Text Analysis

Introduction to Voyant & Topic Modeling Tool

Dr. Sierra Eckert

Princeton University

sceckert@princeton.edu

Text Analysis Tools

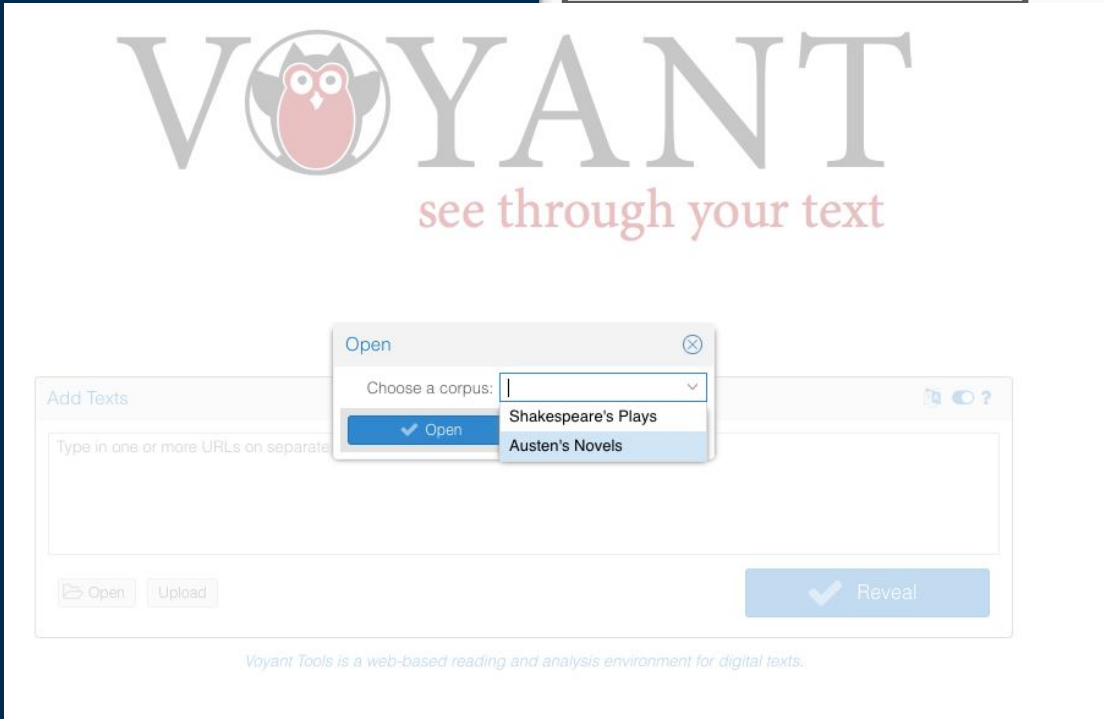
What are some text-analysis tools used for “distant reading” today?

1. **Voyant - a basic dashboard for text analysis**
2. Topic modeling browser

Let's try some text analysis!

Please go to:

voyant-tools.org

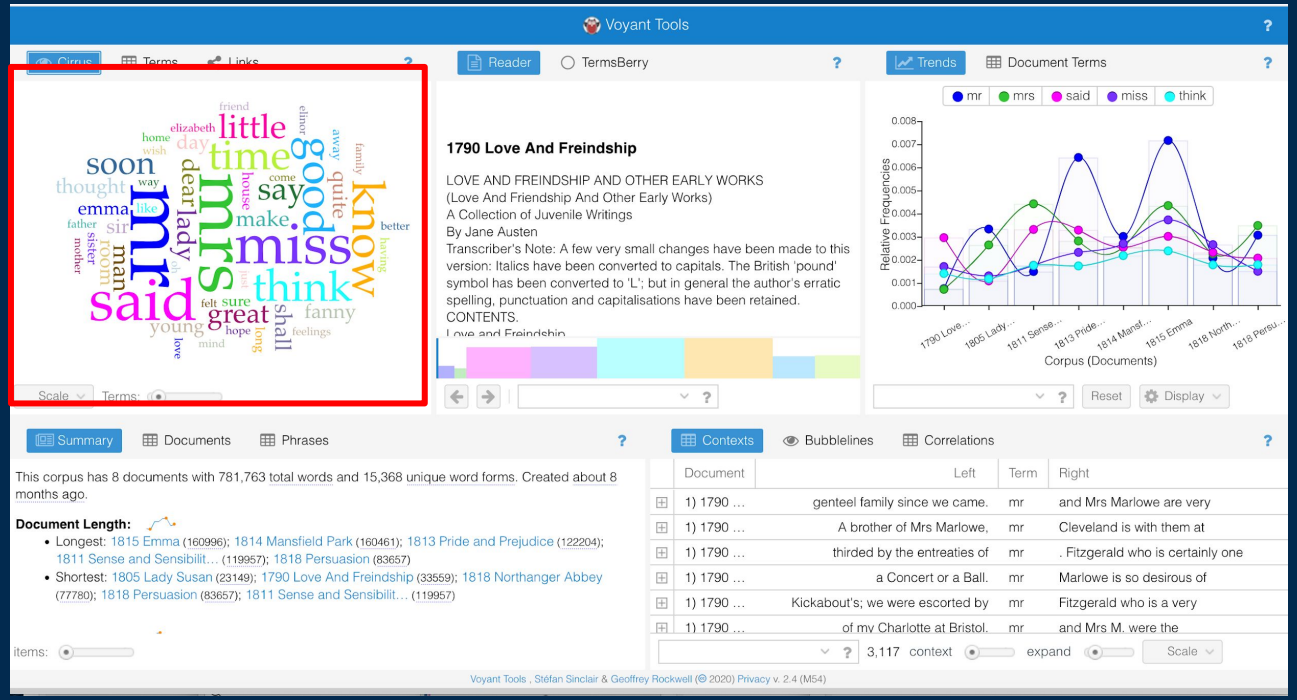


From the **Open** menu, choose the corpus **Austen's Novels**.

Then press **Reveal**.

