



WHAT CAN YOU DO WITH TEXT MINING?

TEXT MINING CAN...

- **Summarize** topics of interest in a group of texts
Analysis method: Topic modeling & Clustering
- **Connect** common keywords among a group of texts
Analysis method: Network analysis
- **Track** sentiment over topic, text source, time period
Analysis method: Sentiment Analysis
- **Identify** names, locations, entities
Analysis method: Natural Language Processing
- **Distinguish** texts in a corpus by a given author (i.e. who authored which federalist paper)
Analysis method: Stylometry
- **Differentiate** poetry from prose
Analysis method: Text Classification
- **Contrast** the vocabulary of different corpora
Analysis method: Keyword/feature extraction
- **Categorize** documents
Analysis method: Document/term clustering

APPLICATION FOR TEXT MINING

SAMPLE USE CASES




TEXT MINING: CULTURAL STUDIES

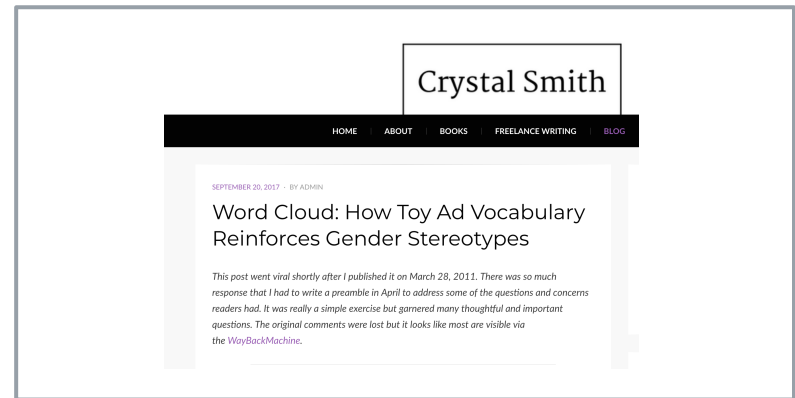
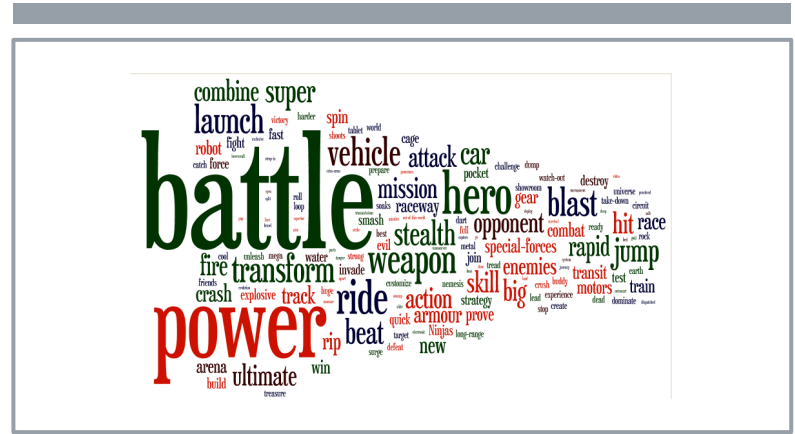
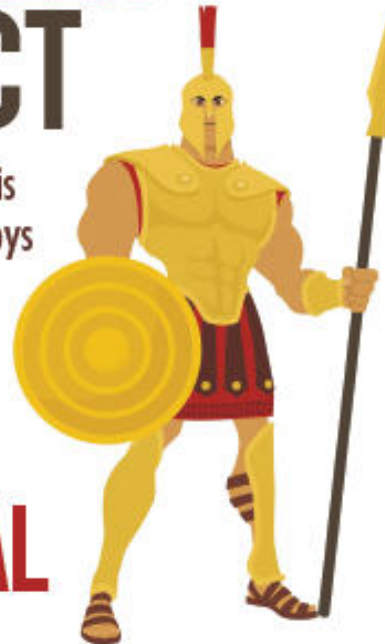
THE ACHILLES EFFECT

What Pop Culture is
Teaching Young Boys
about Masculinity

CRYSTAL
SMITH



**CRYSTAL
SMITH**



Slide adapted from: UNLV “Introduction to Text Analysis”

TEXT MINING: LITERARY NETWORKS

VIRAL TEXTS PROJECT

This site presents data, visualizations, interactive exhibits, and both computational and literary publications drawn from the Viral Texts project, which seeks to develop theoretical models that will help scholars better understand what qualities—both textual and thematic—helped particular news stories, short fiction, and poetry “go viral” in nineteenth-century newspapers and magazines.

Ryan Cordell and David Smith, *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (2017), <http://viraltexts.org>.



