

Text Analysis

Introduction to Voyant & Topic Modeling Tool

Dr. Sierra Eckert

Princeton University

sceckert@princeton.edu

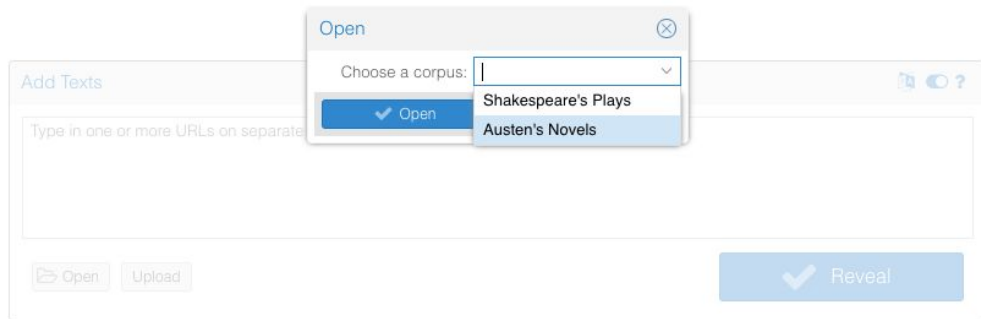
Text Analysis Tools

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



Voyant Tools is a web-based reading and analysis environment for digital texts.

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

Then press **Reveal**.

Voyant Tools

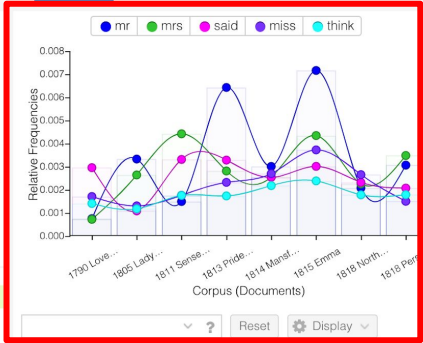
In the upper center square is the **TEXT of your CORPUS**.







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



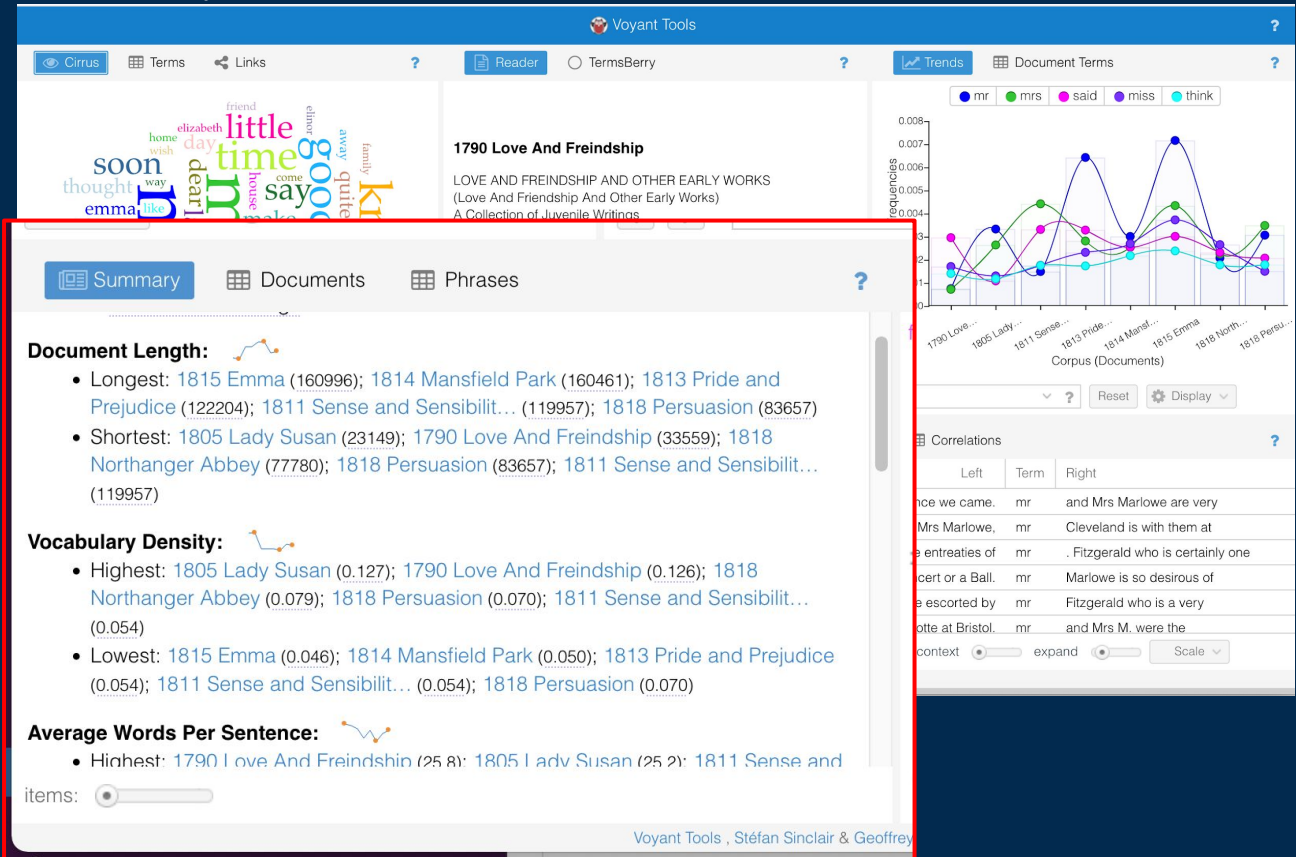
	Document	Left	Term	Right
	1) 1790 ...	genteel family since we came.	mr	and Mrs Marlowe are very
	1) 1790 ...	A brother of Mrs Marlowe,	mr	Cleveland is with them at
	1) 1790 ...	thirded by the entreaties of	mr	. Fitzgerald who is certainly one
	1) 1790 ...	a Concert or a Ball.	mr	Marlowe is so desirous of
	1) 1790 ...	Kickabout's; we were escorted by	mr	Fitzgerald who is a very
	1) 1790 ...	of my Charlotte at Bristol.	mr	and Mrs M. were the

