Access to Billions of Pages for Large-Scale Text Analysis

Peter Organisciak, Boris Capitanu, Ted Underwood, J. Stephen Downie
University of Illinois at Urbana-Champaign

Abstract
Consortial collections have led to unprecedented scales of digitized corpora, but the insights that they enable are hampered by the complexities of access, particularly to in-copyright or orphan works. Pursuing a principle of non-consumptive access, we developed the Extracted Features (EF) dataset, a dataset of quantitative counts for every page of nearly 5 million scanned books. The EF includes unigram counts, part of speech tagging, header and footer extraction, counts of characters at both sides of the page, and more. Distributing book data with features already extracted saves resource costs associated with large-scale text use, improves the reproducibility of research done on the dataset, and opens the door to datasets on copyrighted books. We describe the coverage of the dataset and demonstrate its useful application through duplicate book alignment and identification of their cleanest scans, topic modeling, word list expansion, and multifaceted visualization.

Keywords: non-consumptive research; feature extraction; large-scale text analysis; datasets; text mining

DOI: Citation info is to be added.

Contact: organis2@illinois.edu.

1 Introduction

Individual and consortial library digitization efforts around the world have been scanning a massive number of books and other library items. Projects such as Google Books and the HathiTrust have resulted in non-trivial, petascale collections of the world's published cultural heritage. Such digitization efforts are remarkable for the span of time they represent, often hundreds of years, and their span of cultures. For example, the HathiTrust collection, housed at the University of Michigan, alone comprises 14.7 million volumes taking up 659 TB of disk. Though library digitization efforts are primarily intended for preservation and document access, they also pave the way to new forms of large scale research. As data, they provide for a wealth of linguistic, cultural, historic, or even structural insights, providing researchers evidence for modelling language across more sub-facets and broad time periods. We present the HTRC Extracted Features (EF) dataset, a public research-oriented dataset of page-level features extracted from 4.8 million digitized public-domain books (referred, generally, as *volumes*) provided by the HathiTrust Digital Library. The EF dataset is notable because of its scale, the ease of access allowed by its non-consumptive design, and the ease of use for reproducible research enabled by its preprocessed and cleaned format.

The EF dataset is a general-purpose dataset, oriented toward research uses that necessitate a long view of the published word or breadth across different topics or languages. Here we demonstrate a number of those uses for information science: duplicate book alignment and identification of cleanest scans, topic modeling, word list expansion, and multifaceted visualization.

With typical research datasets the text analysis process starts with feature extraction, followed by computation over those features (e.g. modeling, counting). In contrast, the EF dataset has already completed the costly first part of that process. In addition to eliminating the effort- and resource- costs associated with feature extraction, this type of dataset allows for more readily reproducible experiments. While the EF dataset is currently derived from works that are in the US public domain, there is also a pragmatic access reason for a feature dataset: it abstracts away from the original full-text in a way that does not redistribute access-restricted scans, and offers a roadmap toward datasets from texts that are copyrighted or of unknown status (orphan works). Such use is referred to as non-consumptive because it cannot be enjoyed in a traditional sense by a person, but little is lost from a computational point of view.

The most important markers of a text's meaning are the words, and it follows that of the various features offered in EF, the most broadly useful ones are term frequency counts. The EF dataset offers such counts at the page-level for each page of each volume, tagged by the part-of-speech in the context that they are used. Though positional information is not included, these bags-of-words (BOW) are in a small enough context that they can be quite discriminating in informing a scholar what a page is about.

Particularly useful for clean use of the data, token counts are disambiguated by head, body, and footer on the page. Token counts and other features are noted by section, which makes it very easy to focus solely on content of a page without confounding textual information such as page titles, chapter titles, and page numbers. Other features provided in EF include counts of sentences, lines, and empty lines on the page, page-level language inference, counts of characters that occur at the start and end of lines, and a count of the longest length of capital characters starting a line.

A massive but granularly facetable corpus of scanned texts can support numerous information science uses. In this paper, we demonstrate using EF to:

- *Identifying duplicate books and selecting a best-scanned copy*: By processing a collection of literature in EF into smaller-dimensional 'fingerprints', we show that pairwise similarity between books is tractable for identifying multiple copies of a book in the HathiTrust collection, linking a book not only to its identical printings but to different publications of that book. Not all texts are digitized equally, so we introduce a method to surface candidates for the cleanest copy of a book.

- *Topic modeling concepts across books*: We provide a demonstrative example of how EF can support mixed-model soft clustering with Latent Dirichlet Allocation (LDA) in order to find conceptual 'topics'. Since topic modelling is concerned with coherent conceptual term collocation patterns, various training approaches can be used with EF, including filtering to more interesting parts of speech and training at a page-level document frame.

- *Word list generation*: Using a list of related words unfolded from a seed word is a portable way of following higher-level concepts in a text. It is also used in information retrieval for query expansion, expanding searches beyond the exact keyword of a query. By leveraging a topic model built on literature, we show a simple case of word list generation derived from EF.