A "Stunning" Love Letter to Viral Texts

Like most nineteenth-century newspapers, *The Raftsmen's Journal* sought to connect its readers in rural Clearfield, Pennsylvania with wider worlds of news, information, and literature. Whether published in major metropolitan areas such as New York, Boston, and Philadelphia; in smaller cities such as Wheeling or Nashville; or in rural towns such as Clearfield, nineteenth-century newspapers relied on networks of exchange for much of their content. Newspaper editors subscribed to each others' newspapers, which came to them in the mail on post roads or, later, railroads.

When exchange papers arrived, editors would comb through them to find content their readers would appreciate, which they would then clip out with scissors and paste on sheets for their compositors to set in new type for the next daily, weekly, or irregular edition, sometimes changing the original text in the process. Sometimes a clipping would not be needed immediately, but would be saved for later use; we find clusters of reprinted texts that circulated in this way around the country over years or even decades.

Thus texts of all kinds—including news, fiction, poetry, vignettes, how-to columns, lists, descriptions of scientific and historical curiosities, etiquette, medical and health notes, business advice, parenting advice, recipes, religious affirmations, jokes, and more—circulated around the country, connecting readers from New England to New Orleans to California through shared texts.

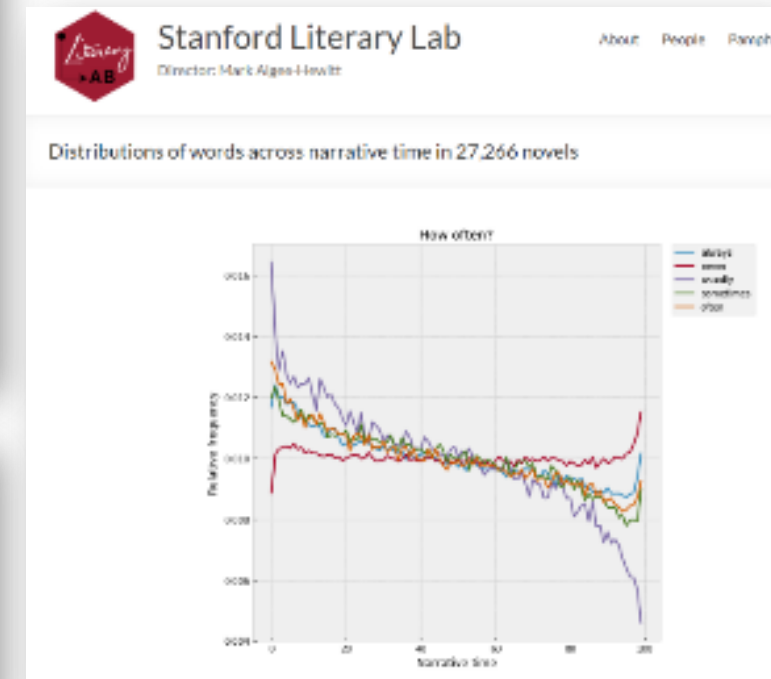
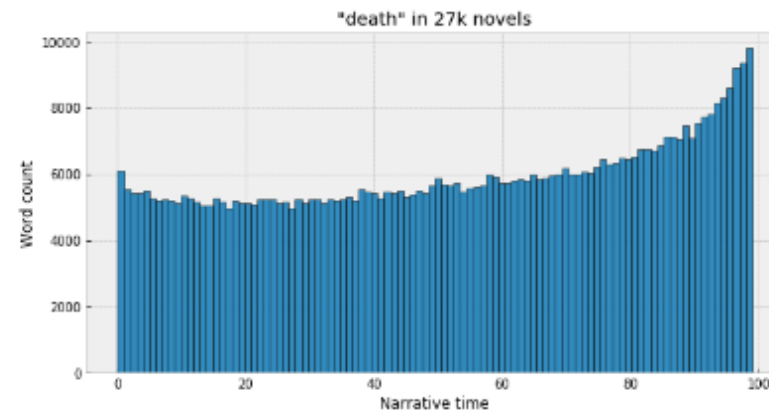
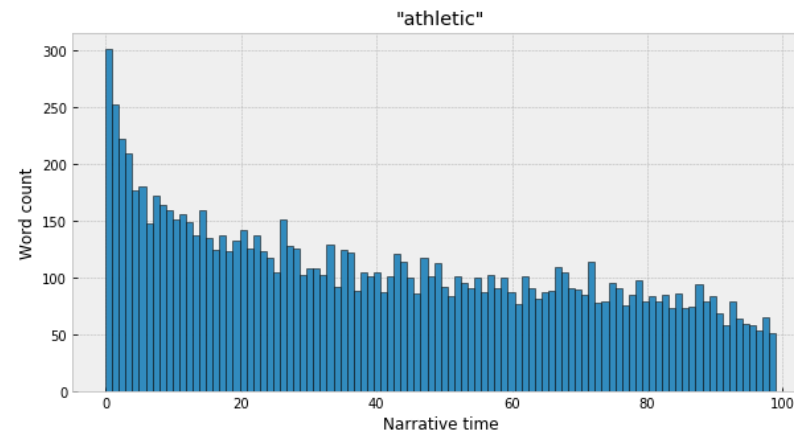
This exhibit is intended to hint at the breadth—and the oddities—of nineteenth-century reprinting that we have found thus far in the Viral Texts Project. If you peruse the page, you will find articles that link to our database, where you can browse versions that appeared in other newspapers, or related pieces.

