent of each other.

All assumptions are validated for the model.

7.3 Qualitative Evaluative

To qualitatively assess the linear model, we rank the ranking on the entire dataset of stories, producing the top 20 results as shown in 7.4.

Headline	Linear Regression Rank
On the Matter of Why Bitcoin Matters	1.4308055329e-05
IsoHunt Resurrected Less Than Two Weeks After \$110 Million MPAA Deal — TorrentFreak	3.44561920863e-05

Target credit card data was sent to server in Russia (- Security)	3.49571358212e-05
Samsung is pulling another Amazon on Android, but this is even bigger Tech News and Analysis	8.44917571025e-05
Paramount stops releasing major movies on film - Los Angeles Times	0.000149120181244
In Googles Shadow, Facebooks Zuckerberg Pursued Oculus Over Several Months, Ending in Weekend Marathon of Dealmaking — Re/code	0.000355726874979
Media Player Winamp Shutting Down on December 20, 2013	0.000620821465603
GE experimenting with '3D painting' to repair metal parts	0.000923684678292
Tech Billionaires Spend Millions on 'Science Oscars' - Businessweek	0.000937239549966
IC ON THE RECORD Statement on Bloomberg News story that NSA knew...	0.00103476217025
Target (Yes, That Target) Wants To Launch An Accelerator In India — TechCrunch	0.00114113617863
Obamacare Website Will 'Work Smoothly' By Late November, Official Says	0.00114429155325
Google's social GPS app Waze now available on Windows Phone — The Verge	0.00117061597783
Google Chairman Eric Schmidt Posts Guide on Converting to Android from iPhone - Mac Rumors	0.00130639697418

News, opinion and aggregation on business, politics, entertainment, technology, global and national The Wire	0.00169311868044
CIA's Financial Spying Bags Data on Americans - WSJ.com	0.00172888177454
US Working Overtime Behind The Scenes To Kill UN Plan To Protect Online Privacy From Snooping — Techdirt	0.00179161996451
LinkedIn kills its Intro email service after less than four months — The Verge	0.00195341395995
Lumia 520 continues to dominate Windows Phone, while Nokia crosses 90% threshold — Windows Phone Central	0.00196713092388
Heres How Googles New Search Results Will Look Under European Antitrust Settlement — Re/code	0.00203380323543

Table 7.4: Top 20 ranked stories using Linear Regression Rank

7.4 Discussion

With an R^2 value of 0.4115, the model provides a satisfactory fit for the data. This suggests that a linear model is appropriate for approximating human opinion about importance. However, because the dataset contains high-leverage data points, more data would help regularize the model.

Interestingly, the best predictors of the judges' rankings are the features that capture the notion of volume, or coverage, of a news story (cardinality, duration, number of articles, and rank) rather than the content of a news story (cohesiveness). Intuitively, it makes sense that stories that are in the public eye (and visible for long

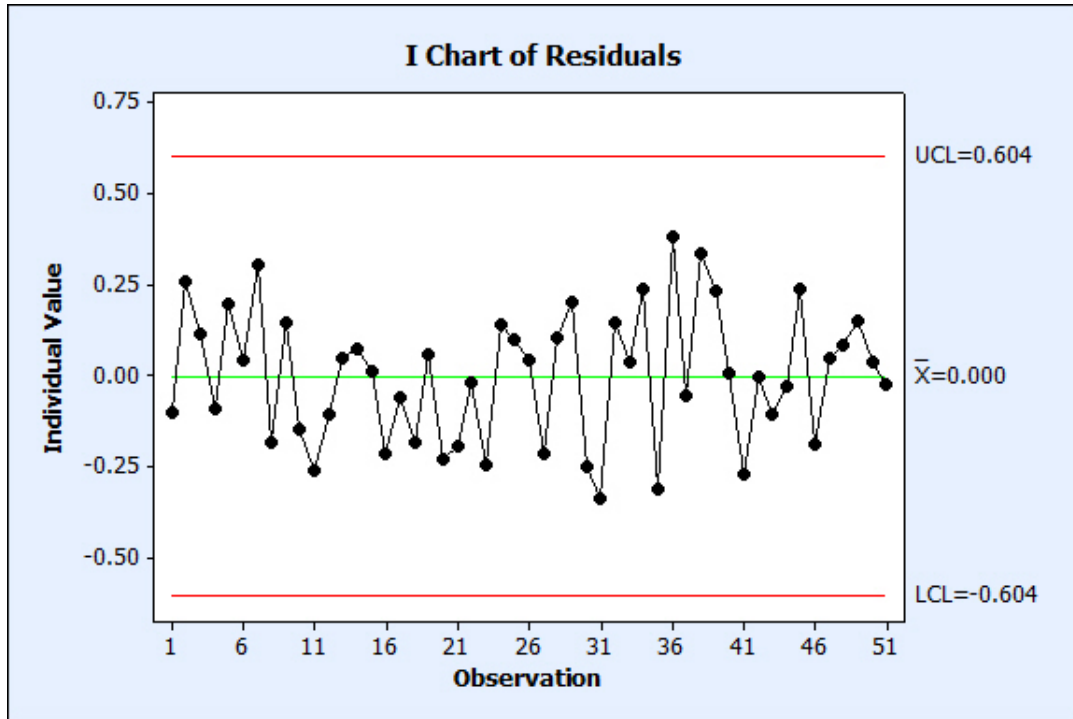


Figure 7.4: Individual Measures Chart for Residuals

periods of time) tend to be the most important. Additionally, cohesiveness had very little predictive power, suggesting that either the hypothesis that high-cohesiveness is less ideal might be false or cohesiveness is a poor predictor in general. This may be because humans do not necessarily observe cohesiveness when deciding what stories are important.

For the qualitative assessment, the linear regression rank produces intuitively-good results. The regression model captures some important stories, including the Target credit card debacle, the problems with the Obamacare website, although the results may not be considered good for all-time, but in the particular time the stories began, they were considered important, and the model reflects that. However, it's arguable that the first story is more important than the second story, so the linear model may give approximate results of what was important.

CHAPTER 8

Weighted Feature Rank Evaluation

To evaluate TSPOONS weighted feature rank equation for importance, we performed two tests measuring the predicted ranking. Although there is no golden standard for this ranking scheme, we can compare the results to the previous method, which relied on human judgment, and we can make qualitative judgments the results.

8.1 Design

We tested the results on the dataset used for the linear regression importance rank (see Section 7.1.1) and compared the results to the human judges. This provides a method for comparing how well the weighted-feature rank method approximates human intuition.

8.2 Results

The top 5 and lowest 5 results for the weighted-feature ranking of the human-ranked dataset are shown in Table 8.2 and Table 8.1, respectively. Full results are

Headline	Chain id	score
Instagram Testing Facebook Places Integration To Replace Foursquare — Fast Company — Business + Innovation	tm_14-03-25_16-00_24	0.274862379
Adobe Expands Its Marketing Cloud With Predictive Tools, iBeacon Support, And More — TechCrunch	tm_14-03-25_15-00_25	0.3719950259
Music Piracy Goes Mobile — Re/code	tm_14-03-25_11-00_16	0.9493952863
TV Check-In Company Viggie Buys Facebook Publisher Wetpaint - Peter Kafka - Media - AllThingsD	tm_13-12-16_20-00_21	1.0650102371
Hortonworks raises \$100M to scale its Hadoop business Tech News and Analysis	tm_14-03-25_17-00_25	1.1556779302

Table 8.1: Bottom 5 Weighted-feature rank for human-rated dataset.

Headline	Chain id	score
Obama To Meet Tech Execs Over NSA Spying, Obamacare Website — TIME.com	tm_13-12-16_18-00_9	466.894967984
Windows Phone 8.1 includes notification center and Siri-like personal assistant — The Verge	tm_13-12-16_12-00_10	754.913960228
Obama to Call for End to N.S.A.s Bulk Data Collection - NYTimes.com	tm_14-03-25_01-00_2	847.238420178
The new HTC One is available in Google Play and Developer editions	tm_14-03-25_12-00_2	866.281192419
Edward Snowden says judge’s ruling vindicates NSA surveillance disclosures — World news — theguardian.com	tm_13-12-16_14-00_3	2343.9207267
Box Files For 250MIPOnFull – YearRevenueOf124M, Net Loss Of \$168M	tm_14-03-24_18-00_2	2577.97083142

Table 8.2: Top 5 Weighted-feature rank for human-rated dataset.

shown in Appendix D.

8.3 Discussion

Since the weighted-feature rank is not normalized to produce a score between 0 and 1, the rankings for all stories vary from less than 1 to over 400. The stories with large ranks usually have multiple sub-clusters and long durations. The highest ranked story had a duration of 39, the second had a duration of 83 hours, and the third had a

duration of 51 while the lowest rank was on the page for one hour. As with the linear regression ranking, cohesiveness appeared to have a small effect on the overall result; instead, the size of the cluster and rank dominated the score. If we assume that media coverage is strongly correlated with importance, the weighted-feature rank provides a decent measure for coverage.

When compared to the human judges rankings, the weighted-feature rank method generally agrees with judges; 4 out of 5 of the top weighted-feature rank's stories are in the top 15 of judges rankings, and 3 out of 5 of the lowest weighted-feature ranked stories are in the bottom 15 of the judges' rankings. Although not a perfect correlation, the methods tend to agree with each other.

However, given the results of both ranking methods, it appears that cohesiveness is not a very good predictor of overall importance, despite the intuition that it may affect the value of a snapshot cluster.

CHAPTER 9

Topical Link Evaluation

Clustering the story chains using DBSCAN results in distinct clusters and noisy points. Story clusters share similar themes, thus stories in a cluster should be thematically similar to each other. To evaluate how well clusters are formed, we performed the following validation.

9.1 Design

DBSCAN takes in two main parameters: ϵ and *MinPts*. To find good clusters, we need to choose the parameters that produce the best clusters. To tune these parameters, we performed several different runs with differing ϵ and *MinPts* values. Because the true clusters are unknown, we evaluated each of the runs by judging the "goodness" of each cluster.

Cluster "goodness" was determined by looking at all of the stories in each cluster, and deciding whether the majority of stories were all thematically similar. The clusters that had the most stories that were thematically related were considered good. The parameters for the run that produced the most "good" clusters overall became

the parameters TSPOONS uses for topical linking.

9.2 Results

After trying many different values for ϵ and *MinPts*, none produced fantastic results. Most of the clusters produced by DBSCAN were large and varied; they did not stick to a theme, but instead were generally all about technical things. With the parameters $\epsilon = 0.75$ and *MinPts* = 10, DBSCAN produced 81 clusters of content and the rest of the stories were considered noise. A sample cluster from this run is shown in Table 9.1.

Exclusive: Kleiner Perkins makes major changes - The Term Sheet: Fortune's deals blogTerm Sheet
Bing Gives IE11 Users A Quick Look At The Top Search Result With New "Pre-Rendering" Feature
Germany wants a German Internet as spying scandal rankles — Reuters
Twitter revises IPO price; pegs \$23-25 per share — ZDNet
Now All Of Snapchat's Investors Are Being Sued By The Man Who Says He Was The Third, Ousted Founder
NSA Transparency Hurts Americans Privacy, Feds Say With Straight Face
Zulily Surges in Market Debut - NYTimes.com
NSA cites Reagan-era executive order to justify collection of cell-phone location data
Chinese hackers spied on Europeans before G20 meeting: researcher — Reuters
NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report — World news — theguardian.com
At the Moment, Netflix Is Just \$6.99, but Only if You're New — Adweek

Dish's 'Virtual Joey' app brings the Hopper DVR experience to LG Smart TVs — The Verge
Google releases Moto G Google Play Edition for \$179/\$199 — 9to5Google
Googles High Handed Bus Memo — TechCrunch
On the Matter of Why Bitcoin Matters
Target Data Breach Went on Longer Than Thought - WSJ.com
YouTube Reportedly Developing a Version for Kids
Cisco CTO Padmasree Warrior Joins Boxs Board — Re/code
Bitcoin Exchange Vircurex Battles Insolvency
Wanna Build a Rocket? NASAs About to Give Away a Mountain of Its Code
Technologys Man Problem - NYTimes.com

Table 9.1: Results for single cluster where $minPts$ is 10 and ϵ is 0.75.

