

# Cultural Analytics

ENGL 64.05

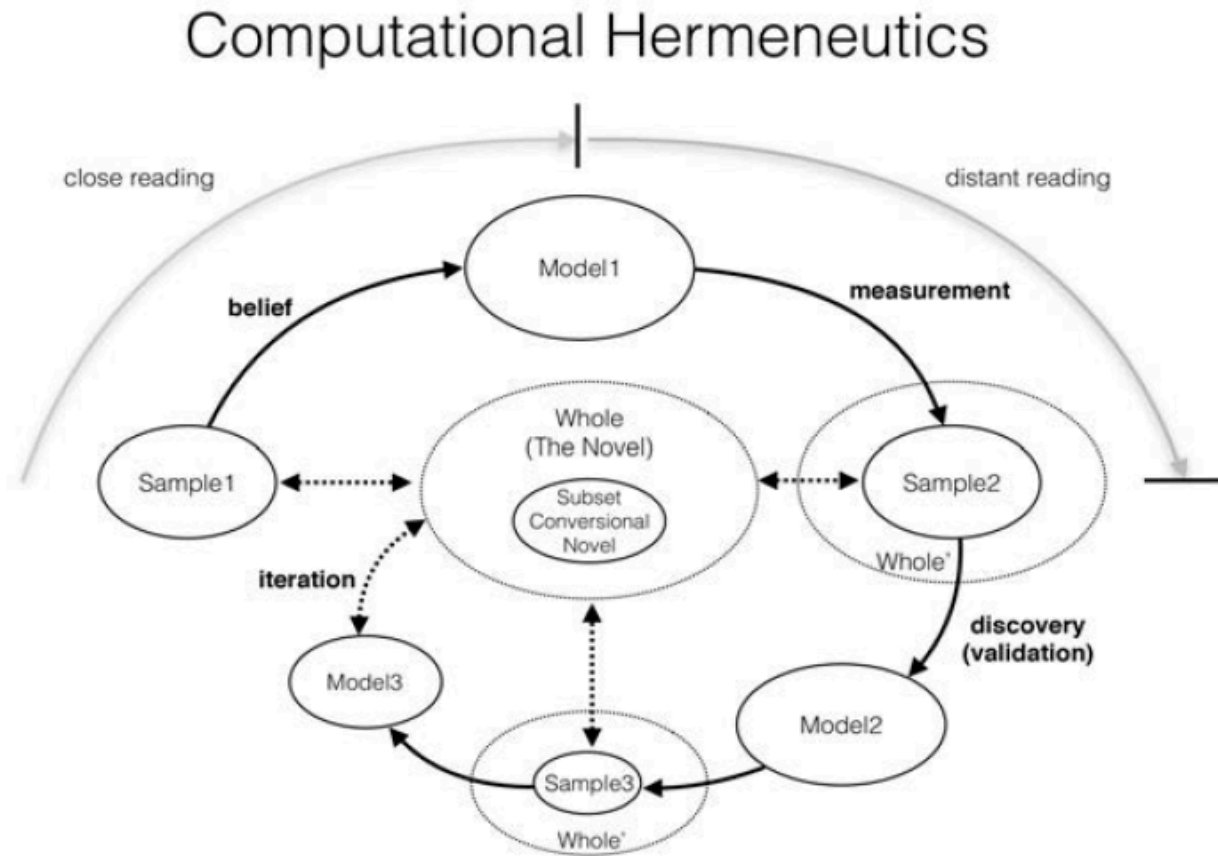
Fall 2019

Prof. James E. Dobson

October 28, 2019



# Andrew Piper's Model



# Archives and Sources

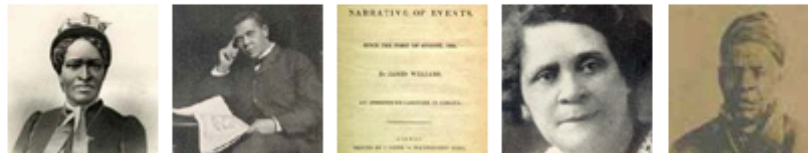
DOCUMENTING  
the *American South*

Google Custom Search



[About](#) | [Collections](#) | [Authors](#) | [Titles](#) | [Subjects](#) | [Geographic](#) | [K-12](#) | [Facebook](#) | [Buy DocSouth Books](#)

