

# FUNDAMENTALS OF TEXT MINING: CURATING, PREPARING, ANALYZING, AND VISUALIZING TEXTUAL DATA

**ELECTRONIC RESOURCES & LIBRARIES 2021**

SARAH KETCHLEY, PHD, UNIVERSITY OF WASHINGTON, GALE


LINDSEY GERVAIS, PHD, GALE

MARGARET WALIGORA, MLIS, GALE

Pronouns: she/her/hers



# IN THIS WORKSHOP WE WILL DISCUSS...

- What text mining is
  - What can you do with it
  - How you do it
- 

## OUR WORKSHOP HAS BEEN INFORMED BY...

- Our work in the digital humanities
- Input from our colleagues and students of digital humanities
- Scholarly articles and literature
- Work of other digital humanists
- Our participation in numerous workshops (Digital Humanities Summer Institute, conferences day-long symposia, hour-long workshops)
- Product development and user experience research

# COURSE OUTLINE

<http://bit.ly/tdm-workshop>

## Course Outline

March 15<sup>th</sup> – March 25<sup>th</sup>

Module 1: Introduction to Text Mining (Sarah)  
Overview of datasets (Maggie)

Module 2: Digital Literacies and Critical Thinking Skills (Lindsey)  
Ideating, Developing and Interpreting Research Questions –Lindsey)

Module 3: Text Cleaning with Lexos (Maggie)

Module 4: Text Analysis with Voyant (Sarah)





# INTRODUCTION TO TEXT MINING & YOUR DATA SETS



# Topology of Digital Humanities Projects

- Maps (GIS)
- Data visualization
- Text mining
- Digital editions of texts
- Virtual exhibits
- 3D imaging and reconstructions
- Creating of multimedia narratives
- Timelines





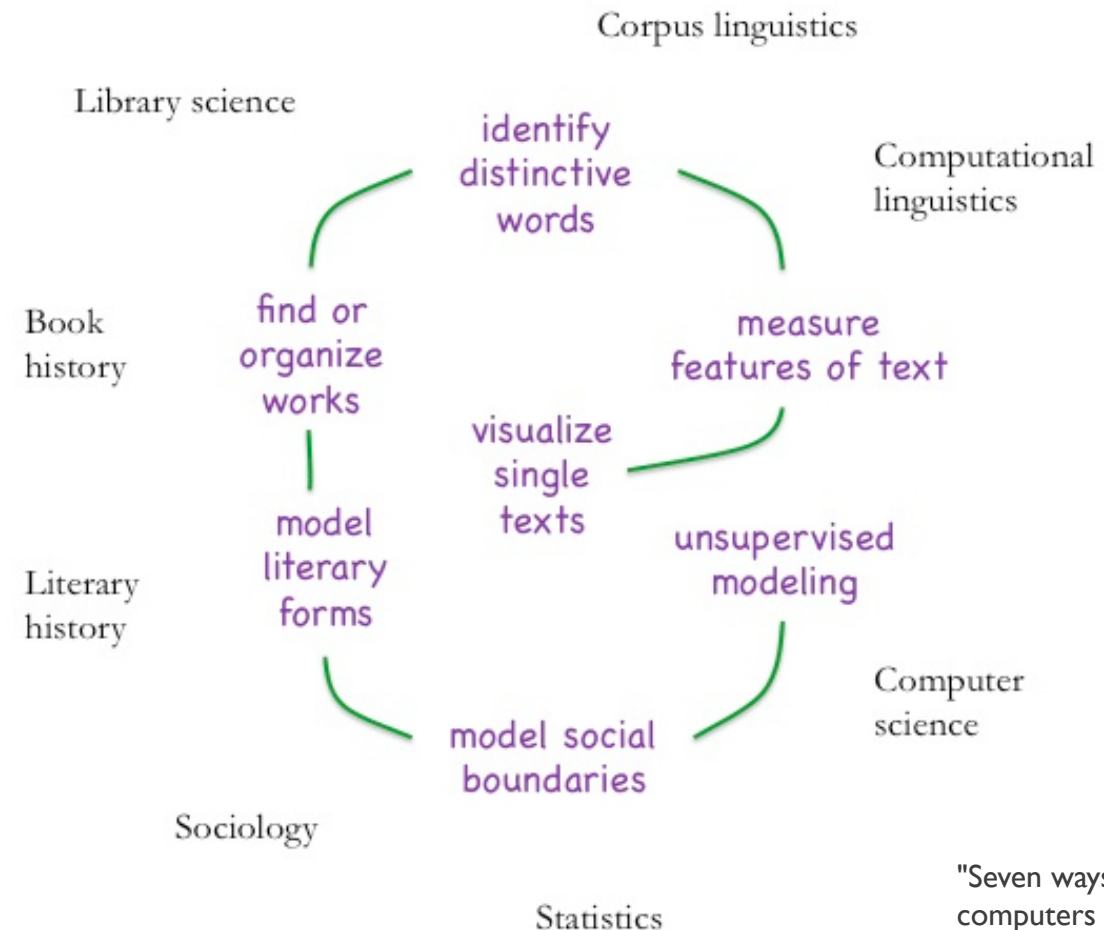
# WHAT IS TEXT MINING?