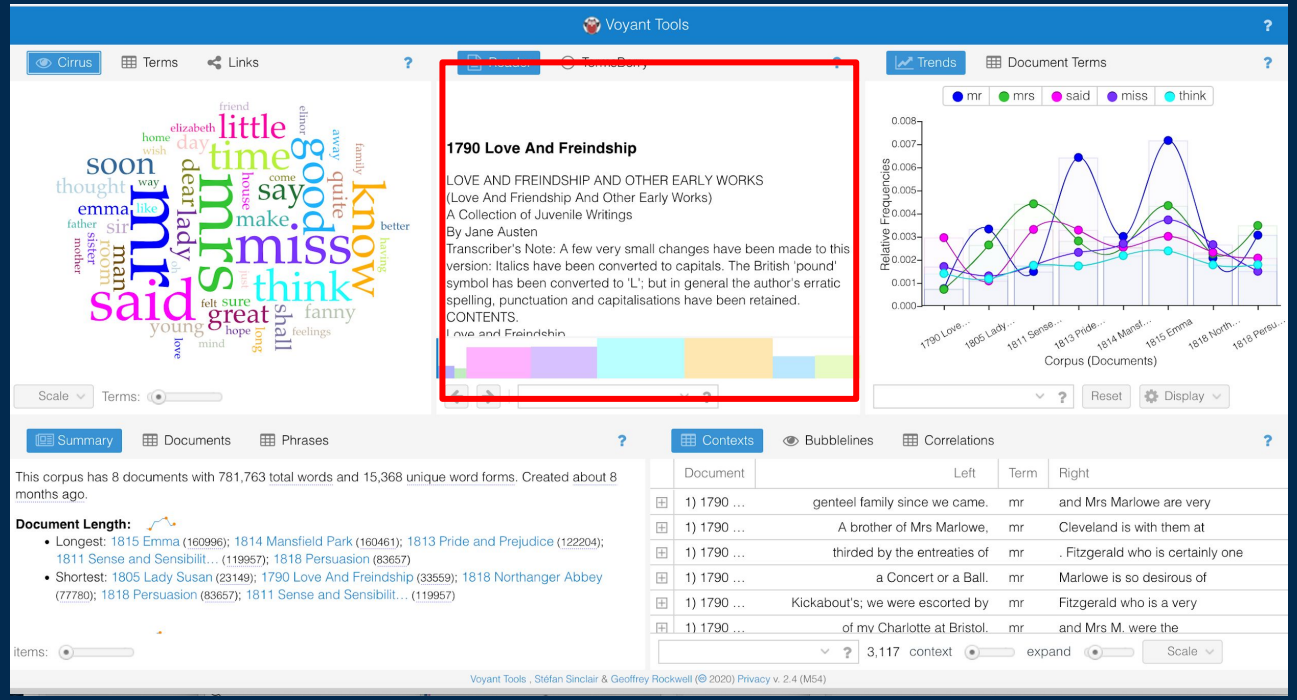
On the upper left is a WORD CLOUD.
This is a list of words sized by their frequency

Voyant Tools



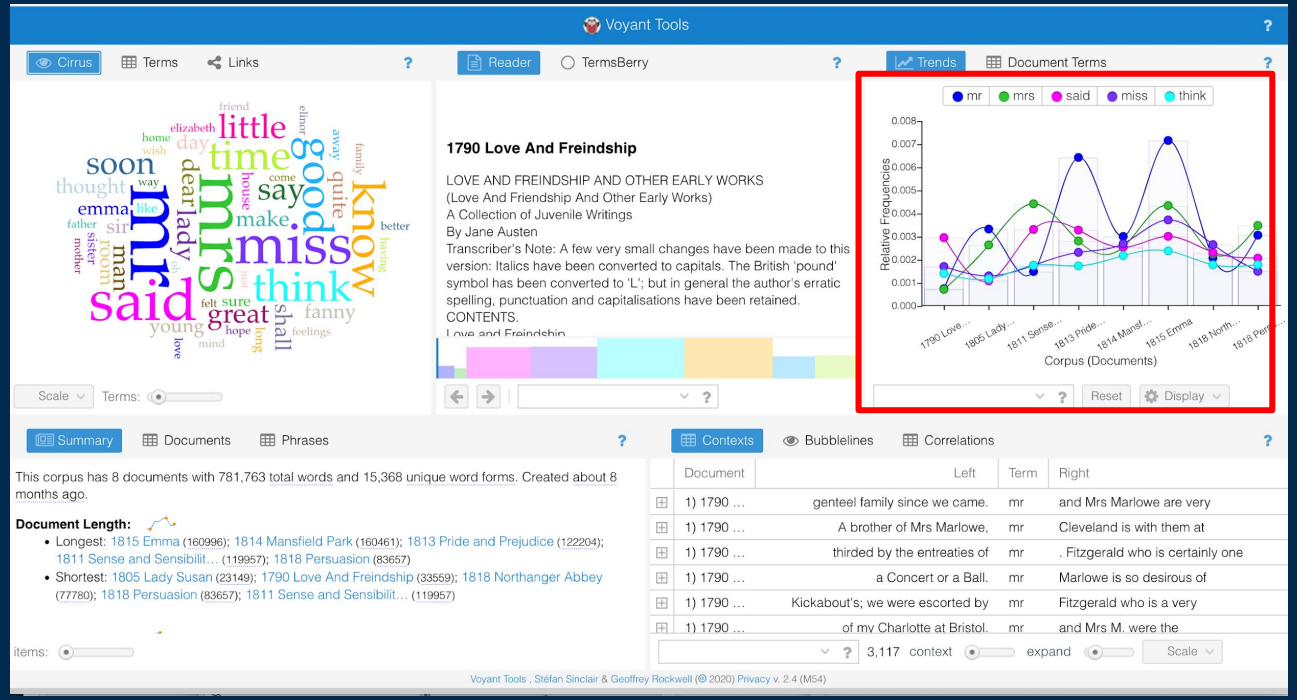
Voyant Tools

In the upper center square is the **TEXT of your CORPUS**.
This gives you the full text that you're analyzing, listed in order that they are labeled (in this case, by date). If you hover over a word, it will tell you how many times it appears in the collection.
Take a look at the text in the box. What do you notice? What kinds of problems might it pose for our analysis?