On a subset of the data, we performed a simple test to check how well DBSCAN would perform on a small dataset with a few known clusters. Using several different parameters for ϵ and $MinPts$, we used a single day's worth of data and looked to see if stories and their sub-stories were clustered together. In none of the cases were the results meaningful.

9.3 Discussion

DBSCAN's approach to clustering fits the notion that there are major groups of stories and smaller, single instance stories, which may not be a part of a larger narrative in the discourse. However, the difficulty in using DBSCAN comes from tuning the parameters on the data: with a high $MinPts$ value, we may be eliminating

story clusters that have fewer points but are still valid clusters. At the same time, the *MinPts* value allows us to only capture the largest, most impactful clusters, leaving out stories that do not relate to major themes in the discourse.

However, the results for DBSCAN were not ideal. This could be due to TF*IDF vectors with too few features or could be a product of `scikit-learn`'s implementation. Another alternative could be that what DBSCAN is catching is general news-like language, which may not be distinguishing enough among stories, although tests with small datasets still did not produce good results, suggesting that there is an issue with either the distance calculation or the feature selection. Since `scikit-learn` controls which features are selected, there is no way for us to make it choose better features. We were unable to produce meaningful clusters with several iterations of ϵ and *MinPts* values which warrants further investigation and reimplementaion.

CHAPTER 10

Querying Evaluation

The query engine retrieves stories by measuring the similarity between the query and the content of the stories and then ranks the resulting salience profiles using either of the ranking schemes. The following sections outline the evaluation of the query engine qualitatively and quantitatively.

10.1 Design

The following sections describe the two evaluation tasks for the query engine: first, we evaluate how well the query engine retrieves a topical query and ranks the results using the two ranking methods; second, we evaluate whether or not query expansion improves the two types of topical queries, event and thematic.

10.1.1 Querying with Importance Ranking

To evaluate the query engine, we perform a sample query "Edward Snowden NSA" and retrieve the results. We first validate the relevance of the top 200 results using the precision score. Recall is not used because the relevance of every query to every

story is unknown.

The precision score (see Equation 6.1) measures the proportion of stories PyLucene returns that are relevant to the given query.

The query is expanded (see Section 4.6) and then the top 200 results are returned. We first evaluate the precision of the results, then we validate the results of the two ranking methods qualitatively.

10.1.2 Thematic vs. Event Querying

To test the effectiveness of thematic and event queries with and without query expansion, we generated 10 thematic queries (see Table 10.1) and 10 event queries (see Table 10.2) and performed querying with and without query expansion. For both of the categories, with and then without expansion, and each of the retrieved stories, we calculated whether or not the story was relevant to the query based on our judgement and the intent (thematic or event) of the query. For each query, TSPOONS retrieved 50 results, whether or not there were 50 relevant links.

Without a golden standard, or knowing how many relevant stories there were for each query, we could not compute recall for the queries. Instead, we measured precision against the true number of links found at each result level, (result 1, result 2, ... etc) to gauge how well the query engine was doing.

For each query, we found the precision of the results at every number of stories from 1 to 50. For example, if the first result was relevant, the precision is 1; if the first was relevant and the second was not, then the precision would be 0.5 or $1/2$, and so on, if we continued incrementing the number of stories until we had evaluated all 50 stories. After we had a precision score at every result count, we took the average precision for the query overall, not including the precision for any result found after

Thematic queries
Edward Snowden NSA
Lavabit
Google Smartwatch
Y-combinator start ups
Google Glass
Netflix Shows
Right to privacy
Bitcoin inventor identity
iphone rumors
Amazon drones

Table 10.1: 10 Thematic Queries used for evaluation

the last relevant result.

Once we had computed the average precision for every query, we took the mean average precision over the set of queries for each category.

10.2 Results

We evaluate the results of the query engine using a sample query and the two ranking methods for importance in Section 10.2.1. The results of the evaluation for thematic vs event queries are shown in 10.2.2.

10.2.1 Results of Querying with Importance Ranking

The top 10 results of the query without importance ranking are shown in Table 10.3. All of the top results are relevant to the expanded query, which included infor-

Event queries
Facebook buys WhatsApp
Google Stock Split
Facebook acquires Oculus Rift
Bitcoin prices drop
Apple ssl error
Microsoft develops siri google competitor cortana
Google makes Glass available for one day
Target breach credit card
Amazon launches fire tv streaming device
Obamacare Website failure

Table 10.2: 10 Event Queries used for evaluation

mation from the Wikipedia pages: *PRISM (surveillance program)*; *National Security Agency*; *The Guardian*; *Glenn Greenwald*; *Global surveillance*; *Global surveillance disclosures (2013—present)*; and *Government Communications Headquarters*. See Appendix C for the full table of results.

The top 10 results ranked with regression importance are shown in Table 10.4. For the full results, refer to Table C.3. The lower the rank, the more important a story is to TSPOONS.

The top 10 results with weighted-feature rankings are shown in Table 10.5. For the full results, refer to Table C.2. The higher the rank value, the more important TSPOONS considers the story.

chain_id	Story Headline
tm_13-10-16_22-00_12	My Next Adventure in Journalism Omidyar Group
tm_13-10-15_21-00_16	Exclusive: Greenwald exits Guardian for new Omidyar media venture — Reuters
tm_14-02-10_14-00_14	The NSA’s Secret Role in the U.S. Assassination Program - The Intercept
tm_14-04-14_22-00_22	Guardian and Washington Post win Pulitzer prize for NSA revelations — Media — The Guardian
tm_14-02-18_05-00_0	Snowden Documents Reveal Covert Surveillance and Pressure Tactics Aimed at WikiLeaks and Its Supporters - The Intercept
tm_14-01-30_02-00_20	US intelligence chief has 30 days to reveal if specific citizens were spied upon — The Verge
tm_13-12-13_18-00_14	The Mission to De-Centralize the Internet : The New Yorker
tm_13-12-16_14-00_3	Edward Snowden says judge’s ruling vindicates NSA surveillance disclosures — World news — theguardian.com
tm_14-03-12_10-00_2	How the NSA Plans to Infect ‘Millions’ of Computers with Malware - The Intercept
tm_14-01-24_02-00_18	Snowden: ‘Not all spying bad’ but NSA program ‘divorced from reason’ - CNET

Table 10.3: Top Ten Results from Regular Query (ranked by similarity)

Chain Id	Headline	Rank
tm_14-02-27_10-00_0	Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ — World news — The Guardian	-0.17538026
tm_13-12-09_01-00_2	Facebook, Google, Twitter, and more create the Reform Government Surveillance coalition — VentureBeat — Security — by Meghan Kelly	-0.27087983
tm_13-12-08_11-00_2	The Biggest Social Network No One Is Talking About: Gamers	-0.3708418
tm_13-11-27_02-00_12	Top-Secret Document Reveals NSA Spied On Porn Habits As Part Of Plan To Discredit 'Radicalizers'	0.08605344
tm_14-03-12_10-00_2	How the NSA Plans to Infect 'Millions' of Computers with Malware - The Intercept	0.08891645
tm_13-12-09_08-00_5	The Biggest Social Network No One Is Talking About: Gamers	0.10520884
tm_13-12-23_22-00_2	The National Security Agency's oversharing problem — Ars Technica	0.15520328
tm_13-12-18_14-00_1	Secret Spy Court Won't Reconsider Phone Data Collection	0.15806629
tm_14-01-02_00-00_9	Edward Snowden, Whistle-Blower - NYTimes.com	0.20830528
tm_14-02-18_05-00_0	Snowden Documents Reveal Covert Surveillance and Pressure Tactics Aimed at WikiLeaks and Its Supporters - The Intercept	0.22473922

Table 10.4: Top 10 results for "Edward Snowden NSA" query ranked by the regression model.

Chain Id	Headline	Rank
tm_13-11-07_10-00_21	U.S. weighs option to end dual leadership role at NSA, Cyber Command - The Washington Post	939.818050039
tm_13-12-10_22-00_4	How to stop spies from piggybacking on commercial Web tracking	928.131469908
tm_13-10-15_21-00_16	Exclusive: Greenwald exits Guardian for new Omidyar media venture — Reuters	83.2169162981
tm_13-12-29_09-00_3	Your USB cable, the spy: Inside the NSAs catalog of surveillance magic — Ars Technica	7706.03414741
tm_14-02-27_10-00_0	Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ — World news — The Guardian	7151.90172481
tm_14-03-31_07-00_14	The Best NSA Fix Comes From the Patriot Acts Author - The Daily Beast	7.39184964175
tm_13-11-04_09-00_3	How we know the NSA had access to internal Google and Yahoo cloud data	699.458715453
tm_14-03-10_15-00_6	Snowden says encryption and oversight are key to protecting the public from surveillance	693.88408905
tm_14-02-18_11-00_5	Spy Chief: We Shouldve Told You We Track Your Calls - The Daily Beast	61.1806744544
tm_13-12-09_01-00_2	Facebook, Google, Twitter, and more create the Reform Government Surveillance coalition — VentureBeat — Security — by Meghan Kelly	5693.5371537

Table 10.5: Top 10 results for "Edward Snowden NSA" query ranked by weighted-feature rank.

When looking at the top 200¹, the precision of the query for "Edward Snowden NSA" was 0.798, where 154/193 returned results we judged to be relevant.

10.2.2 Thematic vs. Event Query Results

The mean average precision for each combination of thematic or event and expanded or non-expanded queries are shown in Table 10.6.

	Thematic	Event
Expanded	0.4529335745	0.1723369144
Non-Expanded	0.5843538702	0.5604667873

Table 10.6: Mean Average Precision of Thematic vs Event Queries with and Without Query Expansion

The variety of the precisions for each query in the four query result categories are shown in Figures 10.1, 10.2, 10.3, 10.4.

10.3 Discussion

We discuss the results for each of the methods described in the previous sections below.

10.3.1 Importance Ranking with Querying

A precision score of 0.798 is greater than chance (0.5), we can say that the majority of the results returned by the query engine are relevant, for this query.

The results for both importance ranking methods show the variation in how each method ranks the stories. The linear regression rank placed some false positives in the

¹There were not 200 stories that PyLucene deemed relevant, so only 193 were returned.