TEXT MINING: NOVELS

AMERICAN FICTION

Positive adjectives and terms about family tend to dominate at the start of novels, and then tail off. Terms relating to death peak at the end of novels. There are some words (they've identified 200) that have a particular narrative "charge" (i.e. they dominate certain stages of a novel more than you'd expect),

David McClure
Stanford Literary Lab



TEXT MINING: HISTORICAL NEWSPAPERS

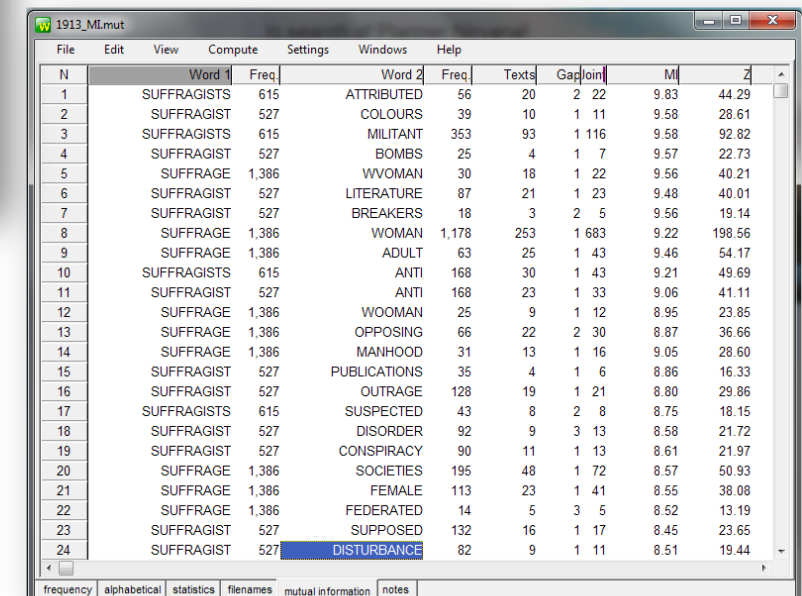
THE LANGUAGE OF BRITISH SUFFRAGE IN THE PRESS

Kat Gupta
University of Roehampton

TO THE EDITOR OF THE TIMES.

Sir,—May I express my entire agreement with the letter of Miss Milner in your issue of this morning? If the recent scenes of rowdiness associated with women's franchise only served to bring ridicule on the self-appointed champions of that cause other women might be well content to let the matter rest there. Unfortunately, such behaviour can only have the most mischievous effect in prejudicing the influence of women in those branches of public life where the beneficial character of their work is universally recognized.

It is often said of women that neither logic nor humour counts among their strongest points. The recent behaviour of the **suffragettes** would appear to support this contention. Mrs. Fenwick Miller's letter in *The Times* this morning is in every way a remarkable document. It opens up an attractive vista of the public results we might expect to follow from the establishment of feminine rule marked by such a judicious and temperate spirit, say, at the Board of Trade or India Office. As an onlooker nothing strikes me as more curious in this controversy than the unreasonable but most feminine desire of the **suffragettes** both to eat and to keep their political and domestic cake. Women cannot expect to have it both ways. They cannot at one and the same time