Text mining is a research practice that involves using **computational analysis** to discover information from **vast quantities** of digital, free-form, natural language, **unstructured text**.

# TEXT MINING IS INTERDISCIPLINARY

“Text mining is an **interdisciplinary** endeavor that also borrows freely from **corpus linguistics** and **computational linguistics**, as well as **social-scientific traditions** like social network analysis...Humanistic text mining seeks to frame questions that contribute meaningfully to existing traditions of humanistic inquiry.”

“Text-Mining the Humanities”  
Matthew L. Jockers & Ted Underwood



“Seven ways humanists are using computers to understand text”  
Ted Underwood

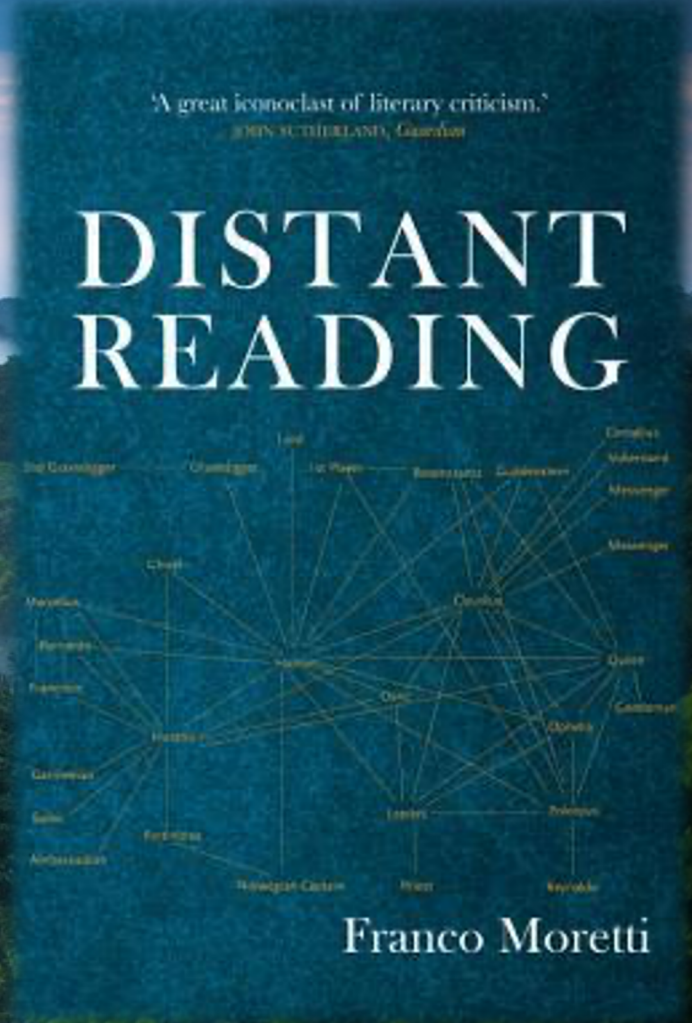
# SYNONYMS?

- Quantitative study of literature (many)
- Algorithmic criticism (Ramsay)
- Digital literary studies (Siemens/Schreibman et al)
- Computer-assisted reading / literary analysis / interpretation (Rockwell/Sinclair)
- Distant reading (Moretti)
- Macroanalysis (Jockers)
- CLS (Computational Literary Studies) (Da)
- ...