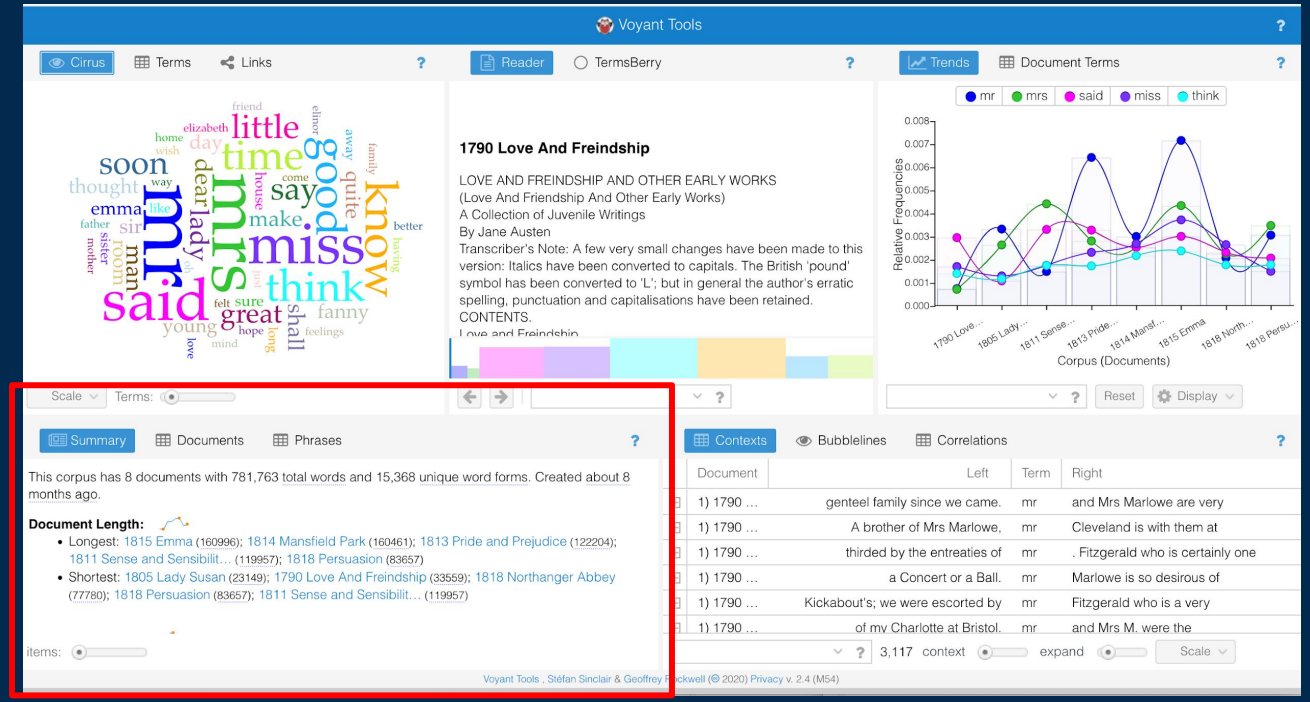
Voyant Tools

In the upper right is the WORD FREQUENCY visualization.
This shows you how frequently a word appears in each of the documents in the corpus.
Try typing in “very” in the search bar at the bottom of this box. What do you notice?



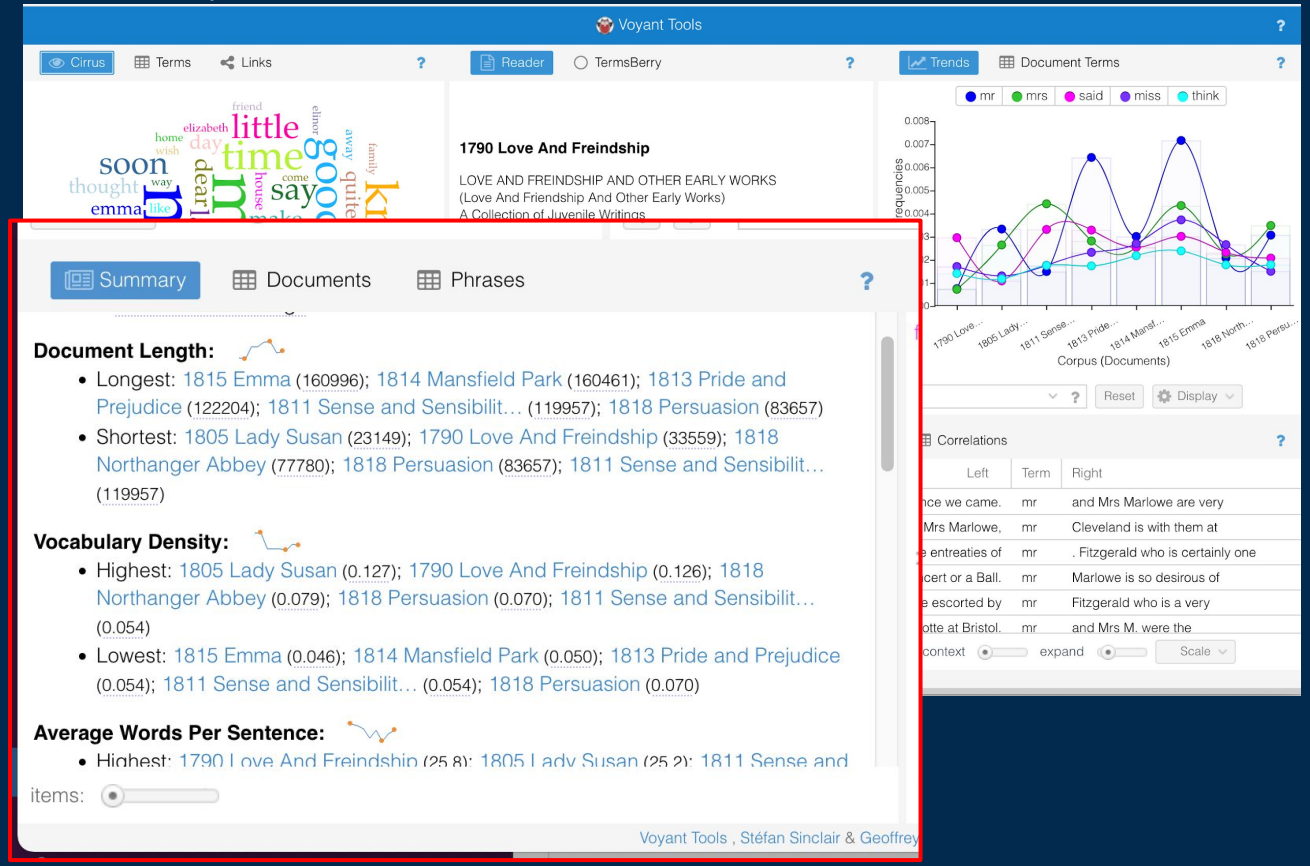
In the bottom left are STATISTICS about your corpus.
Scrolling down in the “Summary” view gives a longer list of some of most distinctive words in each text, the average document and sentence length. “Phrases” allows you to sort by short phrases.