## NORTH AMERICAN SLAVE NARRATIVES



"North American Slave Narratives" collects books and articles that document the individual and collective story of African Americans struggling for freedom and human rights in the eighteenth, nineteenth, and early twentieth centuries. This collection includes all the existing autobiographical narratives of fugitive and former slaves published as broadsides, pamphlets, or books in English up to 1920. Also included are many of the biographies of fugitive and former slaves and some significant fictionalized slave narratives published in English before 1920.

- [Introduction](#)
- [Browse this Collection Alphabetically](#)
- [Browse Images by Subject](#)
- [Scholarly Bibliography of Slave and Ex-Slave Narratives compiled by William L. Andrews, E. Maynard Adams Professor of English](#)
- [Guide to Religious Content in Slave Narratives](#)
- [About this Collection](#)
- [Download full text of this Collection](#)

# Archives and Sources

[HOME](#)[ABOUT](#)[PROJECTS ▾](#)[PUBLICATIONS](#)[NEWS](#)[GET INVOLVED](#)[DATA SETS](#)[Q](#)

## NOVEL450

A collection of 450 novels in German, French, and English that span 1770 to 1930. Each language is represented by 150 novels with a roughly even distribution across time, length, and gender. The data can be downloaded here. And the metadata is here. Please cite:

Piper, Andrew (2016): txtlab Multilingual Novels. figshare.

<https://dx.doi.org/10.6084/m9.figshare.2062002.v3>

## CONTEMPORARY NOVELS

A collection of 1,211 novels published between 2000-2015. They are categorized by the following 6 groups: Bestsellers (BS), Prizewinners (PW), Novels reviewed in the New York Times (NYT), Mysteries (MYST), Romances (ROM), and Science Fiction (SCIFI). Metadata is available here.

Please cite: Andrew Piper and Eva Portelance, “How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading,” Post45 (2016).

## RACE AND FILM

This data set contains character dialogue from 780 Hollywood movies produced between 1970 and 2014. Characters have been labeled by their racial and ethnic identity using IMDB. The data set is available here.

Please cite: Vicky Svaikovsky, Anne Meisner, Eve Kraicer, and Matthew Sims, “Racial Lines: Race Ethnicity and Dialogue in 780 Hollywood Films, 1970-2014.”

## 20C POETRY

A table of derived word counts from a collection of 75,297 English-language poems. A table with the top 20K words is located here and three tables of POS, Hypernyms, and word counts is located here.



# Cultural Analytics



From 1830 until the 1890s, already free and once captive Black people came together in state and national political meetings called "Colored Conventions." Before the War, they strategized about how to achieve educational, labor and legal justice at a moment when Black rights were constricting nationally and locally. After the War, their numbers swelled as they continued to mobilize to ensure that Black citizenship rights and safety, Black labor rights and land, Black education and institutions would be protected under the law.



# Archives and Sources



search for books

- Browse Catalog
- Bookshelves
- Main Page
- Categories
- Contact Info

Project Gutenberg appreciates your donation!

[Donate](#)

- Why donate?

in other languages

- Português
- Deutsch
- Français

hosted by 

## Free eBooks - Project Gutenberg



[Book search](#) · [Book categories](#) · [Browse catalog](#) · [Mobile site](#) · [Report errors](#) · [Terms of use](#)

## Some of the Latest eBooks



## Welcome

**Project Gutenberg** is a library of over 60,000 free eBooks. Choose among free epub and Kindle eBooks, download them or read them online. You will find the world's great literature here, with focus on older works for which U.S. copyright has expired. Thousands of volunteers digitized and diligently proofread the eBooks, for enjoyment and education.