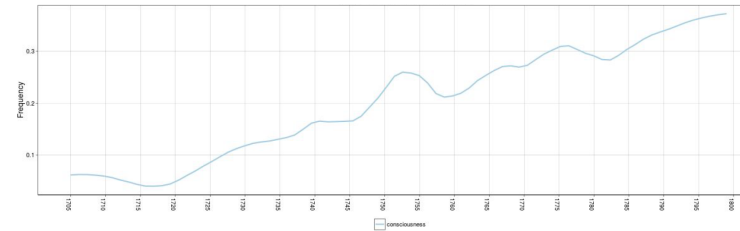
N	Word 1	Freq	Word 2	Freq	Texts	Gap	Join	MI	Z
1	SUFFRAGISTS	615	ATTRIBUTED	56	20	2	22	9.83	44.29
2	SUFFRAGIST	527	COLOURS	39	10	1	11	9.58	28.61
3	SUFFRAGISTS	615	MILITANT	353	93	1	116	9.58	92.82
4	SUFFRAGIST	527	BOMBS	25	4	1	7	9.57	22.73
5	SUFFRAGE	1,386	WVOMAN	30	18	1	22	9.56	40.21
6	SUFFRAGIST	527	LITERATURE	87	21	1	23	9.48	40.01
7	SUFFRAGIST	527	BREAKERS	18	3	2	5	9.56	19.14
8	SUFFRAGE	1,386	WOMAN	1,178	253	1	683	9.22	198.56
9	SUFFRAGE	1,386	ADULT	63	25	1	43	9.46	54.17
10	SUFFRAGISTS	615	ANTI	168	30	1	43	9.21	49.69
11	SUFFRAGIST	527	ANTI	168	23	1	33	9.06	41.11
12	SUFFRAGE	1,386	WOOMAN	25	9	1	12	8.95	23.85
13	SUFFRAGE	1,386	OPPOSING	66	22	2	30	8.87	36.66
14	SUFFRAGE	1,386	MANHOOD	31	13	1	16	9.05	28.60
15	SUFFRAGIST	527	PUBLICATIONS	35	4	1	6	8.86	16.33
16	SUFFRAGIST	527	OUTRAGE	128	19	1	21	8.80	29.86
17	SUFFRAGISTS	615	SUSPECTED	43	8	2	8	8.75	18.15
18	SUFFRAGIST	527	DISORDER	92	9	3	13	8.58	21.72
19	SUFFRAGIST	527	CONSPIRACY	90	11	1	13	8.61	21.97
20	SUFFRAGE	1,386	SOCIETIES	195	48	1	72	8.57	50.93
21	SUFFRAGE	1,386	FEMALE	113	23	1	41	8.55	38.08
22	SUFFRAGE	1,386	FEDERATED	14	5	3	5	8.52	13.19
23	SUFFRAGIST	527	SUPPOSED	132	16	1	17	8.45	23.65
24	SUFFRAGIST	527	DISTURBANCE	82	9	1	11	8.51	19.44

TRACING “CONSCIOUSNESS” IN PHILOSOPHICAL WRITING

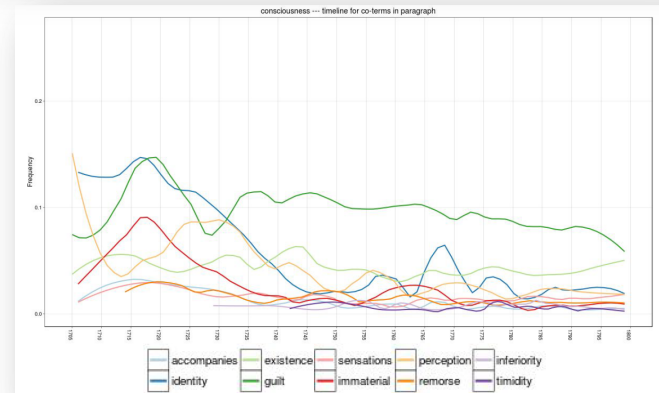
EIGHTEENTH CENTURY
COLLECTIONS. ONLINE

Helsinki Computational History Group
(COMHIS), University of Helsinki
<https://www.helsinki.fi/en/researchgroups/computational-history>

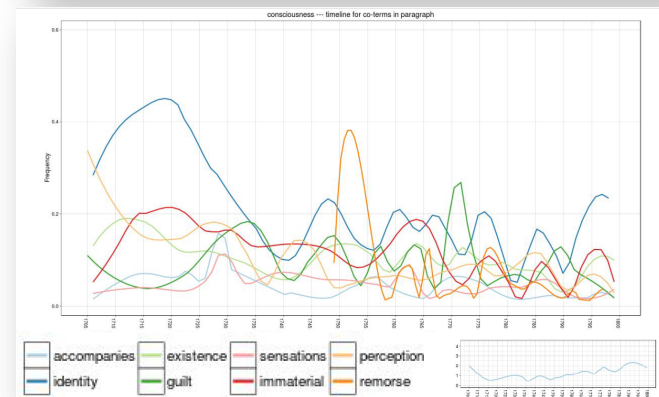
“Consciousness”



Frequency of the appearance of
“consciousness” in each
paragraph per year in the ECCO