- *Multifaceted visualization*: To ease exploratory analysis against the overwhelming scale of 5 million books, we built a multifaceted, interactive visual interface to the EF. Built on top of the open-source tool Bookworm, our publicly-accessible implementation can display trends in different subsets of the data; for example, comparing the use of the word 'lady' in British versus American texts.

The possibilities for use of the EF dataset extend well beyond those that we demonstrate. In information retrieval, for example, it can assist retrieval by augmenting collection models for historical or domain-specific collections, or be used to training structural classifiers for book-parsing to improve index. The EF dataset is also appropriate for supporting research in full-text book search, a direction of research that has been impeded by access to large corpora, the variability of documents, and difficult of evaluation (Kazai, Kamps, Koolen, & Milic-Frayling, 2011). EF's accessibility gives it potential as a test book collection, though its scale and breadth makes it particularly viable for use in developing models of text in different domains, time periods, and languages. In cataloguing, the broad coverage of the dataset can be used for outlier detection to find possibly misclassified texts. Beyond higher-level modelling, we anticipate the value of studying the content itself, for scaling up research questions in computation social sciences and the digital humanities. A deliberate use of EF can follow discourse of a topic over time, look at the rise of a cultural trend or linguistic shift, or observe how the structure of the book has changed.

2 Dataset

The HTRC Extracted Features dataset covers slightly over 4.8 millions volumes, comprised of 1.8 billion pages. Each volume is represented as an individual file, structured in JSON and accessible compressed using the `rsync` utility that is common on Unix-like systems. Details for access are available at `http://dx.doi.org/10.13012/J8X63JT3`.

The volumes represented in the EF dataset are from the HathiTrust Digital Library, a consortium of institutions collecting their digitized collections into a single digital library. This prominently includes volumes from libraries and other institutions across the US, including those scanned by the Google Books

Table 1: Most-represented languages

Language	Count	Percentage of Volumes
English	2705542	58.52%
German	564463	12.21%
French	528817	11.44%
Spanish	142874	3.09%
Italian	125561	2.72%
Latin	114643	2.48%
Japanese	73447	1.59%
Russian	59936	1.30%

Table 2: Most-represented classes (Library of Congress)

LC class	count of volumes
Language and Literature	385044
General and Old World History	244976
Social Sciences	212951
Science	184441
Philosophy, Psychology, and Religion	162939
Law	126709
Technology	120953
General Works	108515

project, but also holds contributions from non-US institutions. The underlying materials were scanned and their full text is parsed from those scans. The EF dataset does not share full text, however: only the quantitative feature counts that may be needed in computation analysis or modeling of the volumes.

In the interest of long-term preservation, the EF dataset release is maintained with a DOI (digital object identifier) through the University of Illinois Library, which is intended to persistently point to the current hosting for the EF data. The dataset is licensed with a Creative Commons Attribution License, which hews closely to academic convention by allowing any form of usage or redistribution in return for attribution.

2.1 Coverage

There are 512.1 billion unigrams across the 1.8 billion pages of the EF dataset. The version of the dataset described here is specific to public domain works, but this was a temporary restriction: we have released an expanded dataset including in-copyright and orphan works, three times larger, since this paper was accepted for publication.

The EF dataset covers 344 languages. As the source materials are primarily digitized from US-based academic libraries, English is the best-represented language, with 58.5% of the collection, though German, French, Spanish, Italian, and Latin all have over 100000 texts represented. The computed part-of-speech tags are only accurate for a subset of the languages.

Temporally, 95% of the dataset covers documents published between the years 1722 to 1987. Figure 1 shows the date distribution. Dates are taken from metadata records for the represented volumes. A troublesome but common classification quirk in libraries is confounding accurately classified dates with century-floored approximate dates (e.g. entering 1800 to denote a 19th century text rather than one published in exactly that year), and our use of bibliographic metadata means these errors are retained in the EF dataset.

Cross-referencing the texts with bibliographic metadata from the HathiTrust, we can find Library of Congress classifications for approximately 49% of the volumes, showing that the texts span all 21 top-level classes. The best-represented is class P (Language and Literature).

The corpus underlying the EF dataset benefits from a mostly indiscriminate digitization policy, meaning that much of the collection cuts broadly across the holdings of the participating institutions. There are benefits to smaller carefully curated digital collections, like the ability to correct individual OCR or

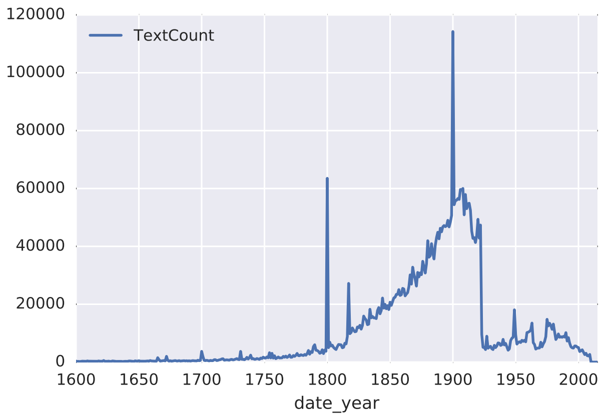


Figure 1: Distribution of volumes in dataset by year.