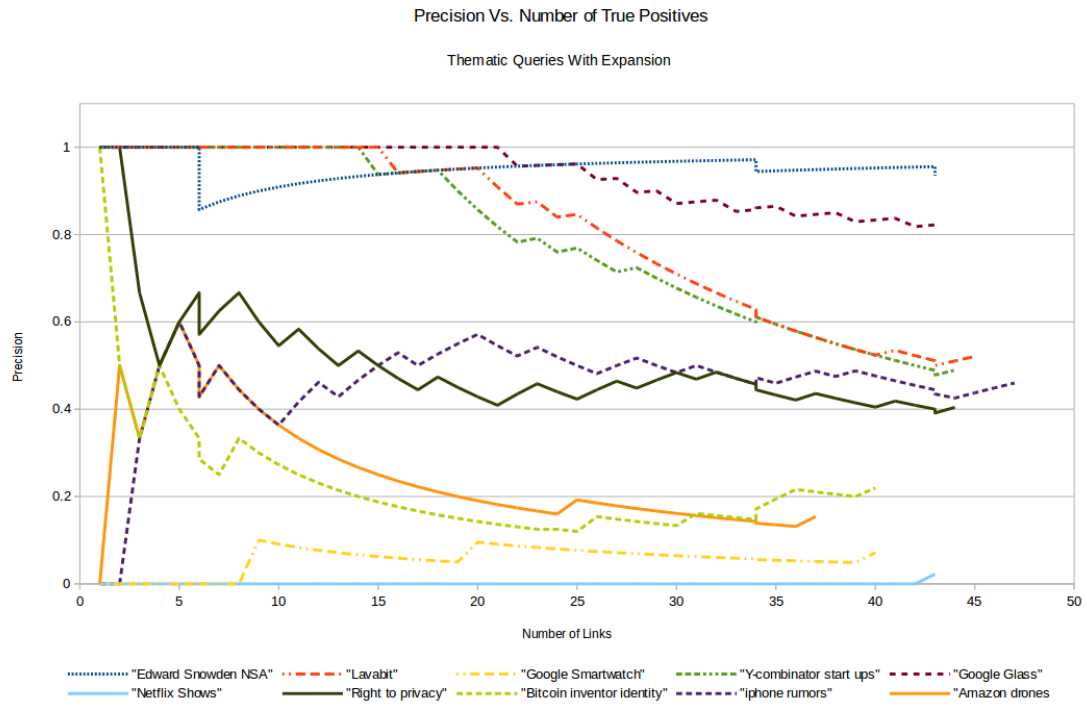


Figure 10.1: Precisions for thematic queries with expansion graphed against number of links.

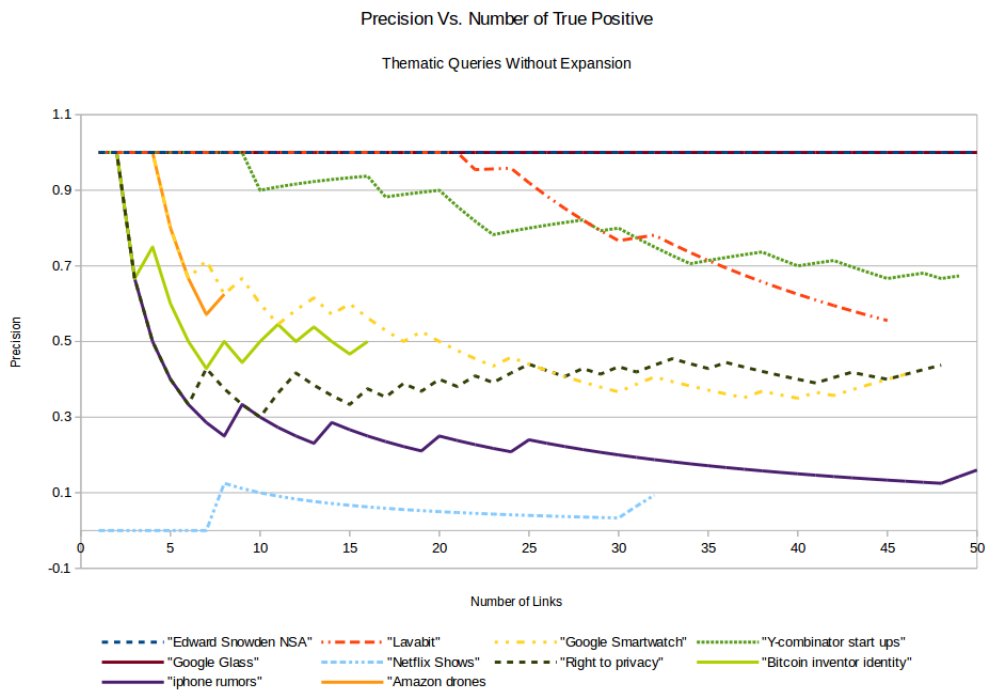


Figure 10.2: Precisions for thematic queries without expansion graphed against number of links.

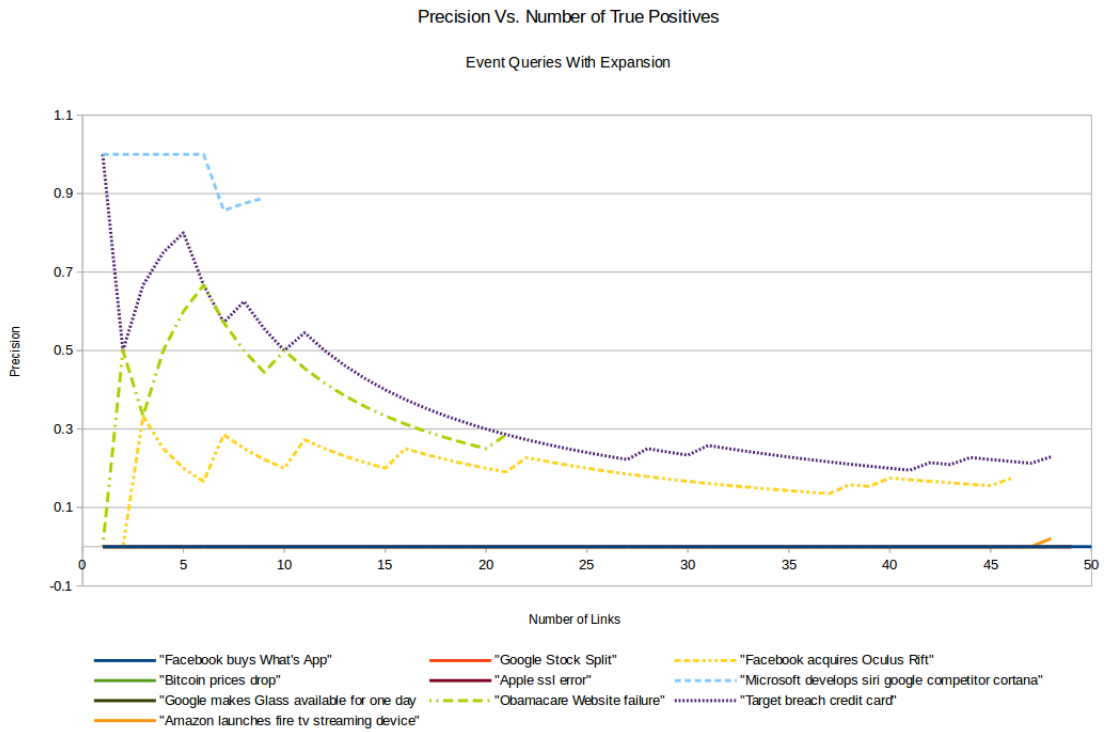


Figure 10.3: Precisions for event queries with expansion graphed against number of links.

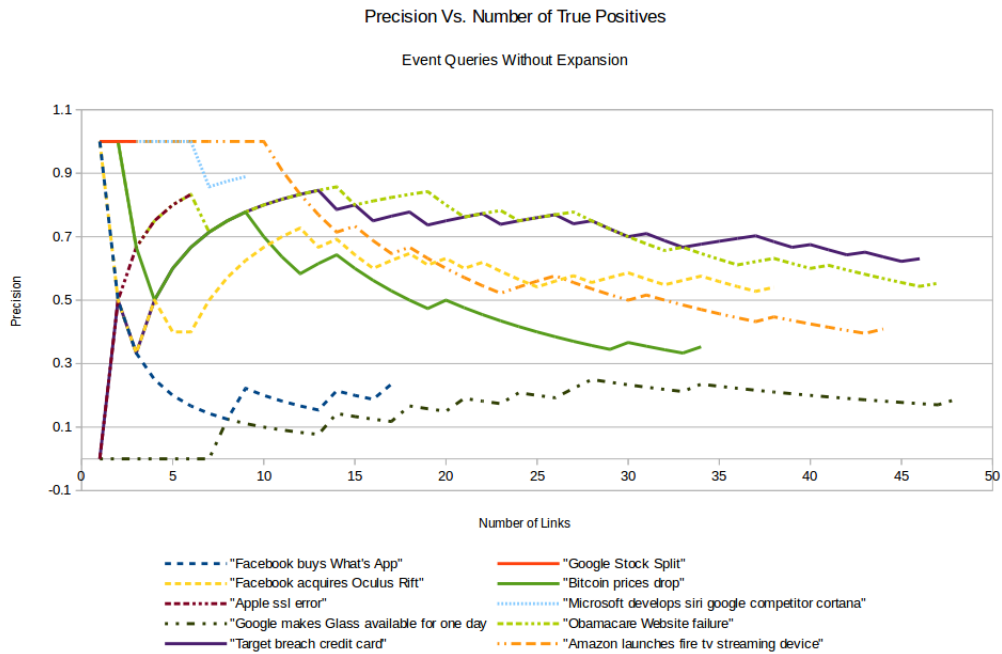


Figure 10.4: Precisions for event queries without expansion graphed against number of links.



Figure 10.5: Precision comparison of event queries with and without expansion.



Figure 10.6: Precision comparison of thematic queries with and without expansion.

results near the top, whereas the weighted-feature rank placed only relevant results in the top 10. This is likely due to linear regression being influenced by many factors, whereas the size of a cluster greatly affects the weighted-feature rank and may result in only large stories making the top spots. Generally, the two methods tend to agree on what is rated higher, but with some variation on the ordering.

10.3.2 Expanded and Non-Expanded Thematic and Event Queries

Overall, using query expansion resulted in lower precision on both event and thematic queries, although the difference between the thematic expanded and unexpanded was not large at about 0.13. The thematic queries did perform better than the event queries overall, which was likely due to the generality of thematic queries versus event queries; thematic queries are less specific and usually broader than event queries, which are looking for specific stories dealing with exact events. Without the recall values, it is difficult to tell if the queries which performed very badly, such as the "Netflix shows" query were due to lack of coverage in the dataset or poor performance of the query engine. Some queries, however, consistently performed well, including those about major stories such as the NSA surveillance and Edward Snowden. This may be due to the volume of stories available in the dataset.

Therefore, query expansion may not be a great approach to querying for stories. Since the rationale for expanding thematic queries was to improve the range of keywords PyLucene could use in order to find results, there may be no need to expand thematic queries, which means they use the same mechanisms as event queries.

Part 4

Conclusions

CHAPTER 11

Future Work & Conclusions

TSPOONS provides an initial framework for analyzing news stories through automated gathering, processing, and structuring news space.

TSPOONS accomplishes its main contributions in the following ways:

(1) *TSPOONS generates a structured view of news space by developing a model for news stories.*

The salience profiles TSPOONS generates provide the structured view of news space by first identifying stories, then breaking them down into their respective features. These profiles pare down the high volume of articles into simple reports of the content, which social scientists can use to more easily process the data.

(2) *TSPOONS generates a detailed profile of each news story, including relevant features like duration, impact, entities involved, and salience.*

TSPOONS generates salience profiles for all stories in the dataset, calculating and extracting useful information, and storing the results for later use.

(3) *TSPOONS provides a query framework for retrieving stories based on topics, stories that began within a time period, and that are about different entities.*

The query engine is able to process different types of queries and sort them by similarity to the topical query or by either of the two importance ranking algorithms.

(4) *TSPOONS provides heuristics for deciding which stories were the most important.*

TSPOONS implements two methods for calculating importance and has the potential to implement other salience calculations. Depending on the application, researchers using TSPOONS may want to approximate a human understanding of importance or they may want to weight pertinent features to discover the most covered stories.