**No fee or registration!** Everything from Project Gutenberg is gratis, libre, and completely without cost to readers. If you find Project Gutenberg useful, please consider a small [donation](#), to help Project Gutenberg digitize more books, maintain our online presence, and improve Project Gutenberg programs and offerings. Other ways to help include [digitizing](#), [proofreading and formatting](#) , [recording audio books](#) , or [reporting errors](#).



[Project Gutenberg Mobile Site](#)

# Archives and Sources



## eBooks and Texts

### Internet Archive

The Internet Archive offers over **20,000,000** freely downloadable books and texts. There is also a collection of **1 million modern eBooks** that may be

MORE

[Share](#)  
[Favorite](#)  
[RSS](#)  
[Play All](#)

[ABOUT](#)
[CONTACT](#)
[BLOG](#)
[PROJECTS](#)
[HELP](#)
[DONATE](#)
[JOBS](#)
[VOLUNTEER](#)
[PEOPLE](#)

[ABOUT](#)
[COLLECTION](#)
[FORUM](#)

DESCRIPTION

The



Internet Archive offers over **20,000,000** freely downloadable books and texts. There is also a collection of **1 million modern eBooks** that may be borrowed by anyone with a free archive.org account.

[Borrow a Book](#)

Created on  
**December 16 2004**



**Jeff Kaplan**  
Archivist

ADDITIONAL CONTRIBUTORS



**AnnaN**  
Member



**binderc**  
Member



**Diana Hamilton**  
Archivist



**jordonz**  
Archivist



**ARossi**  
Archivist



# Archives and Sources

[HTRC Analytics](#) [Algorithms](#) [Data Capsules](#) [Worksets](#) [Datasets](#) [Explore](#) [Help](#) [About](#) [Sign In](#) [Sign Up](#)

## Datasets

**Downloadable, non-consumptive book data.**

HTRC releases research datasets to facilitate text analysis using the HathiTrust Digital Library. While copyright-protected texts are not available for download from HathiTrust, fruitful research can still be performed on the basis of non-consumptive analysis of features extracted from full text. These features include volume-level metadata, page-level metadata, part-of-speech-tagged tokens, and token counts. Additionally, HTRC has partnered with advanced researchers to release a derived dataset, Word Frequencies in English-Language Literature, 1700-1922.

### HTRC Extracted Features Dataset

Page-level features from 15.7 million volumes [v.1.5]

Description	Contents						
<p>The HTRC Extracted Features Dataset v.1.5 is comprised of page-level features for 15.7 volumes in the HathiTrust Digital Library. This version contains non-consumptive features for both public-domain and in-copyright books.</p> <p>Features include part-of-speech tagged term token counts, header/footer identification, marginal character counts, and much more.</p> <p>A full explanation of the dataset's features, motivation, and creation is available at the <a href="#">EF Dataset documentation page</a></p>	<table><tbody><tr><td># of volumes</td><td>15,722,079</td></tr><tr><td>In-copyright</td><td>9,914,509</td></tr><tr><td>Public domain</td><td>5,807,570</td></tr></tbody></table>	# of volumes	15,722,079	In-copyright	9,914,509	Public domain	5,807,570
# of volumes	15,722,079						
In-copyright	9,914,509						
Public domain	5,807,570						
<p><b>Download the data</b></p> <p>All 15.7 million files as well as custom subsets of the EF data are accessible using <code>rsync</code>, as described in the <a href="#">documentation</a>.</p> <p>A sample is available for download through your browser – <a href="#">sample.zip</a> – as well as thematic collections: <a href="#">DocSouth</a> (87 volumes), <a href="#">EEBO</a> (355 volumes), <a href="#">ECCO</a> (505 volumes).</p>	<table><tbody><tr><td># of pages</td><td>5,787,519,444</td></tr><tr><td># of tokens</td><td>2,449,739,213,773</td></tr></tbody></table> <p><b>Resources</b></p> <p>See how scholars are <a href="#">using the EF Dataset</a></p> <p><a href="#">Full documentation</a></p>	# of pages	5,787,519,444	# of tokens	2,449,739,213,773		
# of pages	5,787,519,444						
# of tokens	2,449,739,213,773						