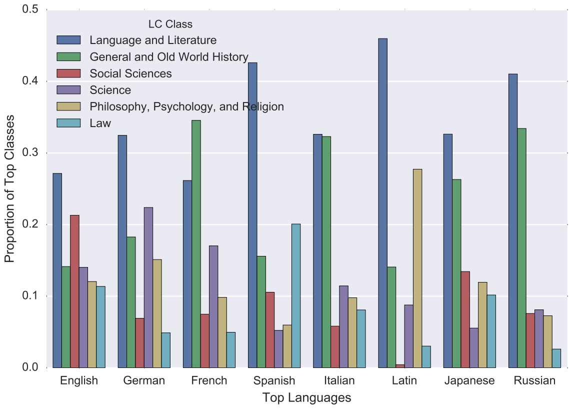
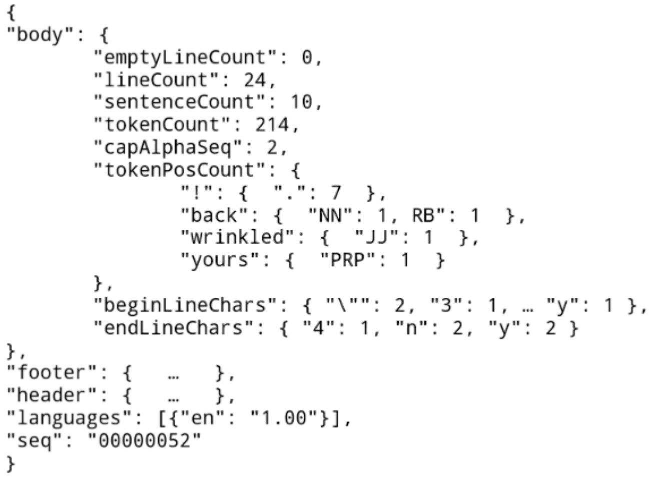


Figure 2: Distribution of the top classes among the top languages.

metadata errors. At the scale of the HathiTrust many such fixes are intractable; instead, individual errors in a book are smoothed over by the larger statistical significance afforded.

Despite the broad coverage, there is no reason to assume that the collection is balanced in its strengths, and understanding potential biases is useful for proper usage of the dataset. While the digitisation was often indiscriminate, the actual holdings of contributing institutions has biases: certain types of text are favoured more by academic institutions (the most common type of contributor), and some texts may be popular enough that the collection holds duplicate copies from multiple contributors. Figure 2 shows the distributions of the top classes within a selection of languages. Note that the classes are distributed differently; for example, Spanish is proportionally well-represented by literature, Latin by philosophy, and German by science. It is unknown how much of these distributions represents collection bias (what materials were held) and which represent the distribution of what is published in that language.

A more significant change in distribution representation happens due to copyright status. All the volumes represented in the EF dataset describe here are US public domain, leading to some some caveats about the dataset. Copyright determinations in the US vary depending on the circumstances of the work, but works published before 1923 are generally in the public domain. As a result of that year's transition from universal to contextual rules, there is a drastic shift in collection coverage at that point. This is seen in the quantity of volumes represented (see 1923 in Figure 1) and in the genres that are seen. For example, volumes in the sciences and social sciences increase proportionally from pre-1923 to 1923-, while history falls and literature falls precipitously.



```
{
"body": {
        "emptyLineCount": 0,
        "lineCount": 24,
        "sentenceCount": 10,
        "tokenCount": 214,
        "capAlphaSeq": 2,
        "tokenPosCount": {
                "!": {  ".": 7  },
                "back": {  "NN": 1, RB": 1  },
                "wrinkled": {  "JJ": 1  },
                "yours": {  "PRP": 1  }
        },
        "beginLineChars": { "\"": 2, "3": 1, … "y": 1 },
        "endLineChars": { "4": 1, "n": 2, "y": 2 }
},
"footer": {   …   },
"header": {   …   },
"languages": [{"en": "1.00"}],
"seq": "00000052"
}
```

Figure 3: Truncated example of a features for a single page.

3 Related Work

The Google Books Ngrams Corpus (Lin et al., 2012; Michel et al., 2011) provides token counts for n-grams, from unigrams to 5-grams, that occur in volumes scanned by the Google Books project. It is comprised of a similar breadth of materials as EF; indeed, a notable portion of the corpus underlying the EF dataset is from Google Books. Where our dataset differs in is in format, providing counts for each page of each volume, while Google's dataset only provides corpus-level counts though with longer phrases. The NGrams corpus includes information on copyrighted volumes, something not yet released publicly for EF. Finally, the EF dataset includes useful cleaning, such as the header/footer extraction, and provides additional features beyond ngrams.

