Text Analysis 101: Voyant Tools

Erica Y. Hayes
Digital Scholarship Librarian
erica.hayes@villanova.edu
Falvey Brown Bag Lunch

Slides & Tutorial:
What is text analysis? Why do text analysis?

“Text Analysis is the search for and discovery of patterns and trends in a corpus of texts. The analysis of those patterns and trends can help researchers uncover previously unseen characteristics of a specific corpus.”

-M. Dabek (2014)
Franco Moretti and distant reading

Figure: Franco Moretti Verso Books
Close reading or microanalysis

“Essentially, close reading means reading to uncover layers of meaning that lead to deep comprehension.” - Nancy Boyles (2013)
Distant reading or macroanalysis

“Distant reading aims to generate an abstract view by shifting from observing textual content to visualizing global features of multiple text(s)” - Stefan Jänicke (2016)
Benefits of both forms of reading and analysis

“It is the exact interplay between the macro and micro scale that promises a new, enhanced, and perhaps even better understanding of the literary record. The two approaches work in tandem and inform each other. Human interpretation of the ‘data,’ whether it be mined at the macro or micro level, remains essential. While the methods of enquiry, of evidence gathering, are different, they are not antithetical, and they share the same ultimate goal of informing our understanding of the literary record, be it writ large or small”

-Matthew Jockers “On Distant Reading and Macroanalysis” (2011)
Guiding questions

1. For texts with which I am already familiar, how can computers help me identify and study interesting things I had not noticed before, or things I had noticed but did not have reasonable means to pursue?

2. How can computers help me identify and understand texts with which I am not familiar with or which I cannot reasonably read?
Getting started: terms, workflows, and challenges

- **Corpus**: a collection of texts.
  - Consider the size of your corpus and your inclusion and exclusion criteria
  - Your corpus must be machine readable for computers. PDF files might need to be OCR’d. Texts may need to be converted to txt or rtf files.
  - Text editors like Notepad++ and Atom can be used for conversion.

- **Data cleaning**: removing information not relevant to your analysis.
  - Examples: chapter numbers and titles, copyright information, headers and footers, page numbers.
  - Programming Languages and tools such as R, Python, and Regular Expressions are necessary with a large corpus.

- **Stopwords**: words you filter out of your analysis
  - Standard lists exist (e.g. Buckley-Salton Stoplist)
  - Tools with built in stopwords. Voyant has a stoplist included
  - Customization of the stoplist is often necessary
Choosing a corpus: size and sources

Sample text data sources:
- Project Gutenberg
- HathiTrust Digital Library
- JSTOR Data for Research
- Corpus of Historical American English
- Corpus of Contemporary American English
- The Royal Society of London Corpus
- Linguistic Data Consortium Corpus

“Quantitative analysis tends to require context before it becomes meaningful. It doesn’t mean much to say that the word “motion” is common in Wordsworth, for instance, until we know whether “motion” is more common in his works than in other nineteenth-century poets. So yes, text-mining can provide clues that lead to real insights about a single author or text. But it’s likely that you’ll need a collection of several hundred volumes, for comparison, before those clues become legible” - Ted Underwood (2012)
Types of text analysis summaries and methods

- **Word Frequency** (How many times does “cedar” show up in the text?)
- **Collocation** (What words show up near “red cedar” in the text?)
- **Concordance** (What are the contexts where cedars are mentioned?)
- **Sentiment Analysis** (Does the author write positively or negatively about cedars?)
- **Topic Analysis** (What does this book discuss? Are cedar trees the main topic of discussion?)
- **N-grams** (Does cedar show up with the word wood a lot? Do these two-, three-, etc. words appear commonly together?)
Voyant Tools is an open-source, web-based application for performing text analysis developed by Stefan Sinclair & Geoffrey Rockwell.

Stefan Sinclair’s & Geoffrey Rockwell’s Voyant Tools
Voyant Tools: Pros & Cons

Positives:

● Great for beginners and available via web browser
● No computer programming experience required
● Allows for quick reading and analysis of texts
● Comparative options to analyze texts in tandem
● A variety of different visualization and text analysis tools

Negatives:

● Web application frequently breaks down/fails to load
● Slow loading time with large documents (web)
● Interface limits number of texts you can easily see altogether and analyze
● The software is not case-sensitive, so it is difficult to differentiate between lower-case and capitalized versions of a word.
Other Text Analysis Tools

- NVivo (proprietary)
- RapidMiner (proprietary)
- AntConc (open-source)
- R programming -- tidyr, topicmodels, janeaustenr, gutenburgr packages (open source)
- Python programming -- Natural Language Processing Toolkit package (open source)
- Serendip--Python (open source)
- InPho Explorer--Python (open source)
Workshop materials: https://bit.ly/3bTri4q

- Pardise Lost Corpus
- Tutorial: Voyant Tools
- Text Analysis 101 Slides
Schedule a consultation or instruction session

Falvey Memorial Library

Email: erica.hayes@villanova.edu

Twitter: @ericayhayes