Slide adapted from:

DHSI 2019 Intro to Comp for Lit Crit @randaelka @DJWrisley





# DISTANT READING VS. CLOSE READING

## FRANCO MORETTI - *STANFORD LITERARY LAB*



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

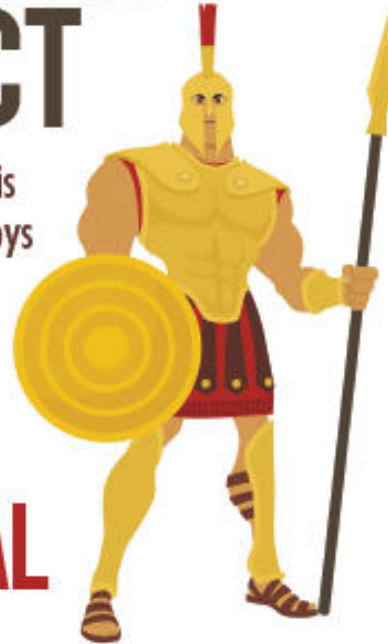
## SAMPLE USE CASES



# TEXT MINING: CULTURAL STUDIES

## THE ACHILLES EFFECT

What Pop Culture is  
Teaching Young Boys  
about Masculinity



CRYSTAL  
SMITH



Crystal Smith

HOME ABOUT BOOKS FREELANCE WRITING BLOG

SEPTEMBER 20, 2017 BY ADMIN

### Word Cloud: How Toy Ad Vocabulary Reinforces Gender Stereotypes

This post went viral shortly after I published it on March 28, 2011. There was so much response that I had to write a preamble in April to address some of the questions and concerns readers had. It was really a simple exercise but garnered many thoughtful and important questions. The original comments were lost but it looks like most are visible via the WayBackMachine.

# 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://viralttexts.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.

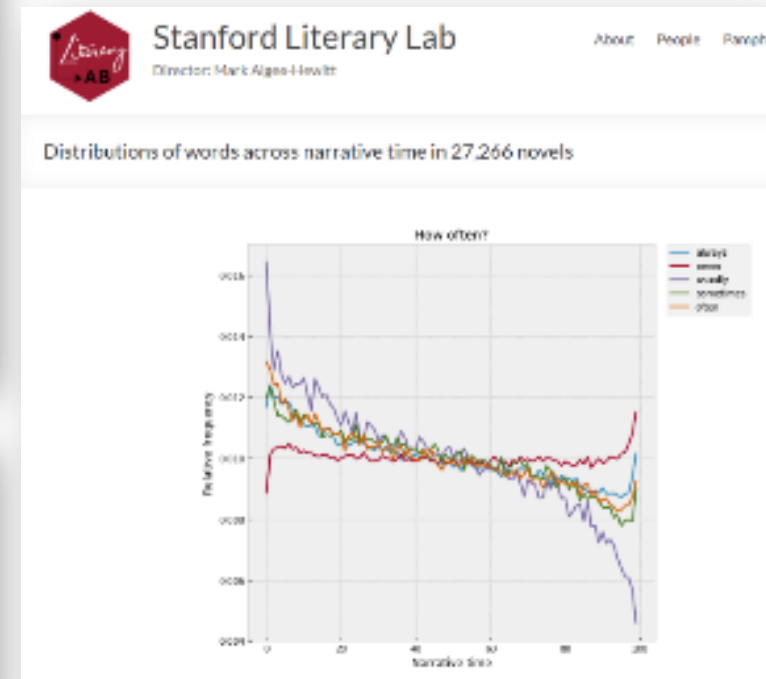
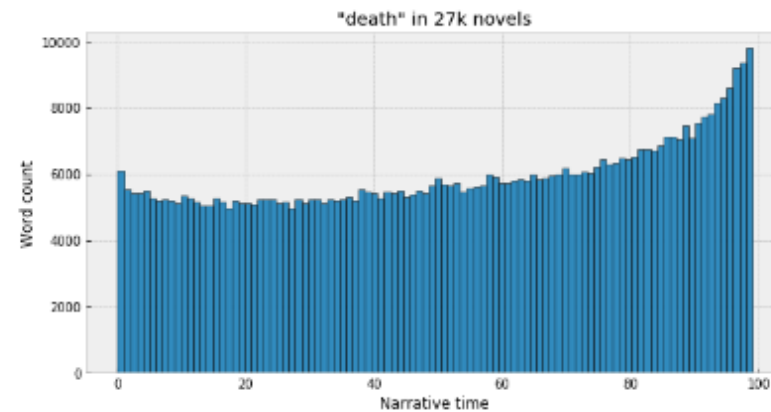
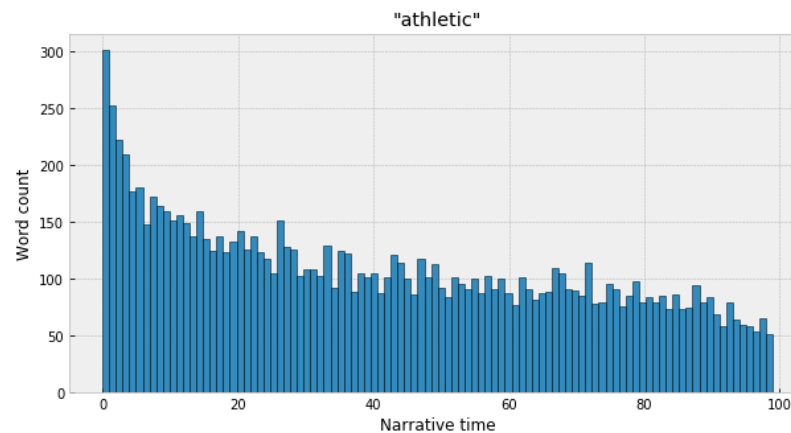
This exhibit is intended to hint at the breadth—and the oddities—of nineteenth-century reprinting that we have found thus far in the Viral Texts Project. If you peruse the page, you will find articles that link to our database, where you can browse versions that appeared in other newspapers, or related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