Data for Research (DfR) from JSTOR (Burns et al., 2009) is another notable historical resource. DfR provide document-level n-gram counts for 1-4 gram counts, as well as $TF * IDF$ weighted lists of discriminatory terms on the documents. These can be downloaded for up to one thousand documents freely, or more with permission. The JSTOR collection holds primarily academic materials with a strength in digitized articles. The EF dataset differs in the scope of the collection, spanning published work more general, and has a different access model, with open access to preprocessed features of the entire collection.

Since the EF dataset was publicly released in 2015, with a smaller demonstrative dataset a year earlier, it has seen some researcher use and redistribution of recombinant parts. Underwood (2014; 2015) inferred genre labels for fiction, poetry, and drama; Forster (2015) inferred gender for the authors of a selection of literature in the collection; Goodwin (2015) trained mixture models of commonly occurring themes in fiction; finally, Mimno (2014) calculated co-occurrence tables for terms that co-occur in each year from 1800 to 1923.

4 Features

The EF dataset contains extracted information about digitized volumes as well as a small amount of metadata. The metadata includes the publication date (`pubDate`), title (`title`), bibliographic language (`language`), imprint information about the publishing context (`imprint`), and a set of identifiers for different contexts (`id`, `htBibUrl`, `handeUrl`, `oclc`). The primary purpose of the dataset is features, so additional metadata must be obtained from secondary sources like the HathiTrust Bibliographic API [1].

At the level of the volume, the only feature provided is `pageCount`, a count of pages in the volume. Other features are provided at the page-level.

At the level of the page, we provide information by section: *header*, *body*, and *footer*. Headers and footers often contain information that is paratextual – related more to the structure of the book rather than

[1] https://www.hathitrust.org/bib_api

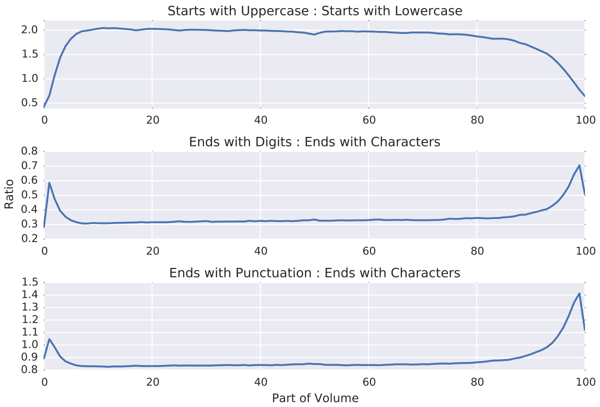


Figure 4: Ratios of notable character types at the start and end of pages through volumes.

the core content – such as headings, titles, and page numbers. They also tend to repeat over multiple pages, resulting in skewed word distributions. Headers and footers are processed using a custom two-pass algorithm that looks for recurring text at the top and bottom of each page.

For each section of the pages, we provide counts for `tokenPosCount`, `tokenCount`, `sentenceCount`, `lineCount`, `emptyLineCount`, `beginLineChars`, `endLineChars`, while `capAlphaSeq` is provided exclusively for the body of each page.

`tokenPosCount` provides an unordered list of all occurring tokens in that section, with counts. Counts are provided by part of speech and are case-sensitive, so 'Jaguar' (`proper noun`), 'jaguar' (`noun`), and 'Jaguar' (`noun`) are disambiguated, as are 'rose' (`verb`) and 'rose' (`noun`). The tokenization and part-of-speech tagging is done by OpenNLP (Apache, 2005), with the part of speech tags following those of the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). `tokenCount` further provides the total count of tokens in the given section of the page, for convenience.

`sentenceCount` and `lineCount` provide counts of sentences and lines, where the former refers to the textual content while the latter refers to the physical structure. Sentence segmentation is done by Apache OpenNLP (Apache, 2005). Sentences that started on a different page are still counted, meaning that a sentence spanning a page break will be counted once for each page. Lines refer to the vertical lines of text physically on the scanned page. Additionally, `emptyLineCount` describes the number of lines that do not contain any content. This is interpreted based on the OCR process for a scanned page, which may vary for volumes scanned by different sources. Multiple consecutive empty lines are not counted, so empty line count in many cases is a proxy toward inferring the number of paragraphs on a page (i.e. $count + 1$).

`beginLineChars` and `endLineChars` count the characters along the left-most and right-most margins of a page, respectively. This information is useful for identifying the type of text on the page (Underwood, 2014). For example, lines of poetry may start with capitalized characters and end with punctuation frequently, prose may have a varied distribution of characters, or a table of contents may have many numeric values at the end of a line.