(5) TSPOONS attempts to group stories into topical clusters for identifying themes within news space.

TSPOONS fails at its final task of topically clustering the news story cluster chains; however, there are many alternate approaches we will implement in the future to achieve the goal, which we describe later.

TSPOONS opens the door for computationally analyzing news stories, but is far from complete. Much of the work done has shown the merit of approaches, but further evaluation is needed to tune and optimize results. The evaluation in this work relied mainly on qualitative analysis and was very limited in this regard. In the future, more rigorous evaluation is needed.

For the importance ranking methods, we can perform better testing to try and improve the R^2 value for the linear regression. In the future, we'd like to use logistic regression and use pairwise comparisons (which helps to eliminate illogical judges) for generating the judge's rankings; these methodologies may provide better results and should be tested. Additionally, more comprehensive tests on querying could shed light on the variability in precision and types of queries. Overall, there are many more ways TSPOONS could be evaluated to improve performance.

Since TSPOONS is still in its early stages, we would like to make improvements to the infrastructure. TSPOONS is limited in the diversity of methodology on some steps because it relies heavily on 3rd party tool kits. For example, although the DBSCAN method from `scikit-learn` works well, it is limited by its input format, requiring smaller dimension arrays for input. Additionally, we were unable to test other clustering methods because we were limited by what was available through `scikit-learn`. In the future, we would like to expand the analysis pipeline to be more modular and less dependant on incomplete tool kits.

There are opportunities for improvement in the data collection as well; the Chainer runs on the entire dataset, rather than iteratively as the parser and scraper do. Modifying the chaining algorithm to run iteratively as data is collected would be fairly simple. TSPOONS also has the potential to pull from multiple sources, such as the sister sites to *Techmeme.com*, with the addition of other parsers. Incorporating different news domains provides more interesting comparisons of new story life-cycles and the validity of the importance ranks TSPOONS used for technical news.

The topical linking through DBSCAN clustering failed to produce good results when using `scikit-learn`'s implementation of DBSCAN. However, we still believe that DBSCAN is the right approach to clustering news stories. In an attempt to favor off-the-shelf implementations, we may have put too much trust in `scikit-learn`, which may have been built with different applications in mind. As a result, we would like to first investigate the distance metric used for ϵ because although we used cosine similarity, `scikit-learn` may have been transforming it into a distance metric, which may not have produced good results. The first change we can make is to precompute the distance matrix for DBSCAN using our own implementation of cosine similarity. Additionally, if this does not work, we want to implement our own version of DBSCAN clustering. As an alternative to DBSCAN, we can also use entity-based clustering, where we match stories based on entity overlap.

Perhaps one of the largest limitations of TSPOONS is that it has no way of visualizing the retrieved salience profiles from the database, except in the raw JSON format. Because TSPOONS outputs in JSON, the front end would only be required to take raw JSON as input. The front-end was not built because returning the raw JSON was sufficient to prove that querying worked, although there are many different kind of visualizations for the data which would benefit those using TSPOONS.

11.1 Threats to Validity

It is important to note that using *Techmeme.com* injects bias into the importance ranking results, since TSPOONS uses many *Techmeme.com*-specific features in the importance ranking calculations. Although, TSPOONS is dependent on *Techmeme.com*, the methods used to derive importance could be modified to accept more general parameters, which may be less dependent on exactly the ranking that the aggregator provides. For example, page position of the story clusters may be omitted and other features such as number of articles and number of articles being written over time could be used instead. Because of the bias, we cannot say that the ranking results are absolute or true in regards to the entire discourse; however, we can evaluate the importance rankings based on what stories are intuitively important to us, as members of the technical community.

As a result of the bias, the validity of TSPOONS is threatened by how well *Techmeme.com* captures news stories in the technical news discourse. The news aggregator may miss important stories in the technical news domain or it may not accurately represent the prominence of the news story. Additionally, the news aggregator may inaccurately promote a news story that is not as interesting or salient as *Techmeme.com* presents it to be. Another threat may be that a story is only considered important because news aggregators like *Techmeme.com* picked the story up

and gave it prominence. All of these situations contribute to bias in the importance rankings and may affect the impact of TSPOONS work. Generally, there is no way to tell exactly how well *Techmeme.com* performs in these situations, but the bias of *Techmeme.com* does not seem extreme from qualitative assessment. In the future, we would like to incorporate alternate news aggregators into the data collection pipeline to temper this bias.

11.2 Final Thoughts

Despite its limitations, TSPOONS is a meaningful first step toward building a fully-automated news story analysis engine. We anticipate that we will continue to work on TSPOONS, improving existing features and building out new aspects to enhance the news analysis engine and enable social scientists to perform extensive analysis of news space. TSPOONS lays the groundwork for computationally measuring news story salience. Its modular design allows for straight-forward extension and experimentation.

Perhaps the most significant accomplishment of TSPOONS is that it enables research in the social sciences, which are in constant need of computational tools to perform analysis on larger scale data. As TSPOONS improves, we hope that it can be used to help scientists answer big questions about how society interacts with the media and consumes news.

BIBLIOGRAPHY

- [1] About techmeme. Available at "<http://techmeme.com/about>".
- [2] Google news. Available at "<https://news.google.com>".
- [3] Mediagazer. Available at "<http://www.mediagazer.com/>".
- [4] Memeorandum: Polical web. Available at "<http://www.memeorandum.com/>".
- [5] Mongodb. Available at "<http://www.mongodb.org/>".
- [6] Uscensus. Online. Available at "<http://www.census.gov/>".
- [7] We smirch: Automatic dirt digger.
- [8] No vc: Why techmemes gabe rivera resists investors, 2012. Available at "<http://go.bloomberg.com/tech-deals/2012-12-04-no-vc-why-techmemes-gabe-rivera-resists-investors/>".
- [9] Alchemyapi deep learning research. online, 2013. Available at "<http://www.alchemyapi.com/resources/deep-learning/>".
- [10] Freebase. online, 2013. Available at "<http://www.freebase.com/>".
- [11] sklearn.feature_extraction.text.tfidfvectorizer. Online, 2013. Available at "http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html".

- [12] Apache lucene - welcome to pylucene. online, 2014. Available at "http://lucene.apache.org/pylucene/".
- [13] Crunchbase: The business graph. Online, 2014. Available at "http://www.crunchbase.com/".
- [14] How techmeme became the must-read news site for everyone in the multibillion-dollar tech industry, 2014. Available at "http://www.businessinsider.com/techmeme-growth-2014-3".
- [15] Musicbrainze. Online, 2014. Available at "http://musicbrainz.org/".
- [16] Opencyc. Online, 2014. Available at "http://www.cyc.com/platform/opencyc".
- [17] sklearn.cluster.dbSCAN. Online, 2014. Available at "http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html-sklearn.cluster.DBSCAN".
- [18] The world cia factbook. Online, 2014. Available at "https://www.cia.gov/library/publications/the-world-factbook/".
- [19] Yago. Online, 2014. Available at "http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/".
- [20] C. C. Aggarwal and C. Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.
- [21] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer international series on information retrieval. Springer US, 2002.
- [22] J. Allan, S. Harding, D. Fisher, A. Bolivar, S. Guzman-Lara, and P. Amstutz. Taking topic detection from evaluation to practice. In *System Sciences, 2005*.

- HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 101a–101a. IEEE, 2005.
- [23] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–45. ACM, 1998.
- [24] J. Devore and N. Farnum. *Applied Statistics for Engineers and Scientists*. Thomson Brooks/Cole, 2005.
- [25] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
- [26] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
- [27] A. Feng and J. Allan. Finding and linking incidents in news. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 821–830. ACM, 2007.
- [28] GeoNames. Geonames.
- [29] W. H. Hsu and S.-F. Chang. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. In *Image Processing, 2006 IEEE International Conference on*, pages 141–144. IEEE, 2006.
- [30] U. A. is Everything. Umbel.
- [31] S. Kioussis. Explicating media salience: A factor analysis of new york times issue coverage during the 2000 us presidential election. *Journal of Communication*, 54(1):71–87, 2004.

- [32] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Data-centric systems and applications. Springer, 2007.
- [33] C. C. Miller. Techmeme offers tech news at internet speed, 2010. Available at "http://www.nytimes.com/2010/07/12/technology/12techmeme.html?_r=0".
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [35] N. Project. Natural language toolkit. online, November 2013. Available at "<http://www.nltk.org/>".
- [36] R. Renhurek. Gensim topic modeling for humans. online, April 2014. Available at "<http://radimrehurek.com/gensim/>".
- [37] G. Rivera. Guess what? automated news doesn't quite work., 2008. Available at "<http://news.techmeme.com/081203/automated>".
- [38] B. Rosario. Latent semantic indexing: An overview. *Techn. rep. INFOSYS*, 240, 2000.
- [39] C. Sahnwaldt. Dbpedia - about.
- [40] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [41] A. Scharl and A. Weichselbraun. An automated approach to investigating the online media coverage of us presidential elections. *Journal of Information Technology & Politics*, 5(1):121–132, 2008.

- [42] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [43] C. van Rijsbergen, S. Robertson, and M. Porter. New models in probabilistic information retrieval. 1980.
- [44] P. Waring. Detecting and linking events.
- [45] A. Wilhelm. Box files for \$250m ipo on full-year revenue of \$124m, net loss of \$168m. Online, mar 2014. Available at "http://techcrunch.com/2014/03/24/box-files-for-250m-ipo-on-full-year-revenue-of-124m-net-loss-of-168m/".
- [46] Y. Zhai and M. Shah. Tracking news stories across different sources. In *Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA '05*, pages 2–10, New York, NY, USA, 2005. ACM.

APPENDIX A

Sample Cluster

```
{
  "_id" : ObjectId("5309806181176c2923e3af87"),
  "chain_links" : [
    {
      "distance" : 0.8888888888888888,
      "exclusion" : 0.8888888888888888,
      "inclusion" : 1,
      "dist_type" : "jaccard",
      "next_cluster" : "tm_14-02-23_01-00_5",
    }
  ],
  "children" : [ ],
  "id" : 5,
  "key" : "tm_14-02-23_00-00",
  "links" : [
    {
      "link_url" : "http://www.bloomberg.com/news/2014-02-22/
        ford-said-to-swap-blackberry-s-qnx-for
        -microsoft-in-sync-system.html",
      "_id" : ObjectId("5309806181176c2923e3af7f"),
      "link_type" : "headline",
      "key" : "c8b43c4471be98995ab5402310aa1265",
      "link_id" : 0
    },
    {
      "link_url" : "http://www.detroitnews.com/article/2014
        0222/BIZ/302220033/Ford-seen-ditching
        -Microsoft-Blackberry-future-cars",