Voyant Tools



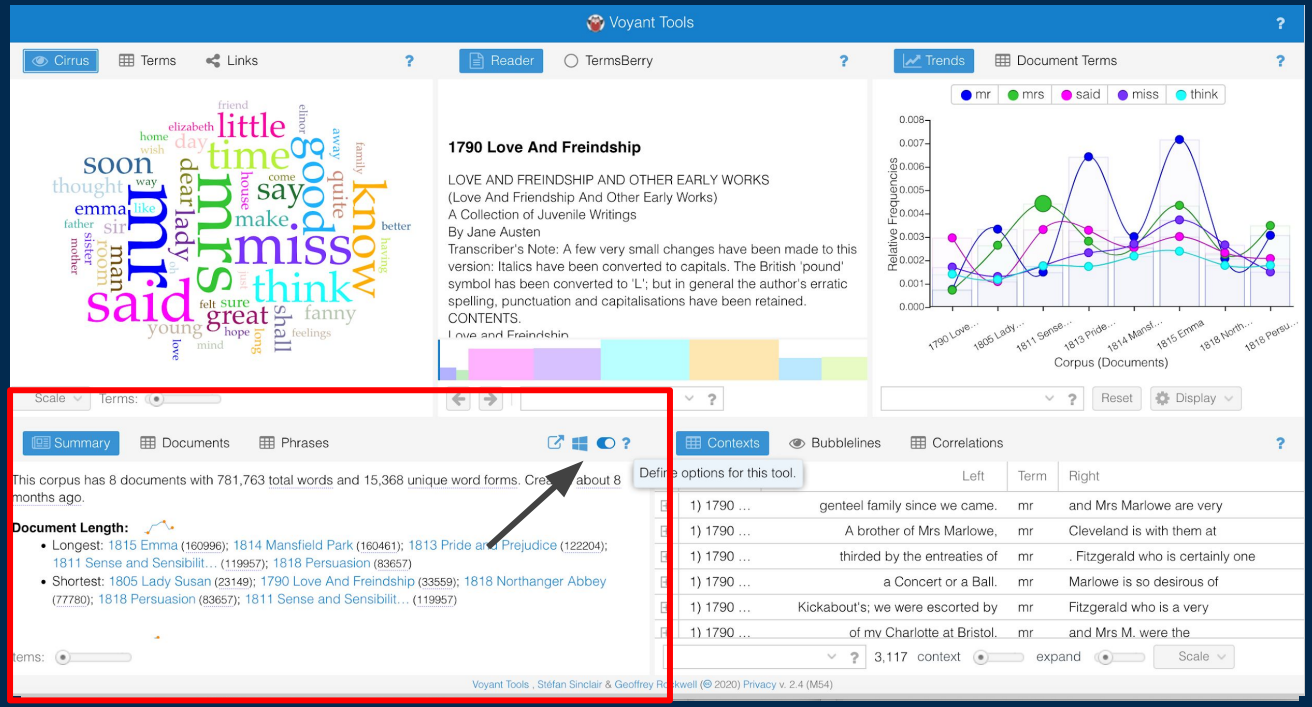
With this same “Summary” box, you’ll have descriptive statistics on the text or corpus (collection of texts) that you’re working with. These include “Document Length”, “Vocabulary Density”, “Average Words Per Sentence” You’ll also see “Most Frequent Words” in the corpus and Most Distinctive Words in each document. If you want to know what exactly these are measuring, click on the question mark in this boxe’s upper right corner.

Voyant Tools



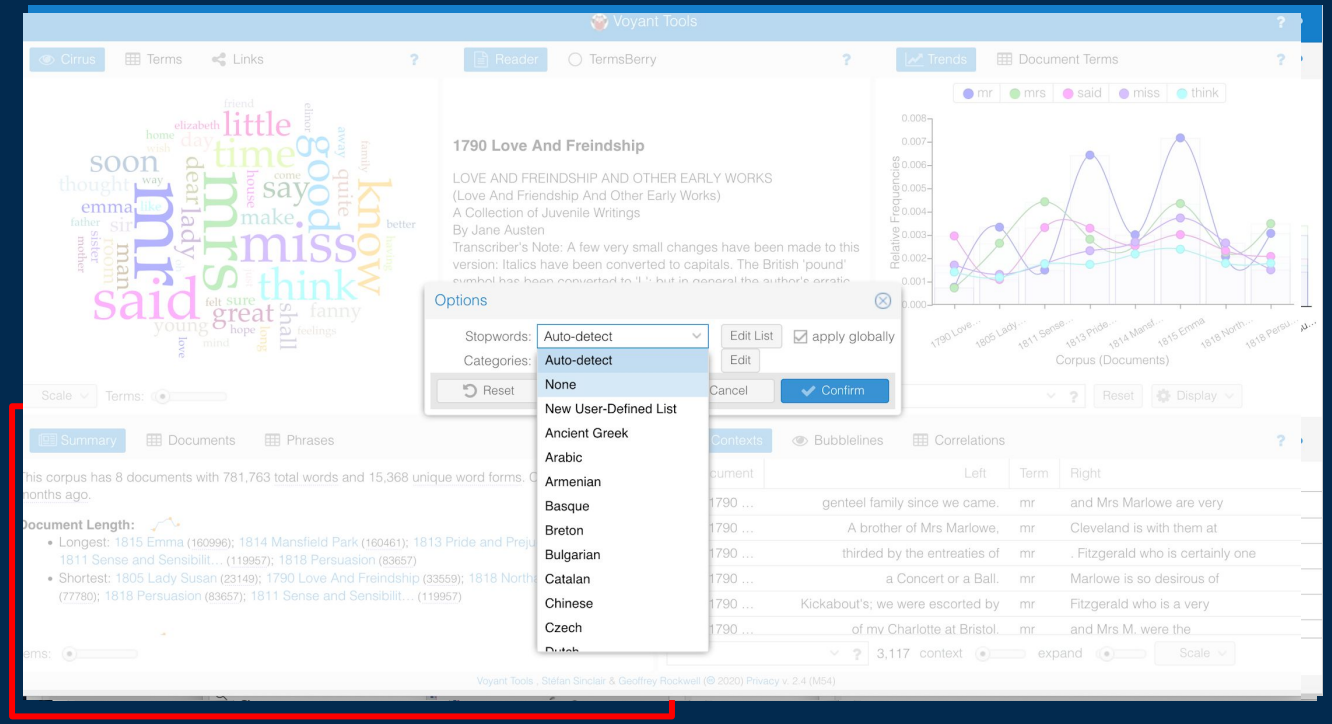
This box also allows you to control the global filters for the toolset.
Hover over the upper right corner of this box and click on the “options” toggle

Voyant Tools



In the options box, click on “None.” Then click Confirm. What happened?
What you just removed was a “stop words” list.
Click on options again, and click on “auto-detect.” Then click on the Edit list button. What do you notice about the words?
When would you want to filter certain words out? When *wouldn’t* you want to remove them? What are the implications?