Finally, `capAlphaSeq` counts the longest length of consecutive alphabetical characters in the given section. Again, this information provides hints as to what type of content is on the page, and lock sequences of capital letters suggest back of the book indexes or title pages.

Provided at the page-level but not separated by section is an inferred language field. Even though there is a bibliographic language classification for each volume, there are instances where it may not be correct, or a book may have multiple languages within it. For this reason, the `languages` feature provides language likelihoods for each page, inferred by software from Shuyo (Shuyo, 2010).

Figure 4 shows the ratios of different notable characters at the start and end of lines, through a book, demonstrating the ability of these features to discriminate between parts of books and eventually between text and paratext. The information shown here is unsupervised, without annotation of what front matter is, what a table of contents is, or what prose is. Still, we see indicators of when books look different at the start and end, where the majority of paratext tends to occurs. In Figure 4, we see that uppercase characters on the left-most side of a page are much more common at the end of a book. At the same time, seeing punctuation or digits at the right-side side of a page is an indicator of the type of content we see at the start and end of a book: perhaps table of contents or a back of the book index. This form of information can be used robustly for classification; for example, it has been used to infer book genre (Underwood, 2014).

5 Tools

Since it is publicly available via `rsync`, structured in JSON, and permissively licensed with the CC-BY Attribution license, the EF dataset can be accessed by anybody and used however they may desire. To aid usage of the dataset, the HTRC Feature Reader library has been released for Python (Organisciak & Capitanu, 2016). The primary goal of this library is to simplify in-memory use of the EF dataset and to provide scaffolding for using it efficiently within the popular Scipy Stack for scientific work using Python.

6 Demonstrative Use

To demonstrate use of the EF dataset, we performed a duplicate text alignment and selection of best copies, topic modeling, word list generation, and multifaceted interactive visualization. These are intended as potential but realistic uses.

6.1 Similarity Between Texts

The EF holds features for each digitized volume copy in the corpus, which includes duplicate copies as well as reprints of the same book. A scholar may want to identify these texts, either to connect a text through its reprintings, make sense of what texts are in an anthology of works, or to filter an analysis to only one copy of each text. The information held in the EF is enough for this task, enabling candidates for duplicate or overlapping texts to be surfaced.

We performed a duplicate candidate evaluation on a set of literature texts (`n=101947`) (Underwood et al., 2015). Specifically, 33 works known from metadata records to have at least twenty duplicate copies were sampled, from which a 'target' text was randomly selected for each work. A similarity ranking was then performed to measure Precision at 20: how many of the twenty most similar texts are the same as the target text.

For this demonstration, we measured similarity by reducing each text to a smaller dimensional representation and performing a similarity measure to find the most similar texts to a target. Specifically, Latent Semantic Analysis was used to interpret dimensions against tf*idf weighted term-document matrices, and euclidean distance was used to measure similarity.

Since similarity was judged using only content from the books, the ground truth was exact metadata matches augmented with hand-checking. The hand-checking was necessary because sometimes the metadata is inconsistent, with typos or variant spellings. One target text was *Gulliver's Travels*, for example, a recently popularized title for a book that was actually published as *Travels into Several Remote Nations of the World. In Four Parts. By Lemuel Gulliver, First a Surgeon, and then a Captain of Several Ships.* An exact title match would not catch that a book by the latter title is nonetheless the same work as a book with the former title.

Measured on 33 sample works with at least 20 known duplicates in the 102k volume test corpus, the average precision at twenty for finding identical texts is $P@20 = 0.748$. While this means that 74.8% of matches are completely the same book at the target, the others are not completely unrelated. 2.8% of the texts returned are published subsets of the target book, while 10.8% are different books by the same authors as the target book. ### Selecting the cleanest single copy of a work

Since duplicate texts occur close together in an EF-trained reduced dimensional space, this can be utilized for a basic selection process for the best scanned copy of a text, which is to say the cleanest, with regards to the OCR quality. We attempted this by averaging a centroid between all the known duplicates of a book, and identifying the book closest to the centroid.

Against 663 works with duplicate copies (totalling 15084 copies), we evaluated the "book closest to the centroid" selection policy by vocabulary size. OCR errors lead to increased numbers of unique words, so we would expect that worse copies of a book would have more unique terms. For our best-copy selection

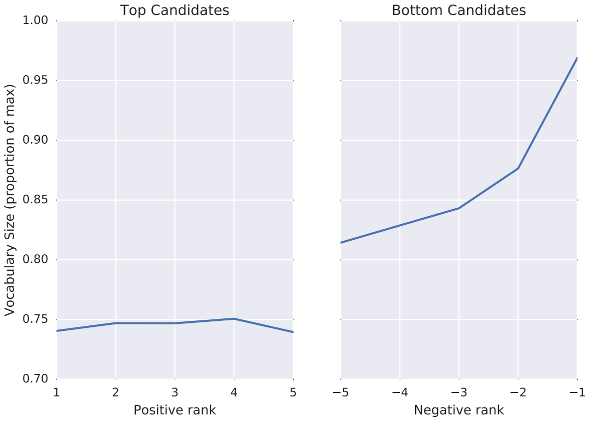


Figure 5: Median vocabulary size for top vs bottom 'best-copy' candidates, as a proportion of the max vocabulary size seem for each book.

process, we find that the top candidate has a median 1403 unique words less than the bottom candidate. Figure 5 shows this relationship to vocabulary holds for the top and bottom candidates.