```

```

    "_id" : ObjectId("5309806181176c2923e3af80"),
    "link_type" : "more",
    "key" : "6be38ec4b53bcf5d61f682b8c8e082b9",
    "link_id" : 1
  },
  {
    "link_url" : "http://www.winbeta.org/news/ford-drops-
      windows-next-gen-sync-systems-favor-
      -blackberry%E2%80%99s-qnx",
    "_id" : ObjectId("5309806181176c2923e3af81"),
    "link_type" : "more",
    "key" : "8c51d22311283cfa77685f2c06dd73dc",
    "link_id" : 2
  },
  {
    "link_url" : "http://crackberry.com/future-ford-sync
      -units-be-powered-blackberrys-qnx-not-
      -microsoft",
    "_id" : ObjectId("5309806181176c2923e3af82"),
    "link_type" : "more",
    "key" : "344627d9a5f74bb616da19160511f187",
    "link_id" : 3
  },
  {
    "link_url" : "http://www.berryreview.com/2014/02/22/rumor
      -ford-plans-to-use-qnx-for-their-sync-system-
      -replacing-microsoft/",
    "_id" : ObjectId("5309806181176c2923e3af83"),
    "link_type" : "more",
    "key" : "9bdf23c9d167f4572ae86b10e21c399d",
    "link_id" : 4
  },
  {
    "link_url" : "http://www.dailytech.com/article.aspx
      ?newsid=34383",
    "_id" : ObjectId("5309806181176c2923e3af84"),
    "link_type" : "more",
    "key" : "5db195ee04981e3dbf9c97584d7665a3",
    "link_id" : 5
  },
  {
    "link_url" : "http://www.wpcentral.com/ford-ditching
      -microsoft-blackberry-qnx-sync-vehicle-systems",
    "_id" : ObjectId("5309806181176c2923e3af85"),
    "link_type" : "more",

```

```
    "key" : "8e1167210ec10894c2647cf2850f1707",
    "link_id" : 6
  },
  {
    "link_url" : "http://twitter.com/qthrul/status/43739144
                  8195874816",
    "_id" : ObjectId("5309806181176c2923e3af86"),
    "link_type" : "tweet",
    "key" : "8e3ae1e8673ecf1de30d470647371a7a",
    "link_id" : 7
  }
],
"parent" : null,
"unique_id" : "tm_14-02-23_00-00_5"
}
```

Figure A.1: Example Snapshot Cluster in clusters collection

APPENDIX B

Human-Rated Story Results

Timestamp	Cluster ID	Headline	Chain Id	Median Rank	Number Stories	Normalized rank
13-12-16.17-00	0	Judge: NSA phone program likely unconstitutional	13-12-16.14-00.3	2	15	0.1333333333
13-12-16.17-01	1	An NSA Coworker Remembers The Real Edward Snowden: A Genius Among Geonituses	13-12-16.12-00.0			
13-12-16.17-02	2	NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report	13-12-16.16-00.4			
13-12-16.17-03	3	60 Minutes: NSA Good, Snowden Bad	13-12-16.12-00.1			
13-12-16.17-04	4	NSA alleges BIOS plot to destroy PCs	13-12-16.12-00.2			
13-12-16.17-05	5	Edward Snowden says judge's ruling vindicates NSA surveillance disclosures	13-12-16.17-00.5			
13-12-16.17-06	6	Windows Phone 8.1 includes notification center and Siri-like personal assistant	13-12-16.12-00.10	9	15	0.6
13-12-16.17-07	7	Photosynth creator and architect of Bing mobile and mapping leaves Microsoft to join Google	13-12-16.13-00.5	9	15	0.6
13-12-16.17-08	8	Ho Ho No! Amazon Apologizes After Customers Lose Access to Christmas Content.	13-12-16.16-00.10	5	15	0.3333333333
13-12-16.17-09	9	As Walt Mossberg exits, WSJ's new personal tech team includes Geoffrey Fowler and new hires Joanna Stern, Wilson Rothman, and Nathan Olivarez-Giles	13-12-16.15-00.12	12	15	0.8
13-12-16.17-10	10	Apple pushes iTunes Radio ad sales, builds real-time bidding exchange for in-app ads	13-12-16.17-00.10	8	15	0.5333333333
13-12-16.17-11	11	AOL's Tim Armstrong reluctantly lets go of Patch as shareholder pressures mount	13-12-16.05-00.4	12	15	0.8
13-12-16.17-12	12	Facebook Launches Donate Button For Non-Profits That Also Collects Billing Info For Itself	13-12-16.17-00.12	5	15	0.3333333333
13-12-16.17-13	13	Two More Executives Leaving BlackBerry	13-12-16.13-00.8	10	15	0.6666666667
13-12-16.17-14	14	Google will not answer to British court over UK privacy claim	13-12-16.06-00.12	5	15	0.3333333333
13-12-16.17-15	15	Amazon Workers in Germany Strike Again	13-12-16.13-00.13	5	15	0.3333333333
13-12-16.17-16	16	Despite higher ARPU generated by smartphones, carriers still complain of subsidy burden	13-12-16.09-00.12	7	15	0.4666666667
13-12-16.17-17	17	Avago to Buy LSI for \$6.6 Billion	13-12-16.14-00.12	13	15	0.8666666667
13-12-16.17-18	18	Crowdfunder Raises Another \$23 Million From Andreessen Horowitz & Others For International Expansion, Enterprise Tools	13-12-16.17-00.18	11	15	0.7333333333
13-12-16.17-19	19	The Download Hits Middle Age (and It Shows)	13-12-16.07-00.4	6	15	0.4

13-12-16.17-20	20	Music streaming services struggle to attract paying subscribers amidst increasing competition	13-12-15-20-00.12			
13-12-16.18-00	9	Obama To Meet Tech Execs Over NSA Spying, Obamacare Website	13-12-16-18-00.9	3	18	0.1666666667
13-12-16.18-01	18	Sprint launches LTE in St. Louis, San Diego; wraps up 2013 with 300 4G cities	13-12-16-18-00.18	9	18	0.5
13-12-16.18-02	19	S.F. rolls out 3 miles of free Wi-Fi along Market Street	13-12-16-18-00.19	7	18	0.3888888889
13-12-16.18-03	20	Violin Memory Fires CEO Basile as IPO and Quarterly Results Disappoint	13-12-16-18-00.20	13	18	0.7222222222
13-12-16.20-00	11	Exclusive: Twitter working on edit feature for tweets	13-12-16-20-00.11	5	21	0.2380952381
13-12-16.20-01	19	Video: Steam Machine beta unboxing, gameplay, teardown and specs	13-12-16-20-00.19	4	21	0.1904761905
13-12-16.20-02	21	TV Check-In Company Viggie Buys Facebook Publisher Wetpaint	13-12-16-20-00.21	13	21	0.619047619
14-03-25.12-00	0	Obama to Call for End to N.S.A.'s Bulk Data Collection	14-03-25-01-00.2	5.5	19	0.2894736842
14-03-25.12-01	1	Bill Would Remove Phone Database From NSA	14-03-25-01-00.3			
14-03-25.12-02	2	HTC announces the new One with depth-sensing camera and larger screen	14-03-25-12-00.2	3	17	0.1764705882
14-03-25.12-03	3	The new HTC One review	14-03-25-12-00.3			
14-03-25.12-04	4	Fitbit Partnership With HTC Gives It A Leg Up In The Fitness Tracking Space	14-03-25-12-00.4			
14-03-25.12-05	5	HTC adds key Sense apps to Play Store to make updating easier	14-03-25-12-00.5			
14-03-25.12-06	6	The new HTC One is available in Google Play and Developer editions	14-03-25-12-00.6			
14-03-25.12-07	7	Apple Testing Related Search Suggestions On The App Store	14-03-25-11-00.2	6	17	0.3529411765
14-03-25.12-08	8	Sony won't use Android Wear, will stick with Smartwatch	14-03-25-11-00.3	6	17	0.3529411765
14-03-25.12-09	9	E-book price fixing settlements land in Amazon customers' inboxes	14-03-25-11-00.11	5	17	0.2941176471
14-03-25.12-10	10	Spotify Pitches College Kids: Half Off Subscriptions While You're in School	14-03-25-11-00.4	7	17	0.4117647059
14-03-25.12-11	12	Hortonworks raises \$100M to scale its Hadoop business	14-03-25-05-00.10	12	17	0.7058823529
14-03-25.12-12	11	Survey shows people who quit Twitter want more filtering and sorting of relevant content	14-03-25-11-00.16	6	17	0.3529411765
14-03-25.12-13	13	Attackers get cash out of ATMs by sending SMS messages	14-03-25-09-00.15	4	17	0.2352941176
14-03-25.12-14	14	Box files IPO: Big losses mask bigger ambitions	14-03-24-18-00.2	14	17	0.8235294118
14-03-25.12-15	15	Hotshot CEO Aaron Levie Will Only Own 4.1% Of Box When It IPOs, Investor DFJ Owns 25.5%	14-03-24-19-00.1			
14-03-25.12-16	16	Box Files For 250M IPO On Full — Year Revenue Of 124M, Net Loss Of \$168M	14-03-25-11-00.9			

14-03-25-12-17	17	YC-Backed Gbatteries Launches BatteryBox, A 50Whr Backup Battery For Mac-Books & Other Gadgets	14-03-25-00-00_7	9	17	0.5294117647
14-03-25-12-18	18	Adobe Expands Its Marketing Cloud With Predictive Tools, iBeacon Support, And More	14-03-25-10-00_13	12	17	0.7058823529
14-03-25-12-19	19	New research may show quantum models fit the data from D-Wave computer, not classical	14-03-25-06-00_17	4	17	0.2352941176
14-03-25-12-20	20	Nokia schedules press event next week, new Windows Phone 8.1 hardware likely	14-03-25-12-00_20	15	17	0.8823529412
14-03-25-12-21	21	Mobile apps eclipse file-sharing services, digital lockers as most widely used source for pirated music	14-03-25-01-00_18	9	17	0.5294117647
14-03-25-12-22	22	Hugo Barra on Xiaomi culture, commitment to Android, and plans to enter the Indian market	14-03-25-00-00_18	15	17	0.8823529412
14-03-25-12-23	23	Nokia MixRadio becomes first global music streaming service to launch in China	14-03-25-07-00_17	15	17	0.8823529412
14-03-25-18-00	0	Facebook to Acquire Oculus VR for approximately \$2 billion	14-03-25-18-00_0	1	18	0.0555555556
14-03-25-18-01	1	Oculus will operate independently within Facebook, focusing first on gaming	14-03-25-18-00_1			
14-03-25-18-02	8	Microsoft makes source code for MS-DOS and Word for Windows available to public	14-03-25-13-00_14	3	18	0.1666666667
14-03-25-18-03	9	IRS: Bitcoin Is Property [FULL RELEASE]	14-03-25-15-00_12	4	18	0.2222222222
14-03-25-18-04	10	Bitcoin Is Property Not Currency in Tax System, IRS Says	14-03-25-15-00_11			
14-03-25-18-05	13	Google Announces Massive Price Drops For Its Cloud Computing Services And Storage, Introduces Sustained-Use Discounts	14-03-25-14-00_10	5	18	0.2777777778
14-03-25-18-06	14	Dorian Nakamoto's neighbor and second Bitcoin user Hal Finney denies helping to invent Bitcoin	14-03-25-17-00_16	9	18	0.5
14-03-25-18-07	15	Cathy Edwards, Co-Founder Of Chomp, Is Leaving Apple On April 11	14-03-25-15-00_22	15	18	0.8333333333
14-03-25-18-08	19	Nvidia unveils next-generation graphics processor with 3D memory	14-03-25-14-00_21	5	18	0.2777777778
14-03-25-18-09	20	Nvidia announce GeForce GTX Titan Z, brings 12GB VRAM for \$3,000	14-03-25-15-00_15			
14-03-25-18-10	21	Big-Data Startup RelateIQ Raises Another \$40 Million	14-03-25-16-00_24	13	18	0.7222222222
14-03-25-18-11	22	Intel Completes Purchase Of Basis Science, Which Will Join Intel's Device's Group	14-03-25-17-00_25	12	18	0.6666666667
14-03-25-18-12	23	Facebook's Open Compute guru Frank Frankovsky leaves to build optical storage startup	14-03-25-14-00_18	13	18	0.7222222222
14-03-25-18-13	24	Instagram Testing Facebook Places Integration To Replace Foursquare	14-03-25-13-00_18	11	18	0.6111111111
14-03-25-18-14	25	Waze Attacked: Technion Students Create Traffic Jam Cyber Attack On GPS App	14-03-25-15-00_23	11	18	0.6111111111