Voyant Tools



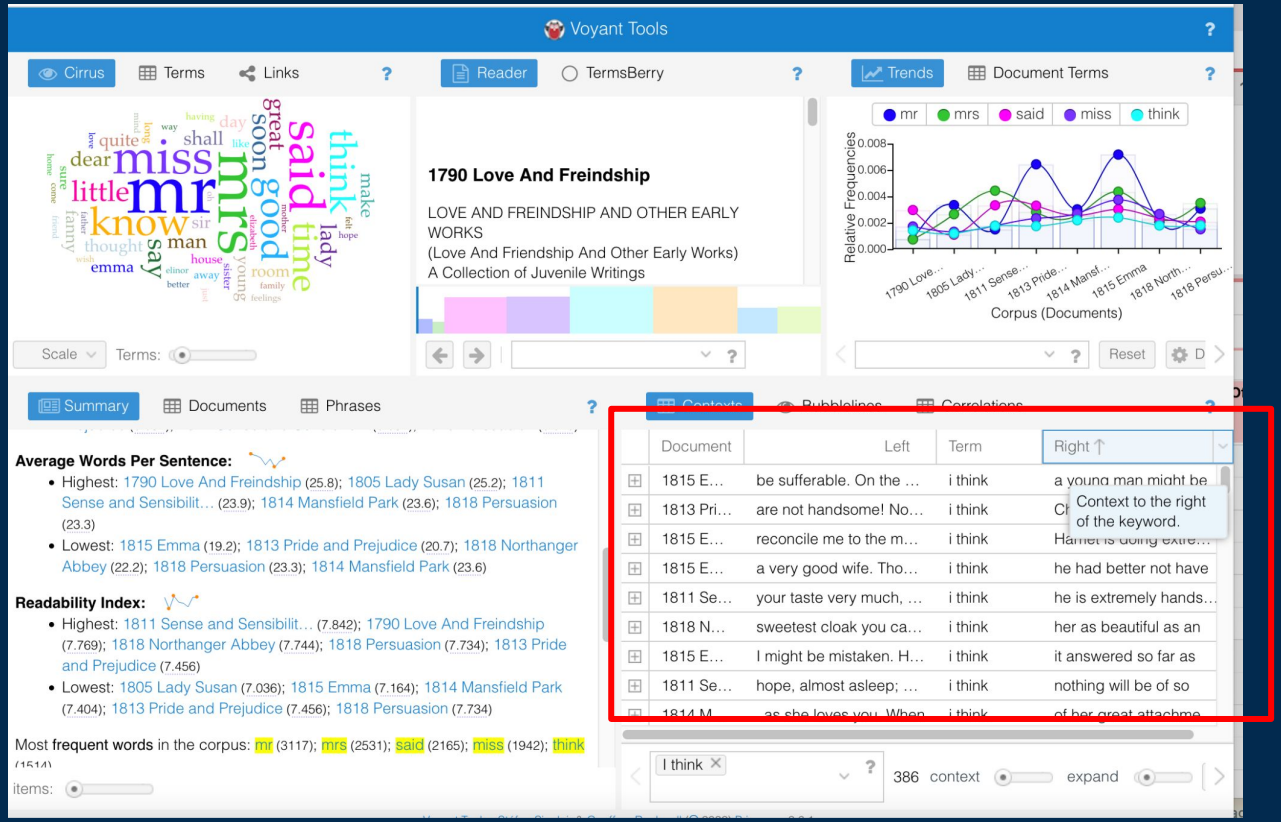
For more about stopwords––their history and their role in computational analysis today–– see this article by Daniel Rosenberg, “Stop, Words.” *Representations* 127, no. 1 (August 1, 2014): 83–92. https://doi.org/10.1525/rep.2014.127.1.83.

Voyant Tools

In the bottom right is a CONCORDANCE.
This gives you the context of words in your corpus as they appear in each document.
Try typing in “gentleman” and sorting by the words that appear on the left.
Toggle to the “Bubblelines” view. Type in “pounds,” “estate,” “money” and “inheritance.” What do you notice?