6.2 Topic Modeling

Topic modeling refers generally to mixed model clustering trained on term co-occurrences, most popularly built using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). The dimensions of a well-trained model can show conceptual coherence, allowing them to be interpreted qualitatively as conceptual *topics*. Topic Modeling is possible with the bag-of-words information provided in the EF, and can benefit from the page-level granularity, stripped headers, and POS tagging in the dataset. We built an example model making use of these features and aided by randomized page sampling and asynchronous prior assumptions on individual topic probabilities.

As topic cohesion is desired with topic modeling, the training process is more discriminating than for dimensionality reduction, as done for our book similarity demonstration earlier. Dimensionality reduction aims to accounts for the most amount of variance in any way possible, but for topics we often strive for conceptually informative groupings of words. For our demonstration of topic modeling, we filtered out a number of word types that are not as interested for generalized topic models, including determiners, personal pronouns, modals, symbols and different forms of adverbs.

Our topic modeling demonstration was built with 400 topics. Topic models are trained on word co-occurrences, so it is better to train with many examples on a smaller document frame than on full books. We used pages as the training frame. After all, a word occurring on the first page of a text may not necessarily reflect a strong relationship to a word on the last page. Another technique we used was randomized, multi-pass sampling. Earlier training texts can exert outsized bias on a topic model, so we trained with multiple passes across the collection, each one sampling only a small part of a book. The first pass only used 1 or 2 pages from each book (specifically, 1/256) to avoid a convergence of topics too early, and subsequent passes gradually increased the per-book page sample size. The choice of training 400 topics was not made by any detailed method. It was intuitively chosen with the motivation of training excess topics, because it is easier to systematically ignore uninteresting extra topics than it is to recover interesting ones that never were trained. Finally, we trained with a set of asynchronous document-topic probabilities that put the majority of topical probability mass in the first few topics. This has the effect of serving as a 'catch-all' for words that are common across the language in general, allowing niche concepts a place for their own less common and often more interesting patterns, as well as a way to identify the less interesting ones.

In LDA, each model can be interpreted as a process that generates words, with different distributions for each term's likelihood to be generated by that model. For example, topic 359 is most likely to generate the words 'sin', 'Christ', 'grace'. Subsequently a text can be assigned a distribution of how likely each topic is to have generated the words in that text. The resulting topics can be used to interpreted in the contexts of

love
39: love, passion, loving, lovers, true, hate, Love, beauty, hearts, tender
359: sin, Christ, grace, Jesus, angels, God, Thy, heaven, be, Let
72: is, be, have, do, good, are, am, know, say, Nay
311: mine, heart, Alas, alas, weep, be, pity, embrace, sorrows, grieve

sad
29: grave, oh, Death, tomb, death, destiny, sadness, manhood, sad, bind
241: tears, heart, eyes, dream, grief, sorrow, arms, death, breast, dead
302: sweet, fair, heart, dreams, kiss, soft, dream, gentle, eyes, smile
212: life, hope, heart, Heaven, is, soul, joy, vain, hopes, woe

queen
325: de, Madame, M., queen, la, France, French, overlook, boudoir, juncture
379: Henry, K., lords, kingdom, War, Hen, crown, Clarence, Lords, Warwick
367: lord, honour, duke, noble, lordship, news, grace, breed, Cardinal, good
10: Ant, Caesar, Antony, Louise, sports, Julius, Cleopatra, Davis, client, Cleo

Figure 6: Top topics representing various terms.

love	**grief**	**forest**	**city**	**philosophy**	**cat**	**night**
passion	sorrow	sunshine	streets	philosopher	stuff	light
loving	dream	shadows	blow	principle	stick	day
lovers	breast	trees	fury	ideas	tail	morning
true	wept	branches	walls	exist	leg	sky
hate	weeping	rays	daring	cases	pockets	stars
beauty	anguish	twilight	roused	conception	reckon	dark
hearts	despair	grove	swords	intellect	pull	earth
tender	bosom	atmosphere	furious	individual	bit	darkness
wisest	bitter	plant	tower	consists	cage	bright

Figure 7: Word lists for a selection of emotional, topical, and setting-based seed words.

texts (e.g. “what are the concepts that make up Pride and Prejudice?”), in a global context (e.g. “what are the different types of topics that we see in literature?”), or at a word level (e.g. “what topics are likely to generate the word”love“?).