# “Text Mining the Humanities”

- Ted Underwood and Matthew Jockers provide an overview of existing text mining techniques for the humanities.
- They place scale at the center of cultural analytics: “Computers are very good at this task, and for some types of questions computational keyword searching is all that is warranted. But what of other questions, questions of scale that have not been asked until quite recently?” (291).
- But do we need to operate at large scale? What is the “scale” and the object of large scale?

# Text Mining: NLP/BoW

- Natural Language Processing (NLP) and bag-of-words (BoW) serve to split techniques of representation of text for Jockers & Underwood.
- But this split obscures other uses of the “raw” input text.
- NLP comes from the social sciences—computational linguistics and information science, primarily.
- NLP techniques include part-of-speech tagging, sentence and document structure.
- Not necessarily literary features. We might be interested in other formal features:
  - poetry: structure of poems
  - fiction: characters and narrative features
  - book-level: paratexts, notes, organization of volume, images, page layout, metadata

# Text Mining: Intrinsic

- Text mining as a mode of intrinsic interpretation means that the only source of meaning is internal to the text.
- Topic modeling, collocations, TTR, etc are intrinsic operations.
- Most, but not all, unsupervised models could be considered intrinsic.
- We can mix modes of analysis using data producing / feature extraction techniques that assume intrinsic analysis (BoW vector models) and then combine with extrinsic methods (dictionaries, labeling of datasets for machine learning).

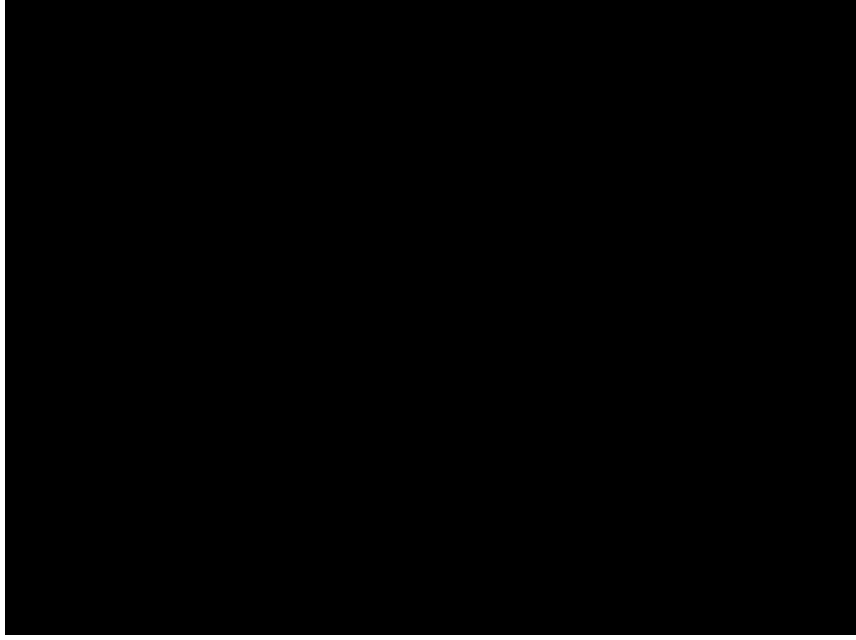
# Text Mining: Extrinsic

- When we use an external dataset, archive, dictionary, etc., we are engaged in an extrinsic mode of interpretation.
- We can use these dictionaries to provide knowledge and context: colors, characters, places, clusters of terms (abstract/hard), sentiment, etc.
- These bring knowledge from outside the text to the text.

# Kurt Vonnegut The Shape of Stories



# Cultural Analytics



# Next Class

Sentiment Mining as Proxy for Plot  
and Historicizing Sentiment