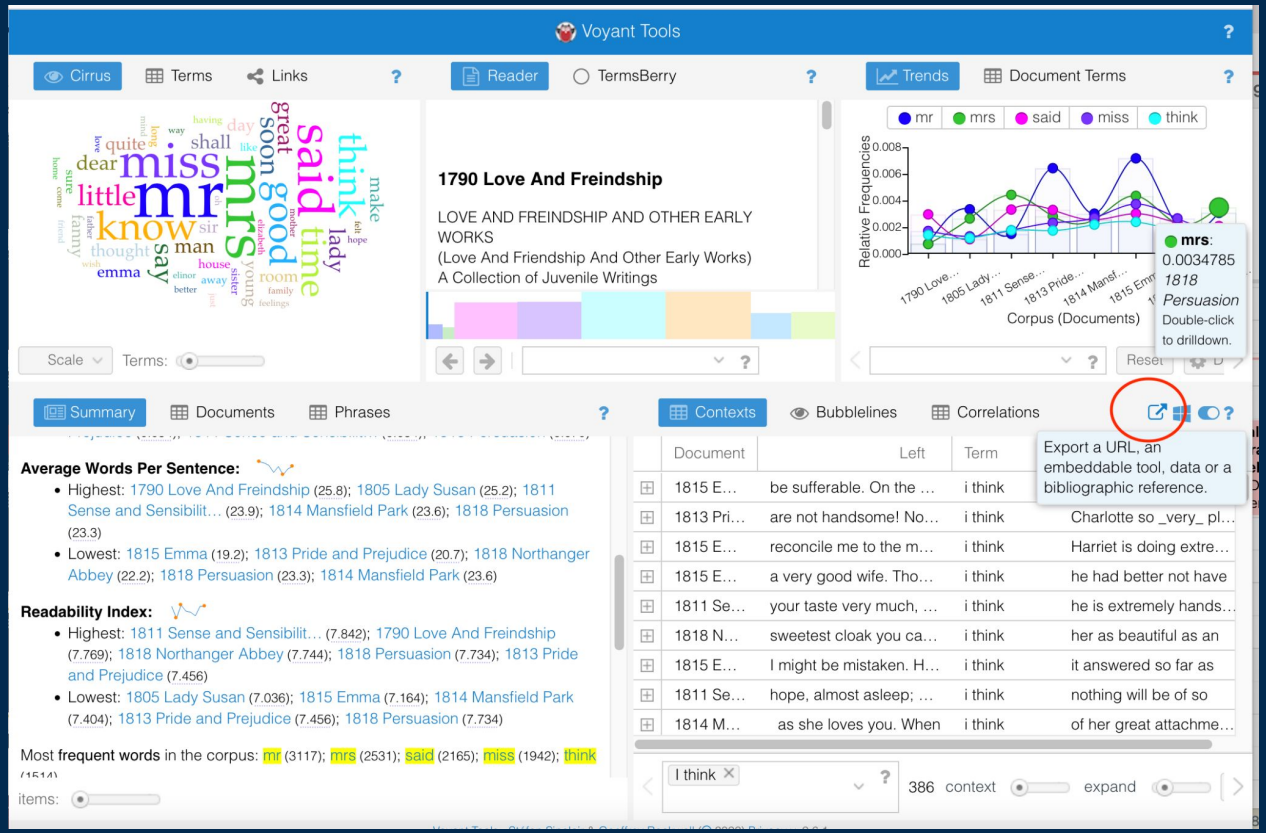
Voyant Tools

You can also look for short phrases in context:
Try typing in “I think” and sorting by the words that appear on the right.
What do you notice?



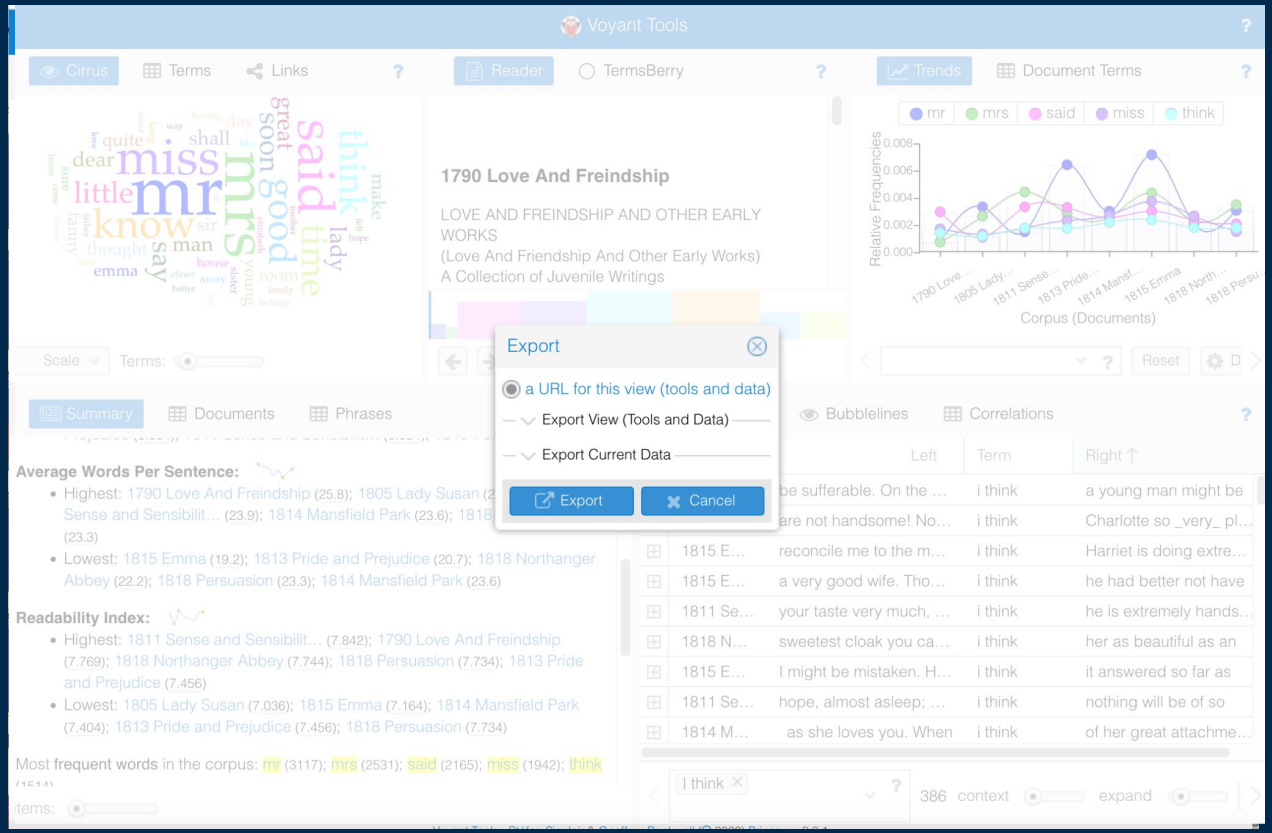
Finally, Voyant will also allow you to download data and visualizations.
Hover over the upper right corner of the Concordance view and click on the arrow and box export view