Figure 6 shows the top topics for three terms: ‘love’, ‘sad’, and ‘queen’. A qualitative interpretation would suggest that the topics are exhibiting some manner of coherence, such as love and God, love and passion, love and sorrow. Since verbs were not filtered, the top topics for ‘love’ include a topic of less interesting generally distributed words like ‘is’ and ‘be’; however, since we trained a catch-all topic, it is possible to programmatically filter such topics by measuring their similarity to the catch-all topic using a probabilistic distance metric like Hellinger Distance.

6.2.1 Word List Generation

Topic modelling can be further leveraged for a use common in information science: expanding a seed word into a list of related terms. In information retrieval this is referred to query expansion, which is used to match search results with appropriate terms beyond the exact query string that an information-seeking user typed into their query. In other areas, words lists are used as a convenient way to track concepts across a text, easier to transfer across collections and easier to interpret than topic models. One popular set of word lists is LIWC (Pennebaker, Francis, & Booth, 2001).

By normalizing word probabilities across topics and comparing their distance using Hellinger Distance, a seed word can be exploded into a list by identifying the words that occur across all topics most similarly to the seed. Figure 7 shows word list for a selection of keywords.

Previous work has leveraged word embedding models for word list generation (Fast, Chen, & Bernstein, 2016), which learn words by their immediate contexts. Skip-gram word embeddings are more conceptually appropriate for word list generation as their goal is to predict context words from a seed word; however, the EF dataset does not provide the positional information necessary for training a clean word embedding model.

6.3 Multi-faceted Visualization

A challenge to working with text data at the scale of EF is that even preliminary exploration is a time-consuming process, presenting a challenge to inductive inquiry. There is a value to being able to quickly