Table B.1: Training set for human judgment of importance

APPENDIX C

Query Results

chain_id	Story Headline
tm_13-10-16_22-00_12	My Next Adventure in Journalism Omidyar Group
tm_13-10-15_21-00_16	Exclusive: Greenwald exits Guardian for new Omidyar media venture — Reuters
tm_14-02-10_14-00_14	The NSA's Secret Role in the U.S. Assassination Program - The Intercept
tm_14-04-14_22-00_22	Guardian and Washington Post win Pulitzer prize for NSA revelations — Media — The Guardian
tm_14-02-18_05-00_0	Snowden Documents Reveal Covert Surveillance and Pressure Tactics Aimed at WikiLeaks and Its Supporters - The Intercept
tm_14-01-30_02-00_20	US intelligence chief has 30 days to reveal if specific citizens were spied upon — The Verge
tm_13-12-13_18-00_14	The Mission to De-Centralize the Internet : The New Yorker

tm_13-12-16_14-00_3	Edward Snowden says judge's ruling vindicates NSA surveillance disclosures — World news — theguardian.com
tm_14-03-12_10-00_2	How the NSA Plans to Infect 'Millions' of Computers with Malware - The Intercept
tm_13-12-07_21-00_2	Snowden and Greenwald: The Men Who Leaked the Secrets
tm_14-01-24_02-00_18	Snowden: 'Not all spying bad' but NSA program 'divorced from reason' - CNET
tm_13-10-27_21-00_11	News, opinion and aggregation on business, politics, entertainment, technology, global and national The Wire
tm_13-12-23_22-00_2	The National Security Agency's oversharing problem — Ars Technica
tm_14-02-18_11-00_5	Spy Chief: We Should've Told You We Track Your Calls - The Daily Beast
tm_13-12-18_14-00_1	Secret Spy Court Won't Reconsider Phone Data Collection
tm_13-10-04_12-00_19	Attacking Tor: how the NSA targets users' online anonymity — World news — theguardian.com
tm_14-04-01_06-00_17	NSA chief's legacy is shaped by big data, for better and worse - Los Angeles Times
tm_13-10-01_00-00_3	NSA stores metadata of millions of web users for up to a year, secret files show — World news — theguardian.com
tm_13-10-30_13-00_4	PRISM already gave the NSA access to tech giants. Here's why it wanted more.
tm_14-01-02_00-00_9	Edward Snowden, Whistle-Blower - NYTimes.com

tm_13-10-21_07-00_13	France in the NSA's crosshair : phone networks under surveillance
tm_14-01-17_07-00_4	Rating Obamas NSA Reform Plan: EFF Scorecard Explained — Electronic Frontier Foundation
tm_13-12-29_09-00_3	Your USB cable, the spy: Inside the NSAs catalog of surveillance magic — Ars Technica
tm_14-03-10_15-00_6	Snowden says encryption and oversight are key to protecting the public from surveillance
tm_13-12-20_11-00_8	N.S.A. Spied on Allies, Aid Groups and Businesses - NY-Times.com
tm_13-12-13_05-00_22	NSA review to leave spying programs largely unchanged, reports say — World news — The Guardian
tm_13-11-04_09-00_3	How we know the NSA had access to internal Google and Yahoo cloud data
tm_13-10-20_14-00_1	NSA Hacked Email Account of Mexican President - SPIEGEL ONLINE
tm_13-11-27_02-00_12	Top-Secret Document Reveals NSA Spied On Porn Habits As Part Of Plan To Discredit 'Radicalizers'
tm_13-12-27_12-00_9	NSA mass collection of phone data is legal, federal judge rules — World news — The Guardian
tm_13-11-08_10-00_15	Exclusive: Snowden persuaded other NSA workers to give up passwords - sources — Reuters
tm_14-03-31_07-00_14	The Best NSA Fix Comes From the Patriot Acts Author - The Daily Beast
tm_13-12-25_05-00_11	Snowden to warn Brits on Xmas telly: Your children will NEVER have privacy The Register

tm_13-12-08_11-00_2	The Biggest Social Network No One Is Talking About: Gamers
tm_14-01-17_13-00_5	President Obama's NSA reforms show both promise and peril — The Verge
tm_13-10-14_20-00_1	The NSA's problem? Too much data. - The Washington Post
tm_14-02-05_00-00_21	The Latest Snowden Revelation Is Dangerous for Anonymous And for All of Us
tm_14-02-08_19-00_2	Snowden Used Low-Cost Tool to Best N.S.A. - NYTimes.com
tm_14-02-27_10-00_0	Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ — World news — The Guardian
tm_14-01-27_16-00_1	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations
tm_13-11-07_10-00_21	U.S. weighs option to end dual leadership role at NSA, Cyber Command - The Washington Post
tm_14-01-27_13-00_7	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations
tm_13-12-10_22-00_4	How to stop spies from piggybacking on commercial Web tracking
tm_14-02-11_05-00_19	Reddit, Mozilla, Tumblr and more gear up for massive NSA protest tomorrow — VentureBeat — Security — by Harrison Weber
tm_13-12-09_08-00_5	The Biggest Social Network No One Is Talking About: Gamers

tm_13-12-09_01-00_2	Facebook, Google, Twitter, and more create the Reform Government Surveillance coalition — VentureBeat — Security — by Meghan Kelly
tm_13-10-09_20-00_14	Schneier on Security: The NSA's New Risk Analysis
tm_13-10-31_23-00_13	Angry Over U.S. Surveillance, Tech Giants Bolster Defenses - NYTimes.com
tm_14-03-11_12-00_2	Feinstein: CIA searched Intelligence Committee computers - The Washington Post
tm_13-11-21_08-00_26	US and UK struck secret deal to allow NSA to 'unmask' Britons' personal data — World news — The Guardian
tm_13-12-10_19-00_14	NSA uses Google cookies to pinpoint targets for hacking
tm_13-10-11_20-00_12	C.I.A. Warning on Snowden in 09 Said to Slip Through the Cracks - NYTimes.com
tm_14-01-23_02-00_0	Independent review board says NSA phone data program is illegal and should end - The Washington Post
tm_14-04-03_16-00_4	The Next Mission — Brendan Eich
tm_14-01-16_15-00_2	NSA collects millions of text messages daily in 'untargeted' global sweep — World news — The Guardian
tm_13-12-16_12-00_1	NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report — World news — theguardian.com
tm_14-01-07_09-00_8	How the NSA Almost Killed the Internet
tm_13-10-25_06-00_8	News, opinion and aggregation on business, politics, entertainment, technology, global and national The Wire

tm_14-03-11_23-00_15	Tim Berners-Lee: 25 years on, the Web still needs work (Q&A) - CNET
tm_13-10-24_16-00_3	Amazon.com Investor Relations: Press Release
tm_13-11-02_14-00_1	Will NSA revelations lead to the Balkanisation of the internet? — World news — The Guardian
tm_13-10-28_05-00_1	Cover Story: How NSA Spied on Merkel Cell Phone from Berlin Embassy - SPIEGEL ONLINE
tm_14-02-12_12-00_13	The Day the Internet Didn't Fight Back - NYTimes.com
tm_13-10-16_18-00_3	U.S. eavesdropping agency chief, top deputy expected to depart soon — Reuters
tm_13-12-14_14-00_4	Lawsuit accuses IBM of hiding China risks amid NSA spy scandal — Reuters
tm_14-04-08_04-00_7	Behind the Scenes: The Crazy 72 Hours Leading Up to the Heartbleed Discovery — Vocativ
tm_14-02-07_13-00_5	NSA Collects 20% or Less of U.S. Call Data - WSJ.com
tm_14-04-11_08-00_2	Obama Lets N.S.A. Exploit Some Internet Flaws, Officials Say - NYTimes.com
tm_14-03-13_16-00_3	The NSA Responds To Allegations It Impersonated Facebook And Infected PCs With Malware — TechCrunch
tm_13-11-05_06-00_2	Google's Eric Schmidt Lambasts NSA Over Spying, Following New Snowden Revelations - WSJ.com
tm_14-03-06_07-00_14	The Satoshi Paradox — Felix Salmon

tm_14-01-30_22-00_21	CSEC used airport Wi-Fi to track Canadian travellers: Edward Snowden documents - Politics - CBC News
tm_13-12-04_16-00_5	NSA tracking cellphone locations worldwide, Snowden documents show - The Washington Post
tm_14-03-24_23-00_6	Obama to Call for End to N.S.A.s Bulk Data Collection - NYTimes.com
tm_14-01-15_14-00_14	Obama to Place Some Restraints on Surveillance - NYTimes.com
tm_14-03-25_18-00_11	Obama to Call for End to N.S.A.s Bulk Data Collection - NYTimes.com
tm_13-11-25_20-00_14	Julian Assange unlikely to face U.S. charges over publishing classified documents - The Washington Post
tm_14-01-16_15-00_3	NSA collects millions of text messages daily in 'untargeted' global sweep — World news — The Guardian
tm_13-10-15_20-00_23	Meet SecureBox, the NSA-Proof Drop Box for Whistleblowers — TIME.com
tm_14-01-04_17-00_7	NSA statement does not deny 'spying' on members of Congress — World news — theguardian.com
tm_14-03-29_13-00_1	GCHQ and NSA Targeted Private German Companies - SPIEGEL ONLINE
tm_14-02-13_22-00_15	Exclusive: Snowden Swiped Password From NSA Coworker - NBC News
tm_13-12-22_12-00_5	White House Tries to Prevent Judge From Ruling on Surveillance Efforts - NYTimes.com
tm_14-01-14_21-00_18	N.S.A. Devises Radio Pathway Into Computers - NYTimes.com

tm_14-01-17_11-00_23	Congressional Reps Ask Bruce Schneier To Explain To Them What The NSA Is Doing, Because The NSA Won't Tell Them — Techdirt
tm_13-11-02_14-00_2	GCHQ and European spy agencies worked together on mass surveillance — UK news — The Guardian
tm_13-10-12_22-00_4	C.I.A. Warning on Snowden in 09 Said to Slip Through the Cracks - NYTimes.com
tm_13-12-16_17-00_5	Edward Snowden says judge's ruling vindicates NSA surveillance disclosures — World news — theguardian.com
tm_14-02-19_18-00_3	Whatsapp and \$19bn Benedict Evans
tm_14-03-25_19-00_10	Obama to Call for End to N.S.A.s Bulk Data Collection - NYTimes.com
tm_13-10-02_19-00_12	NSA chief admits agency tracked US cellphone locations in secret tests — World news — theguardian.com
tm_14-01-18_21-00_9	Obamas restrictions on NSA surveillance rely on narrow definition of spying - The Washington Post
tm_14-01-22_13-00_13	Verizon publishes first transparency report, reveals 320,000 total law enforcement requests — The Verge
tm_14-03-22_14-00_12	NSA Spied on Chinese Government and Networking Firm Huawei - SPIEGEL ONLINE
tm_13-11-01_02-00_12	Feinstein Releases Fake NSA Reform Bill, Actually Tries To Legalize Illegal NSA Bulk Data Collection — Techdirt
tm_13-12-17_14-00_4	NSA goes on 60 Minutes: the definitive facts behind CBS's flawed report — World news — theguardian.com
tm_13-10-24_01-00_24	Merkel Calls Obama Over Suspicions US Tapped Her Mobile Phone - SPIEGEL ONLINE

tm_13-12-09-01-00_1	Reform Government Surveillance
tm_13-11-23-07-00_5	NSA infected 50,000 computer networks with malicious software - nrc.nl
tm_13-10-02-12-00_7	Silk Roads mastermind allegedly paid \$80,000 for a hitman. The hitman was a cop.