Voyant Tools



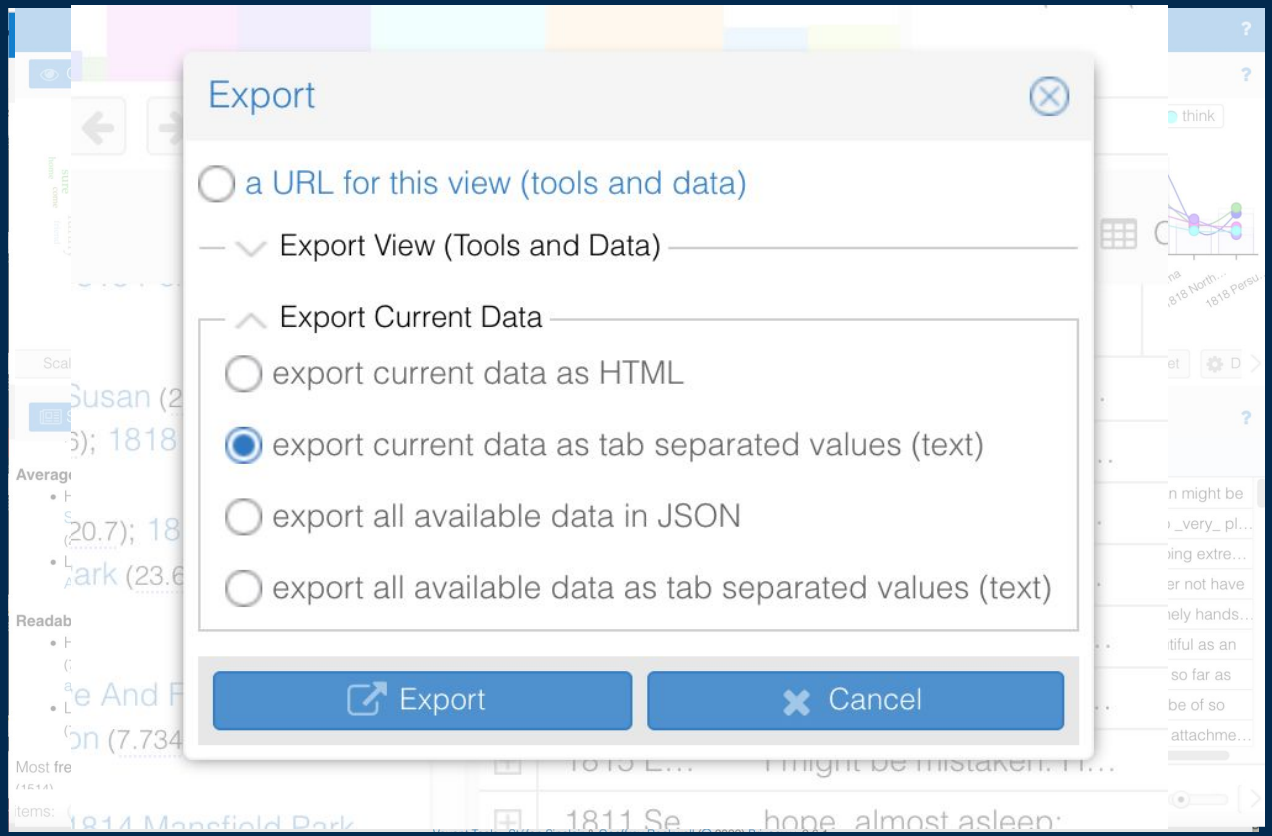
Voyant Tools

Finally, Voyant will also allow you to download data and visualizations.
Hover over the upper right corner of the Concordance view and click on the arrow and box export view
You should see a pop up menu.



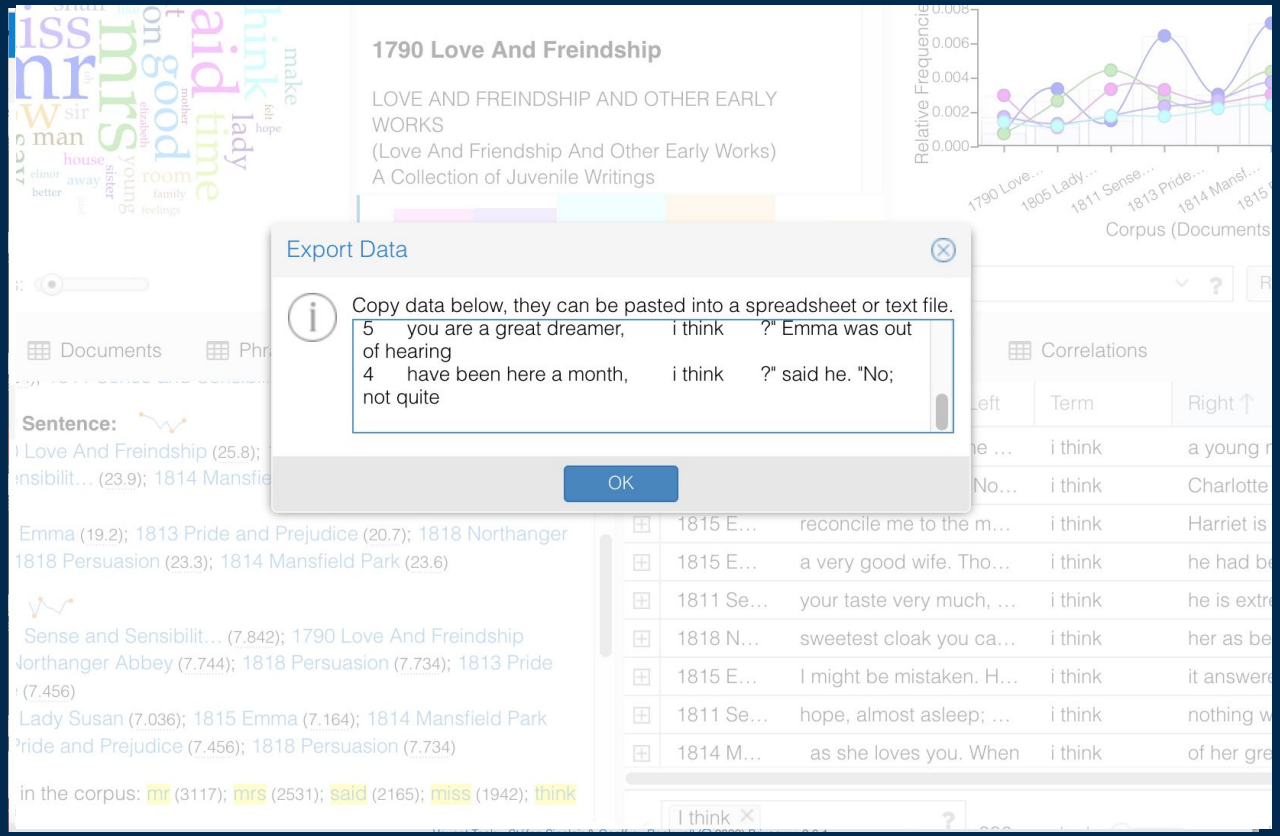
Voyant Tools

Finally, Voyant will also allow you to download data and visualizations.
Hover over the upper right corner of the Concordance view and click on the arrow and box export view
You should see a pop up menu. Click on the third option to “Export Current Data” and select tab sep. values



Voyant Tools

Exporting the current data as tab separated values (text) will give you a second popup window with data formatted in TSV that can can be copied into a spreadsheet (like Excel or Googlesheets) or a simple text editor



Voyant Tools

Take a minute to play around some of the features. Toggle the amount of words in the CONCORDANCE, or the “items” in the STATISTICS box.

Brainstorm a few questions that you could explore with this kind of interface.

What kind of questions could you ask?
What kind of questions could you not ask?

For further research:

If you want to experiment with Voyant or the topic model browser, go to:

tinyurl.com/4xha6a5a

to download a dataset of US Inaugural Addresses (1789-2021).

Then follow the same steps.as slides 2-17.