Frequency of individual co-terms
that appear within paragraphs
regarding “consciousness” in al of
ECCO



Frequency of individual co-terms that
appear within paragraphs regarding
“consciousness” in “philosophers’
corpus”

IDENTIFYING COMMONPLACES AND OTHER FORMS OF TEXT REUSE AT SCALE

EIGHTEENTH CENTURY
COLLECTIONS. ONLINE

Dr Glen Rice (and team) Australian National
University
<http://dh2016.adho.org/abstracts/343>

an Hour of virtuous Liberty, Is worth a whole Eternity in Bondage
- Joseph Addison

Trigrams: hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage

an hour, of virtuous liberty Is worth a whole eternity in bondage
- James Thomson

Trigrams: hour_virtuous_liberty, virtuous_liberty_eternity, liberty_eternity_bondage

Most frequently commonplaced authors (ECCO, Literature & Language Module)

- | | |
|----------------------------|---|
| 1. Shakespeare, William | 16. Gildon, Charles |
| 2. Horace | 17. Young, Edward |
| 3. Pope, Alexander | 18. Congreve, William |
| 4. Milton, John | 19. Rider, William |
| 5. Virgil | 20. Cibber, Colley |
| 6. Ayscough, Samuel | 21. Griffith, Mrs. (Elizabeth) |
| 7. Bysshe, Edward | 22. Fénelon, François de Salignac de... |
| 8. Ovid | 23. Goldsmith, Oliver |
| 9. Terence | 24. Fenning, Daniel |
| 10. Dryden, John | 25. Addison, Joseph |
| 11. Becket, Andrew | 26. Walker, John |
| 12. Thomson, James | 27. Voltaire |
| 13. Cicero, Marcus Tullius | 28. Garrick, David |
| 14. Jonson, Ben | 29. Cibber, Theophilus |
| 15. Chambers, Ephraim | 30. Enfield, William |