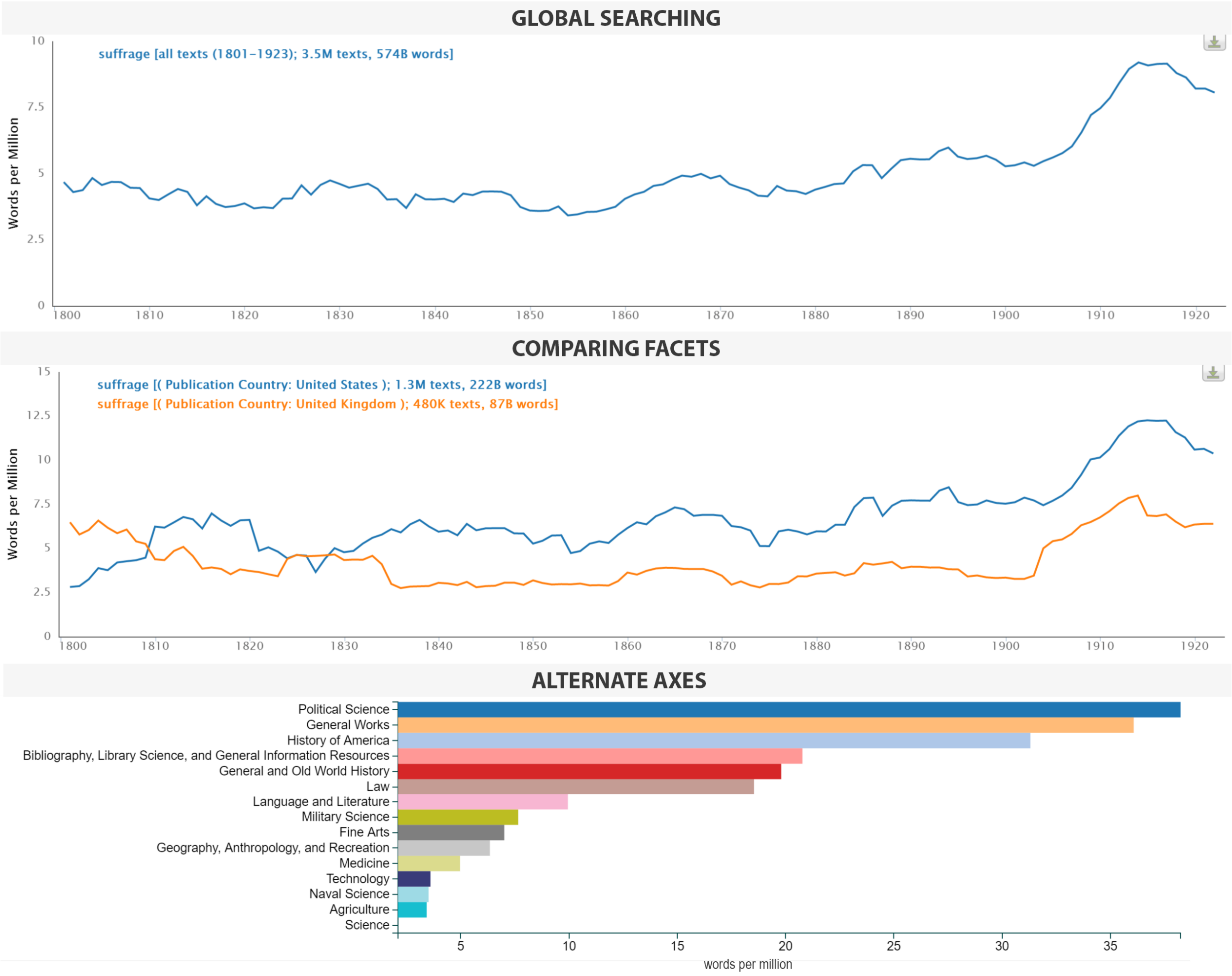


Figure 8: In-browser screenshots of three views of Bookworm-based EF visualization: a corpus wide search for ’suffrage’ (top), a comparison of the same search in British and American texts (center), and a non-date based search, showing relative usages of the term across Library of Congress classes (bottom,truncated).

explore trends throughout the collection to assess their tractability as a study topic. To support exploratory data analysis, we adapted the EF dataset for the collection visualization tool Bookworm, an evolution of work presented in (Michel et al., 2011). This implementation of Bookworm on the EF dataset is available publicly [2].

On the EF implementation of Bookworm, it is possible to observe longitudinal trends across all 5 million texts, or subfacets such as publication country or class (Figure 8 top and middle). It is also possible to observe trends that are not year-based (Figure 8 bottom), or even to query the system backend for the raw numbers to visualize yourself.

While the metadata was augmented from additional sources, the data underlying this visualization tool is from EF. One of the limitations here is that the solely unigram token counts in the EF keep the visualization from allowing phrases as search queries.

7 Conclusion

The EF dataset is designed to support a breadth of different research questions by providing access to millions of books in an open and straightforward way. Its strengths lie in the coverage of its collection – multiple

[2] http://analytics.hathitrust.org/bookworm

languages, varied domains, and spanning hundreds of years – as well as its preprocessed features format, which saves time and computational resources while also providing a standardized foundation for supporting different research needs. These circumstances make the EF dataset valuable for large-scale textual needs, such as topic modeling and similarity measurements between books.

The principle guiding the creation of the Extracted Features dataset is that of non-consumptive access, which seeks ways to nurture effective large-scale text research within the constraints of intellectual property laws. Recent work on the EF dataset has released the same features for 13.7 million volumes, including those that are in-copyright or of unknown status. This release was made possibly only by the non-consumptive structure of the texts.

We demonstrated the malleability of the EF dataset for text mining and analysis through a selection of example uses. Though the public release of this dataset, researchers in many domains can leverage it in pursuit of their own work.

References

Apache. (2005). *The opennlp project.* Retrieved from http://opennlp.apache.org/index.html

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Burns, J., Brenner, A., Kiser, K., Krot, M., Llewellyn, C., & Snyder, R. (2009). Jstor-data for research. In *Research and advanced technology for digital libraries* (pp. 416–419). Springer.

Fast, E., Chen, B., & Bernstein, M. (2016). Empath: Understanding topic signals in large-scale text. *arXiv preprint arXiv:1602.06979*.

Forster, C. (2015). *A walk through the metadata: Gender in the hathitrust dataset.* Retrieved from http://cforster.com/2015/09/gender-in-hathitrust-dataset/ (Blog)

Goodwin, J. (2015). *Creating a topic browser of hathitrust data.* Retrieved from http://jgoodwin.net/blog/creating-hathitrust-topic-browser (Blog)

Kazai, G., Kamps, J., Koolen, M., & Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: Impact of hit design on comparative system ranking. In *Proceedings of the 34th international acm sigir conference on research and development in information retrieval* (pp. 205–214). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2009916.2009947 doi: 10.1145/2009916.2009947

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

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

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

Mimno, D. (2014). *Word counting, squared.* Retrieved from http://www.mimno.org/articles/wordsim (Blog)

Organisciak, P., & Capitanu, B. (2016). Text mining in python through the htrc feature reader. *Programming Historian.* Retrieved from http://programminghistorian.org/lessons/text-mining-with-extracted-features

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

Shuyo, N. (2010). *Language detection library for java.* Retrieved from http://code.google.com/p/language-detection/

Underwood, T. (2014, 12). *Understanding genre in a collection of a million volumes, interim report.* Retrieved from https://dx.doi.org/10.6084/m9.figshare.1281251.v1

Underwood, T., Capitanu, B., Organisciak, P., Bhattacharyya, S., Auvil, L., Fallaw, C., & Downie, J. S. (2015). *Word frequencies in english-language literature, 1700-1922* (Vol. 0.2). HathiTrust Research Center. Retrieved from http://doi.acm.org/10.13012/J8JW8BSJ doi: 10.13012/J8JW8BSJ