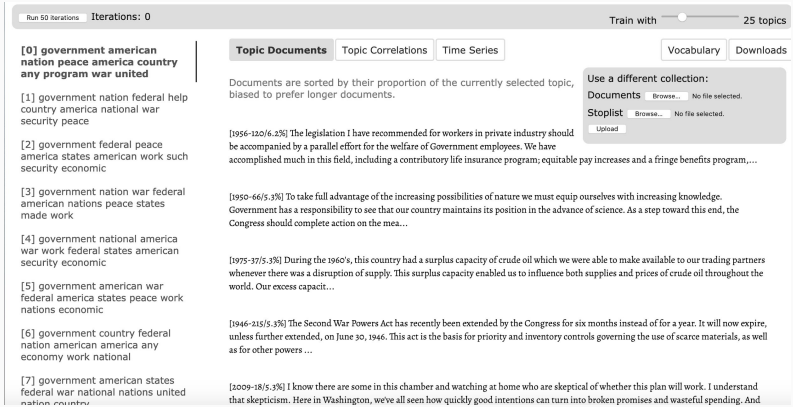
Topic Modeling

Please go to:

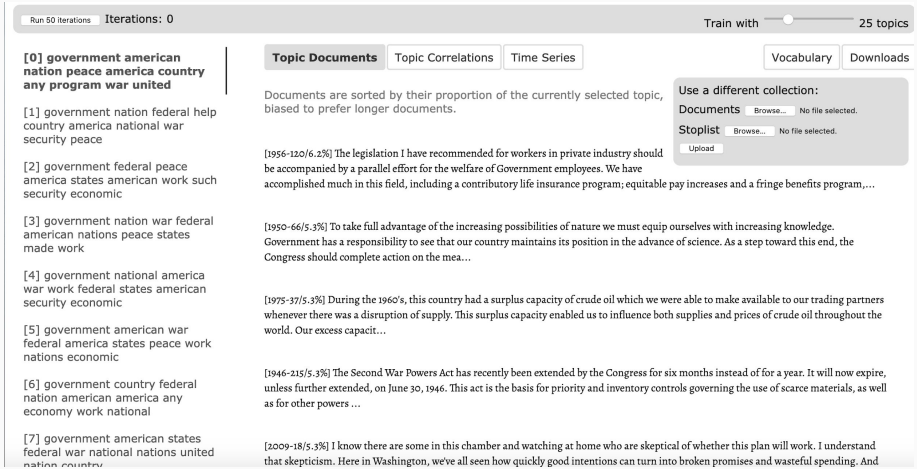
bit.ly/3kL754J

to familiarize yourself with the topic model tool. When you're ready, click on the RUN MODEL button:

You should see a screen like this:



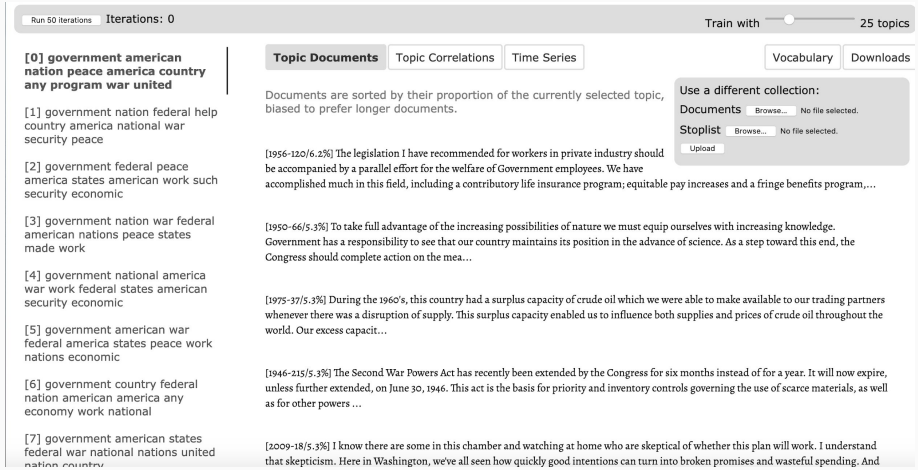
Topic Modeling



This is a topic model of US State of the Union Addresses.

- On the top left is the “run iterations” button. This is what starts your model. On the top right are the NUMBER of topics your model will find
- In the far left column are “topics” ––the categories that the algorithm has found.
- In the center are short snippets of the “documents” (here, the SotU speech texts)
- In the right hand column is the control panel for loading in your own set of documents

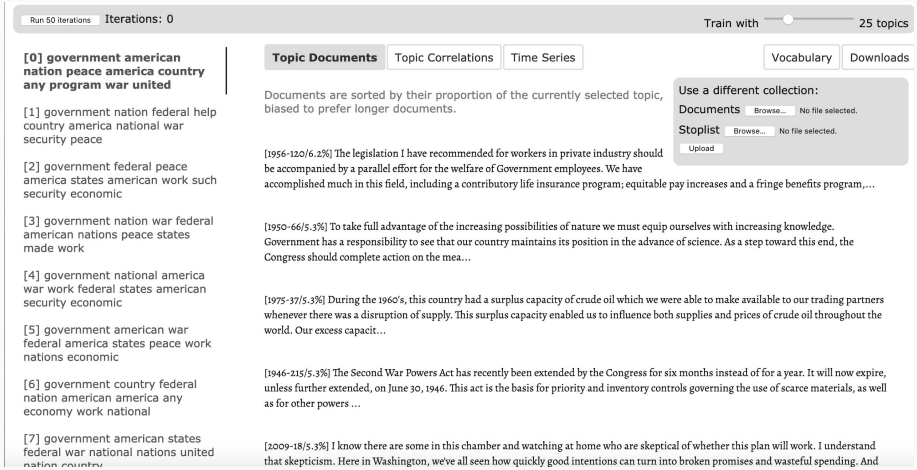
Topic Modeling



Things to keep in mind when using topic modeling.

- The algorithm is generating the number of topics that you tell it to. So if I tell my algorithm, “find 10 related clusters of words in this document” it will return 10 clusters. If I tell it to find 50, it will find 50.
- Topic models work iteratively and probabilistically. This means the model will change slightly every time you run it.
    - Try changing the number of topics to 10 (and click Run 50 iterations). What do you notice changes?
    - Try changing the number of topics to 70.
- Clicking on a “topic” in the right will bring up all of the documents that this topic appears in.

Topic Modeling



- Play around with the settings: run the topic model several more times (running hundreds of iterations), change the number of topics.
- Toggle to the Topic Correlations button
  - This shows you “topics” that occur together (which might suggest similarity between topics, or even that the topics themselves are not, in fact, distinct “topics”)
  - What do you notice?
- Toggle to the Time Series.
  - This shows you how this topic is distributed across the range of SotU addresses
  - What do you notice?

Email me! sceckert@princeton.edu

Thank you!

Resources:

Ted Underwood, "Topic modeling made just simple enough" (2012)
Ben Schmidt, "When You Have MALLET, everything looks like a nail" (2012)

Thank you!