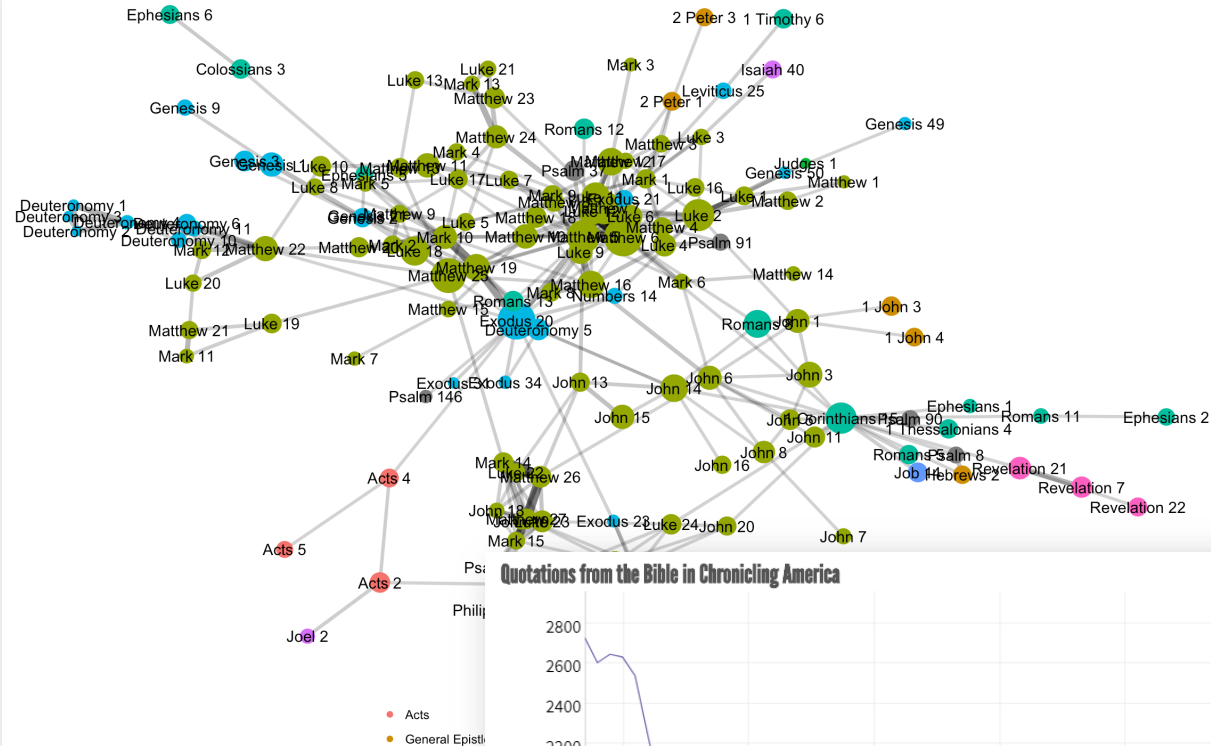
TEXT MINING: HISTORICAL NEWSPAPERS

AMERICA'S PUBLIC BIBLE: BIBLE QUOTATIONS IN U.S. NEWSPAPERS

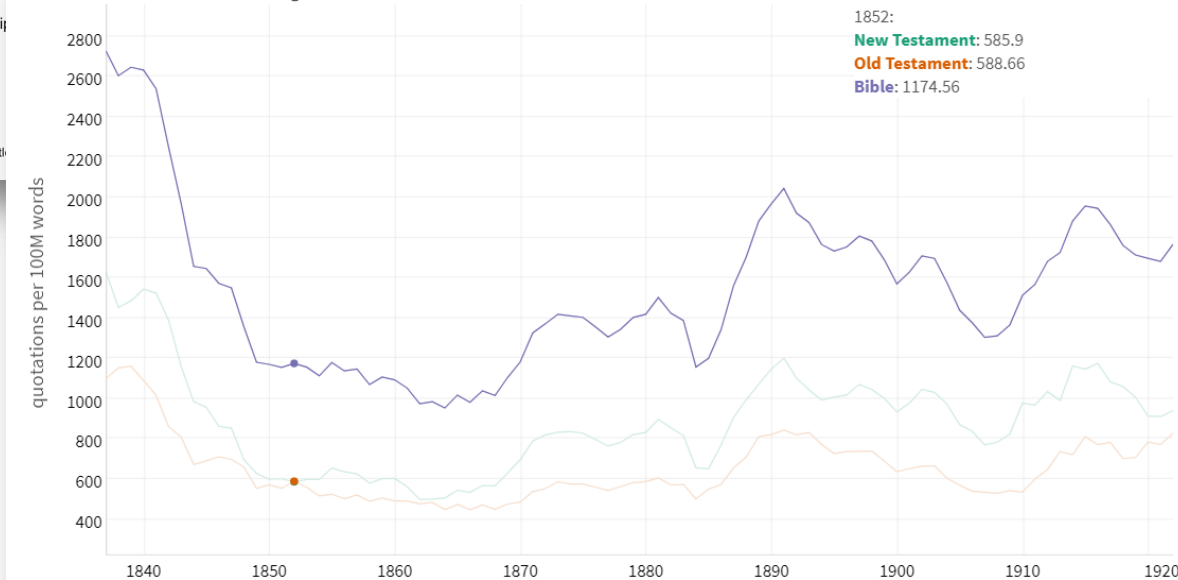
The project “tracks Biblical quotations in American newspapers to show how the Bible was used for cultural, social, religious, and political purposes, and how it was a contested yet common text.”

Professor Lincoln Mullen
History, George Mason University
<http://americaspublicbible.org/>

Biblical passages frequently quoted together



Quotations from the Bible in Chronically America

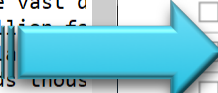


The general trends, however, tell us much less than the patterns for individual verses. Consider this handful of verses, each of which has a pattern that differs from the general trend.⁸



HOW TO MINE TEXTS

© 2014 Pearson Education, Inc. or its affiliate(s). All rights reserved. Pearson Education, Inc., 501 Boylston Street, Boston, MA 02116