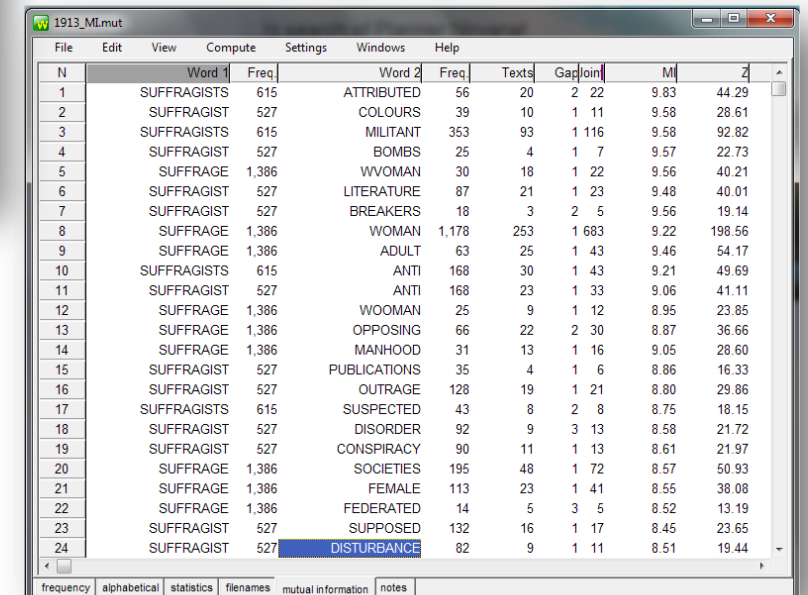
## THE LANGUAGE OF BRITISH SUFFRAGE IN THE PRESS

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 rowdyism 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