Table C.1: Results of the "Edward Snowden NSA" query

Chain Id	Headline	Rank
tm_13-11-07-10-00-21	U.S. weighs option to end dual leadership role at NSA, Cyber Command - The Washington Post	939.818050039
tm_13-12-10-22-00-4	How to stop spies from piggybacking on commercial Web tracking	928.131469908
tm_13-10-15-21-00-16	Exclusive: Greenwald exits Guardian for new Omidyar media venture — Reuters	83.2169162981
tm_13-12-29-09-00-3	Your USB cable, the spy: Inside the NSAs catalog of surveillance magic — Ars Technica	7706.03414741
tm_14-02-27-10-00-0	Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ — World news — The Guardian	7151.90172481
tm_14-03-31-07-00-14	The Best NSA Fix Comes From the Patriot Acts Author - The Daily Beast	7.39184964175
tm_13-11-04-09-00-3	How we know the NSA had access to internal Google and Yahoo cloud data	699.458715453
tm_14-03-10-15-00-6	Snowden says encryption and oversight are key to protecting the public from surveillance	693.88408905

tm_14-02-18_11-00_5	Spy Chief: We Shouldve Told You We Track Your Calls - The Daily Beast	61.1806744544
tm_13-12-09_01-00_2	Facebook, Google, Twitter, and more create the Reform Government Surveillance coalition — VentureBeat — Security — by Meghan Kelly	5693.5371537
tm_14-02-18_05-00_0	Snowden Documents Reveal Covert Surveillance and Pressure Tactics Aimed at WikiLeaks and Its Supporters - The Intercept	567.45701367
tm_14-01-17_13-00_5	President Obama’s NSA reforms show both promise and peril — The Verge	553.397466319
tm_13-10-30_13-00_4	PRISM already gave the NSA access to tech giants. Heres why it wanted more.	5131.46505906
tm_13-12-27_12-00_9	NSA mass collection of phone data is legal, federal judge rules — World news — The Guardian	473.411004444
tm_13-12-08_11-00_2	The Biggest Social Network No One Is Talking About: Gamers	4653.14961949
tm_13-10-27_21-00_11	News, opinion and aggregation on business, politics, entertainment, technology, global and national The Wire	426.83302139
tm_13-10-20_14-00_1	NSA Hacked Email Account of Mexican President - SPIEGEL ONLINE	424.293695819
tm_14-01-24_02-00_18	Snowden: 'Not all spying bad' but NSA program 'divorced from reason' - CNET	41.4096821636
tm_13-10-01_00-00_3	NSA stores metadata of millions of web users for up to a year, secret files show — World news — theguardian.com	392.273645405

tm_13-12-09-08-00_5	The Biggest Social Network No One Is Talking About: Gamers	3675.46295766
tm_13-10-11-20-00_12	C.I.A. Warning on Snowden in 09 Said to Slip Through the Cracks - NYTimes.com	341.517266147
tm_13-10-14-20-00_1	The NSA's problem? Too much data. - The Washington Post	3289.52728933
tm_13-12-13-05-00_22	NSA review to leave spying programs largely unchanged, reports say — World news — The Guardian	317.368144538
tm_13-12-25-05-00_11	Snowden to warn Brits on Xmas telly: Your children will NEVER have privacy The Register	31.8582984436
tm_14-03-12-10-00_2	How the NSA Plans to Infect 'Millions' of Computers with Malware - The Intercept	3088.83088826
tm_13-11-08-10-00_15	Exclusive: Snowden persuaded other NSA workers to give up passwords - sources — Reuters	301.777058988
tm_13-10-04-12-00_19	Attacking Tor: how the NSA targets users' online anonymity — World news — theguardian.com	2919.2721451
tm_13-12-18-14-00_1	Secret Spy Court Wont Reconsider Phone Data Collection	2883.75203127
tm_13-12-10-19-00_14	NSA uses Google cookies to pinpoint targets for hacking	287.549139431
tm_14-02-11-05-00_19	Reddit, Mozilla, Tumblr and more gear up for massive NSA protest tomorrow — VentureBeat — Security — by Harrison Weber	283.26446682
tm_14-01-30-02-00_20	US intelligence chief has 30 days to reveal if specific citizens were spied upon — The Verge	28.2926513903

tm_14-03-11_12-00_2	Feinstein: CIA searched Intelligence Committee computers - The Washington Post	274.517538086
tm_13-12-23_22-00_2	The National Security Agency's oversharing problem — Ars Technica	2690.05736537
tm_13-12-07_21-00_2	Snowden and Greenwald: The Men Who Leaked the SecretsRead	265.85401666
tm_14-01-27_13-00_7	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations	2439.71517391
tm_13-11-27_02-00_12	Top-Secret Document Reveals NSA Spied On Porn Habits As Part Of Plan To Discredit 'Radicalizers'	2429.38472182
tm_13-12-16_14-00_3	Edward Snowden says judge's ruling vindicates NSA surveillance disclosures — World news — theguardian.com	2343.9207267
tm_13-10-16_22-00_12	My Next Adventure in Journalism	225.298147097
tm_14-01-17_07-00_4	Rating Obamas NSA Reform Plan: EFF Scorecard Explained — Electronic Frontier Foundation	1796.14035228
tm_13-10-21_07-00_13	France in the NSA's crosshair : phone networks under surveillance	179.767808655
tm_13-12-13_18-00_14	The Mission to De-Centralize the Internet : The New Yorker	1755.25017006
tm_14-02-10_14-00_14	The NSA's Secret Role in the U.S. Assassination Program - The Intercept	15.3342199721

tm_14-04-01_06-00_17	NSA chief's legacy is shaped by big data, for better and worse-Los Angeles Times	13.2876303236
tm_14-01-27_16-00_1	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations	1267.76079122
tm_14-02-08_19-00_2	Snowden Used Low-Cost Tool to Best N.S.A. - NYTimes.com	1253.05108021
tm_13-12-20_11-00_8	N.S.A. Spied on Allies, Aid Groups and Businesses - NYTimes.com	1242.83532001
tm_14-02-05_00-00_21	The Latest Snowden Revelation Is Dangerous for Anonymous And for All of Us	1151.05047109
tm_14-01-02_00-00_9	Edward Snowden, Whistle-Blower - NYTimes.com	1064.15256872
tm_13-10-09_20-00_14	Schneier on Security: The NSA's New Risk Analysis	1.94858884138
tm_14-04-14_22-00_22	Guardian and Washington Post win Pulitzer prize for NSA revelations — Media — The Guardian	0.487813005814

Table C.2: Top 50 Ranked query results for "Edward Snowden NSA" query ranked by weighted-feature rank.

Chain Id	Headline	Rank
tm_14-02-27_10-00_0	Optic Nerve: millions of Yahoo webcam images intercepted by GCHQ — World news — The Guardian	-0.17538026

tm_13-12-09_01-00_2	Facebook, Google, Twitter, and more create the Reform Government Surveillance coalition — VentureBeat — Security — by Meghan Kelly	-0.27087983
tm_13-12-08_11-00_2	The Biggest Social Network No One Is Talking About: Gamers	-0.3708418
tm_13-11-27_02-00_12	Top-Secret Document Reveals NSA Spied On Porn Habits As Part Of Plan To Discredit 'Radicalizers'	0.08605344
tm_14-03-12_10-00_2	How the NSA Plans to Infect 'Millions' of Computers with Malware - The Intercept	0.08891645
tm_13-12-09_08-00_5	The Biggest Social Network No One Is Talking About: Gamers	0.10520884
tm_13-12-23_22-00_2	The National Security Agency's oversharing problem — Ars Technica	0.15520328
tm_13-12-18_14-00_1	Secret Spy Court Won't Reconsider Phone Data Collection	0.15806629
tm_14-01-02_00-00_9	Edward Snowden, Whistle-Blower - NYTimes.com	0.20830528
tm_14-02-18_05-00_0	Snowden Documents Reveal Covert Surveillance and Pressure Tactics Aimed at WikiLeaks and Its Supporters - The Intercept	0.22473922
tm_13-12-10_22-00_4	How to stop spies from piggybacking on commercial Web tracking	0.22622397
tm_13-12-20_11-00_8	N.S.A. Spied on Allies, Aid Groups and Businesses - NYTimes.com	0.25548821

tm_14-01-27_13-00_7	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations	0.31193068
tm_13-12-16_14-00_3	Edward Snowden says judge's ruling vindicates NSA surveillance disclosures — World news — theguardian.com	0.316159
tm_13-10-20_14-00_1	NSA Hacked Email Account of Mexican President - SPIEGEL ONLINE	0.31935312
tm_13-10-11_20-00_12	C.I.A. Warning on Snowden in 09 Said to Slip Through the Cracks - NYTimes.com	0.39467007
tm_14-02-08_19-00_2	Snowden Used Low-Cost Tool to Best N.S.A. - NYTimes.com	0.45977518
tm_13-12-27_12-00_9	NSA mass collection of phone data is legal, federal judge rules — World news — The Guardian	0.59517127
tm_14-03-10_15-00_6	Snowden says encryption and oversight are key to protecting the public from surveillance	0.61757927
tm_13-10-14_20-00_1	The NSA's problem? Too much data. - The Washington Post	0.64768917
tm_13-11-04_09-00_3	How we know the NSA had access to internal Google and Yahoo cloud data	0.65923126
tm_14-02-05_00-00_21	The Latest Snowden Revelation Is Dangerous for Anonymous And for All of Us	0.66638704
tm_13-12-10_19-00_14	NSA uses Google cookies to pinpoint targets for hacking	0.69625531
tm_13-10-27_21-00_11	News, opinion and aggregation on business, politics, entertainment, technology, global and national The Wire	0.72251499

tm_14-01-17_13-00_5	President Obama's NSA reforms show both promise and peril — The Verge	0.77344311
tm_14-01-27_16-00_1	Snowden docs reveal British spies snooped on YouTube and Facebook - Investigations	0.78103998
tm_14-03-11_12-00_2	Feinstein: CIA searched Intelligence Committee computers - The Washington Post	1.20251013
tm_14-01-17_07-00_4	Rating Obamas NSA Reform Plan: EFF Scorecard Explained — Electronic Frontier Foundation	1.28019109
tm_13-12-07_21-00_2	Snowden and Greenwald: The Men Who Leaked the SecretsRead	1.39414049
tm_13-10-01_00-00_3	NSA stores metadata of millions of web users for up to a year, secret files show — World news — theguardian.com	1.60655418
tm_13-11-08_10-00_15	Exclusive: Snowden persuaded other NSA workers to give up passwords - sources — Reuters	1.70427913
tm_13-10-16_22-00_12	My Next Adventure in Journalism	1.72668713
tm_13-12-29_09-00_3	Your USB cable, the spy: Inside the NSAs catalog of surveillance magic — Ars Technica	1.92734746
tm_13-10-30_13-00_4	PRISM already gave the NSA access to tech giants. Heres why it wanted more.	1.95441091
tm_14-02-11_05-00_19	Reddit, Mozilla, Tumblr and more gear up for massive NSA protest tomorrow — VentureBeat — Security — by Harrison Weber	2.04732128