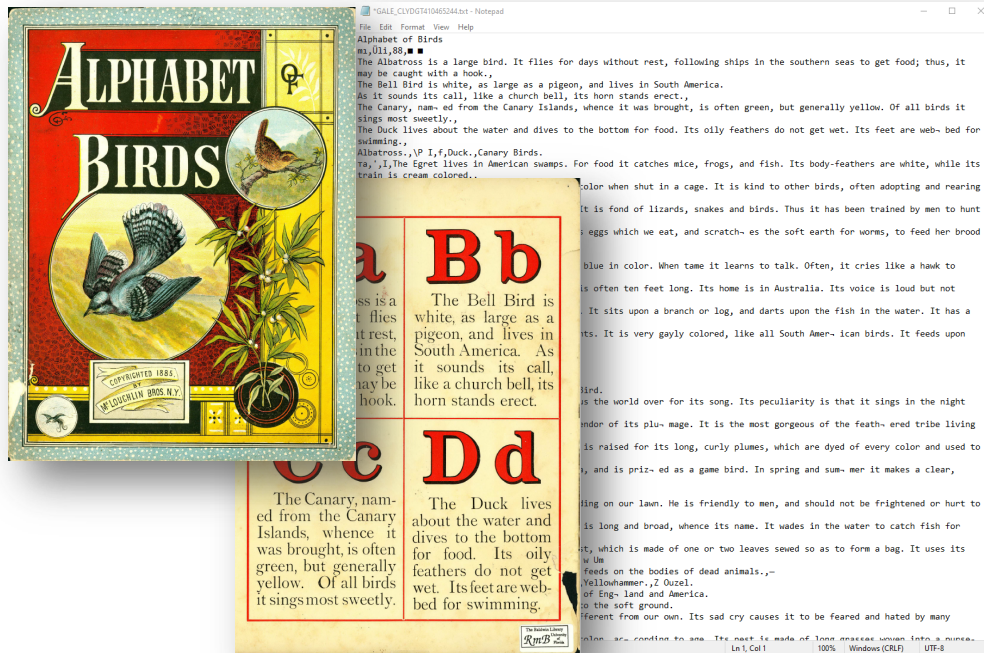
CHOOSING A TOOL OR METHOD

- Data questions:
 - What input/format does this tool require?
- Collaboration questions:
 - Is it easy to share in-progress material with others? (if you need to)
- Accessibility questions:
 - How does this tool work for people using assistive technology?
 - How does this tool work for people who are in locations with low bandwidth/internet access?
- Sustainability questions:
 - Can you download/export your material from this tool once you put it in?
 - Who made this tool? Who are their audiences? What is their revenue stream? (i.e., how long is this tool likely to last?)
 - What are they going to do with the data you put into their tool?

TYPES OF TEXT YOU CAN MINE

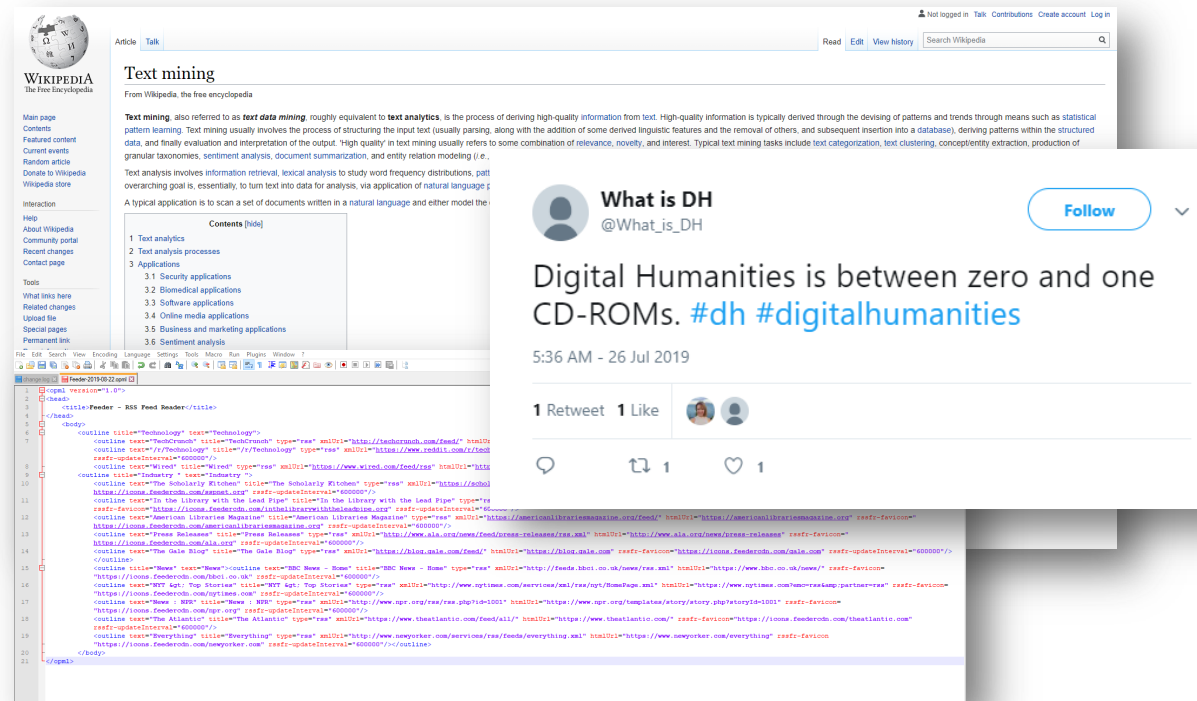
Digitized Texts

Physical documents that are digitized and processed using optical character recognition or manually keyed to create a digital facsimile.



Native Digital Texts

Texts created in a digital format for the purpose of being accessed on an electronic device.