1913\_ML.mut

N	Word 1	Freq	Word 2	Freq	Texts	Gap	Join	M	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	WOMAN	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

frequency alphabetical statistics filenames mutual information notes



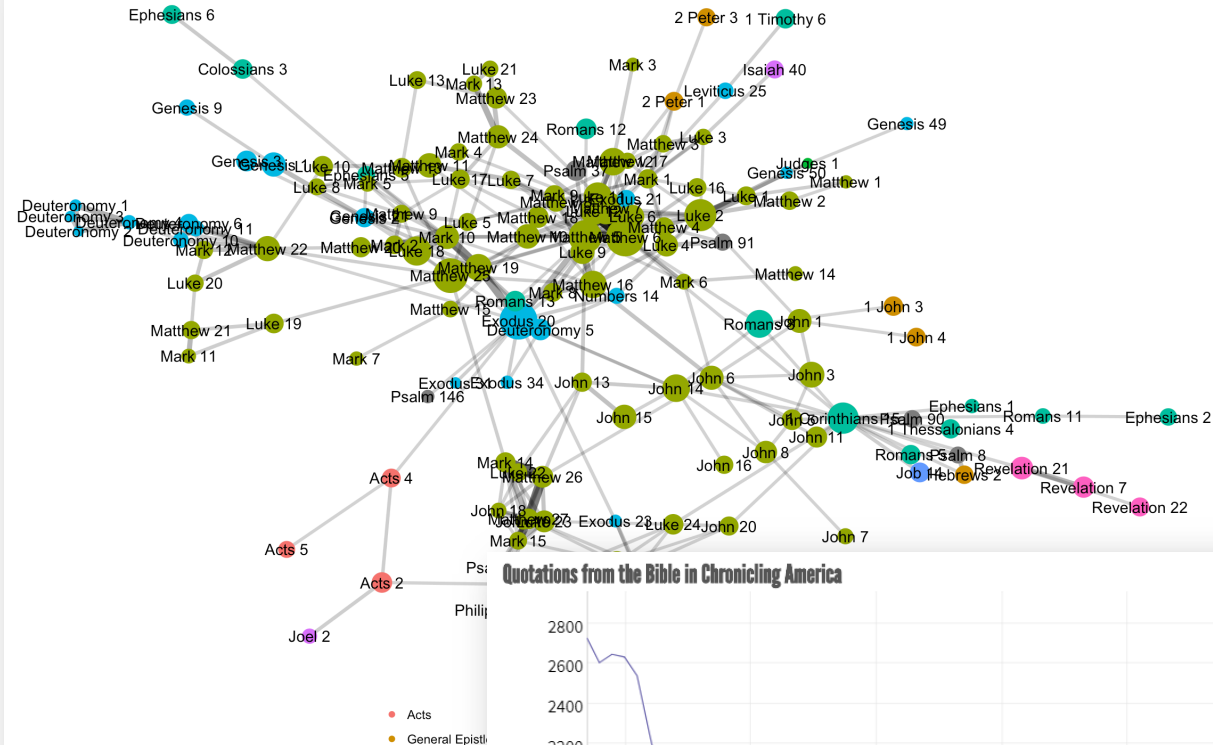
# TEXT MINING: HISTORICAL NEWSPAPERS

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

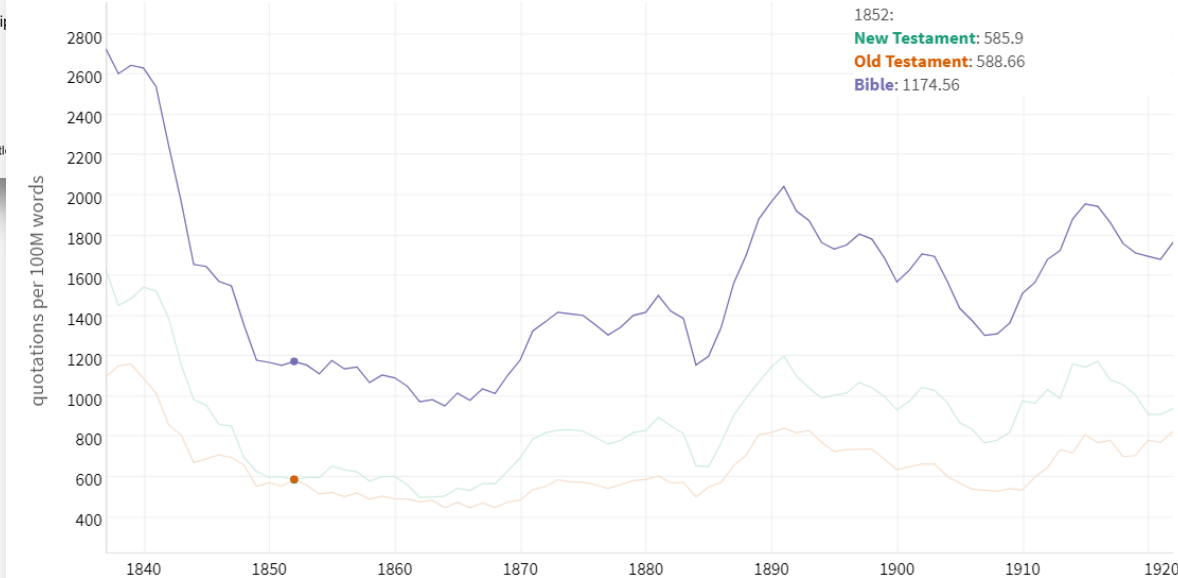
The project “tracks Biblical quotations in American newspapers to show how the Bible was used for cultural, social, religious, and political purposes, and how it was a contested yet common text.”