tm_13-10- 21-07-00_13	France in the NSA's crosshair : phone networks under surveillance	2.13907217
tm_13-11- 07-10-00_21	U.S. weighs option to end dual leadership role at NSA, Cyber Command - The Washington Post	2.24568569
tm_14-02- 18-11-00_5	Spy Chief: We Shouldve Told You We Track Your Calls - The Daily Beast	2.29232104
tm_13-12- 13-05-00_22	NSA review to leave spying programs largely unchanged, reports say — World news — The Guardian	2.59542463
tm_13-12- 13-18-00_14	The Mission to De-Centralize the Internet : The New Yorker	3.45489878
tm_13-12- 25-05-00_11	Snowden to warn Brits on Xmas telly: Your children will NEVER have privacy The Register	3.640524
tm_13-10- 04-12-00_19	Attacking Tor: how the NSA targets users' online anonymity — World news — theguardian.com	4.52523438
tm_13-10- 15-21-00_16	Exclusive: Greenwald exits Guardian for new Omidyar media venture — Reuters	5.62568808
tm_14-03- 31-07-00_14	The Best NSA Fix Comes From the Patriot Acts Author - The Daily Beast	6.90953005
tm_14-02- 10-14-00_14	The NSA's Secret Role in the U.S. Assassination Program - The Intercept	6.93791211
tm_14-04- 01-06-00_17	NSA chief's legacy is shaped by big data, for better and worse-Los Angeles Times	6.984847
tm_13-10- 09-20-00_14	Schneier on Security: The NSA's New Risk Analysis	7.02368894

tm_14-01- 24_02-00_18	Snowden: 'Not all spying bad' but NSA program 'divorced from reason' - CNET	7.32241117
tm_14-01- 30_02-00_20	US intelligence chief has 30 days to reveal if specific citizens were spied upon — The Verge	9.93134805
tm_14-04- 14_22-00_22	Guardian and Washington Post win Pulitzer prize for NSA revelations — Media — The Guardian	nan

Table C.3: Top 50 results for "Edward Snowden NSA" query ranked by the regression model.

APPENDIX D

Weighted-Feature Rank Full Results

Headline	Chain_id	Weighted-Feature Rank Score	duration	Average number of articles	Highest Rank (Normalized)	Rank Duration	Average Cluster Cohesiveness
Instagram Testing Facebook Places Integration To Replace Foursquare — Fast Company — Business + Innovation	tm_14-03-25_16-00_24	0.274862379	3	4	0.9230769231	1	0.0636525737
Adobe Expands Its Marketing Cloud With Predictive Tools, iBeacon Support, And More — TechCrunch	tm_14-03-25_15-00_25	0.3719950259	1	9	0.9615384615	1	0.0265372192
Music Piracy Goes Mobile — Re/code	tm_14-03-25_11-00_16	0.9493952863	5	3	0.6666666667	1	0
TV Check-In Company Viggie Buys Facebook Publisher Wetpaint - Peter Kafka - Media - AllThingsD	tm_13-12-16_20-00_21	1.0650102371	3	5	0.8333333333	1	0.0077077056
Hortonworks raises \$100M to scale its Hadoop business Tech News and Analysis	tm_14-03-25_17-00_25	1.1556779302	1	12	0.9259259259	1	0.057826881
Instagram Testing Facebook Places Integration To Replace Foursquare — Fast Company — Business + Innovation	tm_14-03-25_13-00_21	1.1843456989	9	4	0.56	1	0.0636525737
Violin Memory Fires CEO Basile As IPO and Quarterly Results Disappoint - Arik Hesseldahl - News - AllThingsD	tm_13-12-16_18-00_20	1.4544497721	9	9	0.9166666667	1	0.045117474
Cathy Edwards, Co-Founder Of Chomp, Is Leaving Apple On April 11 — TechCrunch	tm_14-03-25_17-00_16	1.4603968841	1	5	0.5925925926	1	0.0258751743
10,000 GitHub users inadvertently reveal their AWS secret access keys	tm_14-03-25_07-00_17	1.8022569529	7	4	0.7619047619	1	0.0276757557
Waze Co-Founder Skips Google to Try Startup World Again	tm_14-03-25_15-00_22	2.3414478522	5	5	0.6923076923	1	0.0161284922
Avago to Buy LSI for \$6.6 Billion - NYTimes.com	tm_13-12-16_14-00_12	4.1398050273	1	12	0.75	1	0
Secure Domain Foundation Debuts to Fight Internet Domain-Based Threats	tm_14-03-25_09-00_15	4.2195271253	5	6	0.6086956522	1	0.0406287646
Crowdfilt Raises Another \$23 Million From Andressen Horowitz & Others For International Expansion, Enterprise Tools — TechCrunch	tm_13-12-16_17-00_18	4.4532762316	13	7	0.8095238095	1	0.0626230757

A Stream of Music, Not Revenue - NYTimes.com	tm_13-12-16_07-00_4	4.9783635729	21	8	0	1	0.0011533176
Attackers get cash out of ATMs by sending SMS messages - SC Magazine	tm_14-03-25_11-00_11	6.5277171905	7	8	0.4583333333	1	0.0496305837
Facebooks Open Compute guru Frank Frankovsky leaves to build optical storage startup Tech News and Analysis	tm_14-03-25_14-00_17	6.9897697177	7	9	0.68	1	0.0613791573
Nokia schedules press event next week, new Windows Phone 8.1 hardware likely — The Verge	tm_14-03-25_12-00_20	8.1552858847	7	12	0.52	1	0.0620203143
Quantum Computing Research May Back Controversial Company - NYTimes.com	tm_14-03-25_06-00_17	8.6127780864	13	4	0.347826087	1	0
Xiaomi's Hugo Barra: True world phones in 2 years, Android all the way - CNET	tm_14-03-25_00-00_18	12.1029091945	25	5	0.3043478261	1	0.0427536593
Google Deal With Luxottica Will Bring Glass to Ray-Ban, Oakley - WSJ	tm_14-03-25_01-00_18	12.7987390808	1	27	0.8181818182	1	0.0553357454
Sprint launches LTE in St. Louis, San Diego; wraps up 2013 with 300 4G cities Tech News and Analysis	tm_13-12-16_18-00_18	15.1650363187	21	14	0.5833333333	1	0.0486776759
Adobe Expands Its Marketing Cloud With Predictive Tools, iBeacon Support, And More — TechCrunch	tm_14-03-25_10-00_13	16.3420534401	11	11	0.375	1	0.0301412652
Facebook to Acquire Oculus — Facebook Newsroom	tm_14-03-25_18-00_0	18.2953681603	1	13	0	1	0.0438843113
Microsoft makes source code for MS-DOS and Word for Windows available to public - The Official Microsoft Blog - Site Home - TechNet Blogs	tm_14-03-25_13-00_17	19.3684056436	9	10	0.3076923077	1	0.0468685437
Nvidia unveils next-generation graphics processor with 3D memory — VentureBeat — Gadgets — by Dean Takahashi	tm_14-03-25_15-00_12	19.8065400476	5	14	0.4615384615	1	0.0316119256
Amazon Workers in Germany Strike Again - NYTimes.com	tm_13-12-16_15-00_12	21.0717750883	11	13	0.7058823529	1	0
Hortonworks raises \$100M to scale its Hadoop business Tech News and Analysis	tm_14-03-25_05-00_10	29.0275174213	25	12	0.3333333333	1	0.057826881

Amazon Workers in Germany Strike Again - NYTimes.com	tm_13-12-16_13-00_13	29.0408997659	15	13	0.625	1	0
Nvidia announce GeForce GTX Titan Z, brings 12GB VRAM for \$3,000 — PC Gamer	tm_14-03-25_14-00_21	29.8795628591	7	24	0.4615384615	1	0.0383894023
Exclusive: Twitter working on edit feature for tweets	tm_13-12-16_20-00_11	34.2355132348	11	11	0.3913043478	2	0.0468467143
Shameless Carriers — Monday Note	tm_13-12-16_09-00_12	42.5595332866	17	20	0.6428571429	2	0.0934430375
YC-Backed Gbatteries Launches BatteryBox, A 50Whr Backup Battery For MacBooks & Other Gadgets — TechCrunch	tm_14-03-25_00-00_7	43.1608252551	27	9	0.2173913043	1	0.0546051454
S.F. rolls out 3 miles of free Wi-Fi along Market Street - SFGate	tm_13-12-16_18-00_19	47.4180921332	27	15	0.5652173913	1	0.0603421867
Steam Machine specs, unboxing and gameplay videos posted by beta users — BGR	tm_13-12-16_20-00_19	57.4708902644	39	17	0.1923076923	1	0.0337541064
Sony won't use Android Wear, will stick with Smartwatch (for now) - CNET	tm_14-03-25_14-00_10	58.7310160998	7	20	0.4	1	0.0542827578
Two More Executives Leaving BlackBerry - WSJ.com	tm_13-12-16_13-00_8	82.2447885446	23	17	0.5	1	0.0292730318
Sony won't use Android Wear, will stick with Smartwatch (for now) - CNET	tm_14-03-25_11-00_3	92.7244941302	13	21	0.125	1	0.0542998054
Google will not answer to British court over UK privacy claim — World news — The Guardian	tm_13-12-16_06-00_12	100.059505078	29	17	0.2857142857	1	0.0626365521
Amazon Apologizes After Customers Lose Access to Christmas Content - Peter Kafka - Media - AllThingsD	tm_13-12-16_16-00_10	102.019186833	19	21	0.380952381	1	0.0342843073
Facebook Launches Donate Button For Non-Profits That Also Collects Billing Info For Itself — TechCrunch	tm_13-12-16_17-00_12	103.244990548	25	21	0.5652173913	1	0.0553043225
Spotify Pitches College Kids: Half Off Subscriptions While You're in School — Re/code	tm_14-03-25_11-00_4	112.324956037	13	23	0.1666666667	1	0.031427237
AOL Chiefs White Whale Finally Slips His Grasp - NYTimes.com	tm_13-12-16_05-00_4	123.886077827	33	18	0.2142857143	1	0
Apple Testing Related Search Suggestions On The App Store MacStories	tm_14-03-25_11-00_2	183.032979369	13	26	0.0833333333	1	0.0511700188

Apple Is Building an RTB Platform to Sell In-App Ads — Adweek	tm_13-12-16_17-00_10	224.715448853	37	17	0.2692307692	1	0.0510419427
A Microsoft Star Goes to Google - NYTimes.com	tm_13-12-16_13-00_5	244.625775272	41	21	0.2083333333	2	0
Obama To Meet Tech Execs Over NSA Spying, Obamacare Website — TIME.com	tm_13-12-16_18-00_9	466.894967984	37	30	0.2307692308	1	0.0456620461
Windows Phone 8.1 includes notification center and Siri-like personal assistant — The Verge	tm_13-12-16_12-00_10	754.913960228	43	38	0.2083333333	1	0.0625767494
Obama to Call for End to N.S.A.s Bulk Data Collection - NYTimes.com	tm_14-03-25_01-00_2	847.238420178	33	54	0	11	0.0604446023
The new HTC One is available in Google Play and Developer editions	tm_14-03-25_12-00_2	866.281192419	11	77	0	4	0.0416338994
Edward Snowden says judge's ruling vindicates NSA surveillance disclosures — World news — theguardian.com	tm_13-12-16_14-00_3	2343.9207267	83	70	0	2	0.0422667419
Box Files For 250M/POOnFull — YearRevenueO/124M, Net Loss Of \$168M	tm_14-03-24_18-00_2	2577.97083142	39	53	0	7	0.0335500791

Table D.1: Weighted-feature rank for human-rated dataset.