PLACES TO GET TEXT

Digitized Texts

- [Internet Archive](#)
- [Project Gutenberg](#)
- [Google Books](#)
- [Hathi Trust](#)
- [JSTOR Data for Research](#)
- [PubMed Open Access Subset](#)
- [Open American National Corpus](#)

Native Digital Texts

- Email
- HTML
- RSS Feeds
- [Twitter](#)
- Wikipedia
- Data Liberation Front
- [New York Times API](#)

Dataset Repositories

- [Kaggle](#)
- [English-corpora.org](#) (BYU)
- [Data is Plural](#) (Jeremy Singer-Vine)
- [DH Toychest](#) (Alan Liu)

PLACES TO MINE TEXTS

Programming Languages

- [Python](#) (Text Cleaning & Statistical Analysis)
- [R](#) (Statistical Analysis & Visualization)
- Javascript (Visualization)
- GeoJSON (Geo-mapping)

Other helpful links:

- [TAPor](#)

Software Libraries

- [MALLET](#) (Topic Modeling)
- [spaCy](#) (Natural Language Processing)
- [Scrapy](#) (extracting the data from websites)
- [Transkribus](#)

Out-Of-The-Box

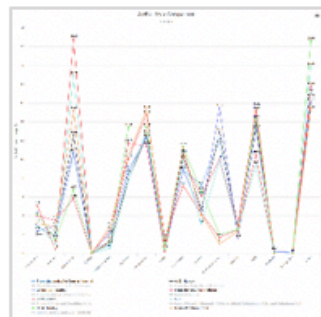
- [Voyant](#)
- [Lexos](#)
- [Juxta](#)
- [AntWord Profiler](#)
- [Textometrie \(TXM\)](#)
- [Textal](#)
- [Gephi](#)
- [Palladio](#)

ANALYSIS METHODS IN THE DSL

QUANTITATIVE ANALYSIS

Parts of Speech uses natural language processing of syntax to recognize, and tag parts of speech. In this implementation of Parts of Speech Tagger you may review how authors use of speech varies over time.

Open Source developer: [spaCy](#)



Named Entity Recognition (NER) recognizes and extracts proper and common nouns from documents using a Parts of Speech tagging method, and outputs them as lists of grouped by entity "type". Some "entity types" available for extraction are: people (including fictional), groups (nationalities, religious, or political), organizations, locations, products, works of art, dates, among others. This implementation uses **spaCy's** Named Entity Recognition model. Learn more [here](#)>



An **Ngram** is a term, or collocation of terms, found in your content set. You set the range or number of terms ('N') you wish to consider in your analysis. Then, the frequency of those Ngrams is counted and displayed for analysis.

Ngram examples:

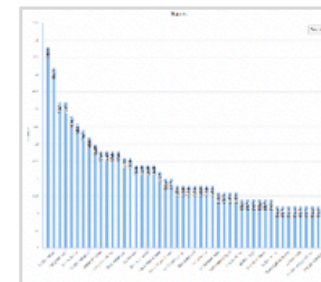
N=1: Unigram

"a", "the", "turtle", "frankenstein"

N=2: Bigram

"on the", "turtle dove", "mary shelley"

Learn more [here](#)>



QUALITATIVE ANALYSIS

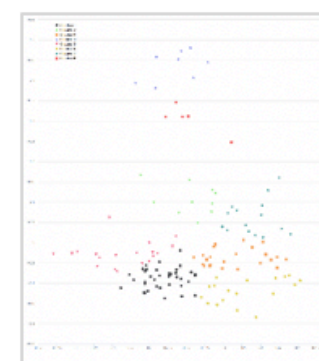
Sentiment analysis determines a tally of the positive or negative words within each document of a content set. It uses the **AFINN** [lexicon](#) (dictionary of words and their sentiment value) to compile sentiment scores for each phrase, which are then compiled to produce a document-level sentiment value. By establishing polarity within the texts (i.e. positive/negative word association), this tool can classify the documents in your content set between positive to negative sentiment.



Clustering analyzes the documents from a content set using statistical measures and methods to group them around particular features or attributes. This implementation of clustering leverages the **k-means** algorithm to create clusters of documents according to similar words contained within each document of your content set.

Open Source developer: [Scikit-Learn](#)

Learn more [here](#)>



Topic modelling allows users to analyze a large corpus of unstructured (OCR) text. A "topic," often referred to as a "bag of words," is a collection of terms that frequently co-occur in your collection of documents. **Mallet** uses Latent Dirichlet allocation (LDA) models to extract contextual clues in order to connect words with similar meanings, as well as differentiate between words that are spelled similarly but have differing meanings. Learn more [here](#)>