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

Biblical passages frequently quoted together



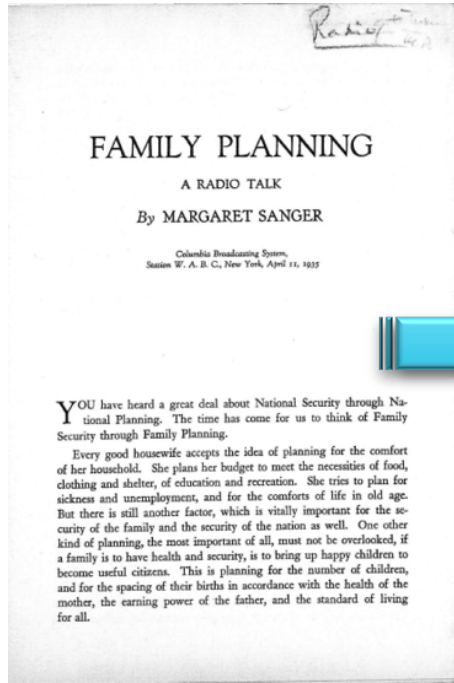
Quotations from the Bible in Chronicing America



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.<sup>8</sup>



# HOW TO MINE TEXTS



Digitized primary source

### Raw OCR text

```
File Edit Format View Help
t? ^ Δ FAMILY PLANNING A RADIO TALK By MARGARE
ower of the father, and the standard of living
In order to space the arrival of the children s
plan his family in accordance with how many ch
, in large part, why we have a high maternal n
unable to provide for the children already bor
a ft mother for the children already born and
hese countries?
r ■ J The Feder
to t' the thous
'~%л* - Senats
of the countr
'f Γ WSF -
;S m i ; "v l
í' - T. ■- -mr*
e put into prac
íkw7 «qWISCT' m
ncy is dangerou
Y -, Ý - ' ЧШ:
le to per- sons
```

Cleaned OCR text

#	Term	Count
1	birth	292
1	control	289
1	health	187
1	children	182
1	mrs	128
1	family	119
1	people	112
1	mothers	106
1	year	89
1	dr	87
1	life	86
1	need	83
1	medical	81
1	child	76

Statistical output



Visualization output



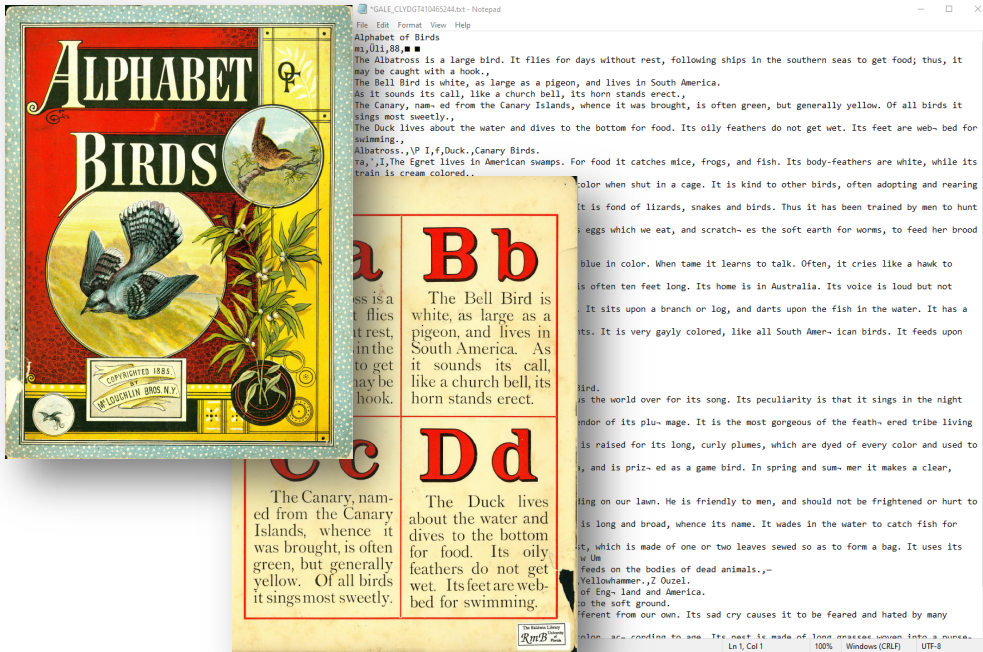
# CHOOSING A TOOL OR METHOD

- 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

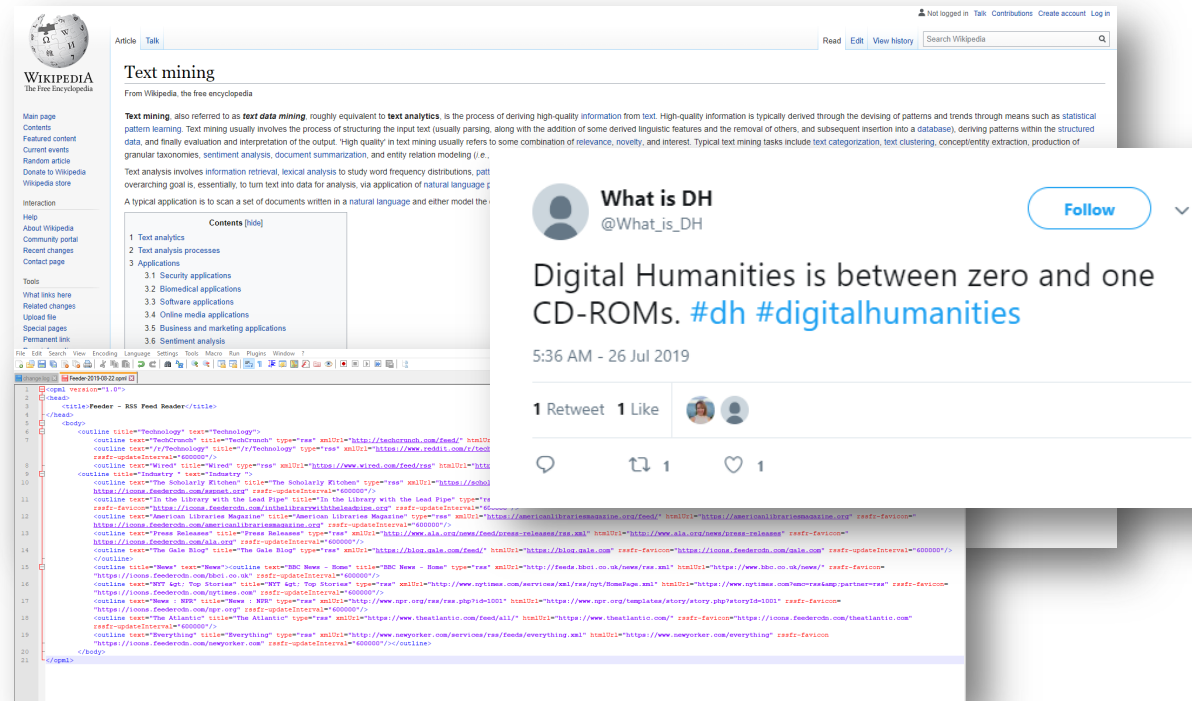
## Digitized Texts

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](#)

## AVAILABLE DATASETS FOR THIS WORKSHOP

1:

Adult British Fiction

2:

Watergate Scandal  
Newspaper Coverage

3:

Inaugural Presidential  
Speeches

4:

Feeding America:  
The Historic  
American Cookbook

5:

#1 Billboard Hits

6:

19<sup>th</sup> C. Sunday School  
Texts