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.

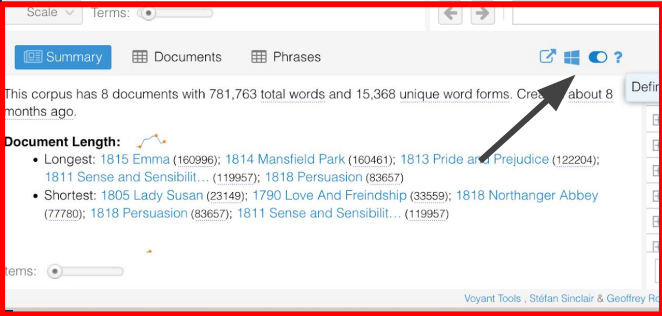
[illegible]

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 box’s upper right corner.

Voyant Tools



Voyant Tools



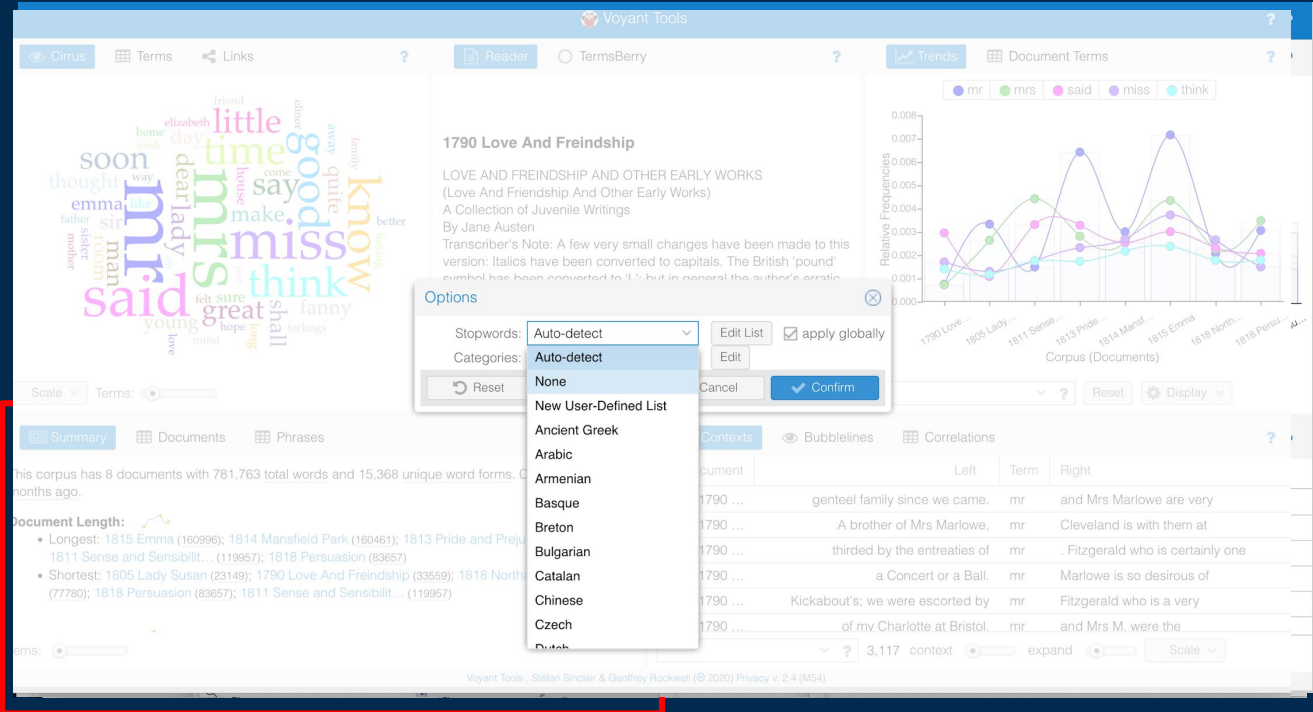
Voyant Tools

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?



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>.

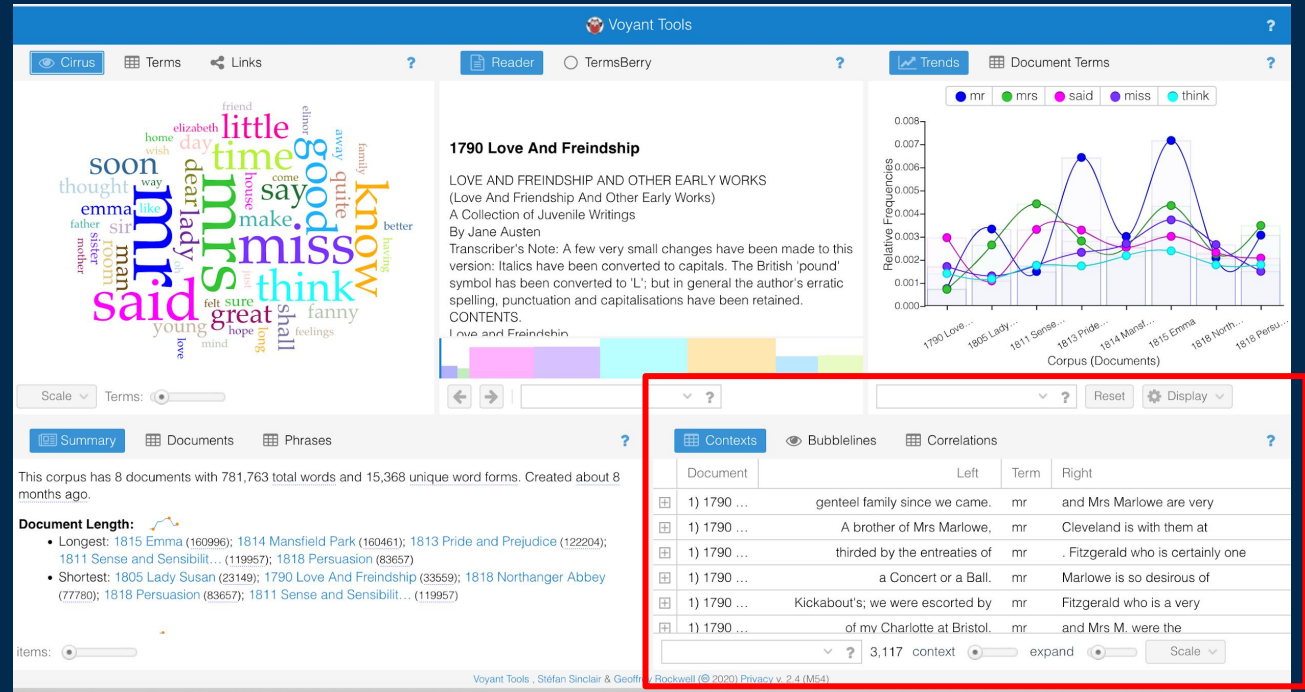
Voyant Tools

In the bottom right is a CONCORDANCE.

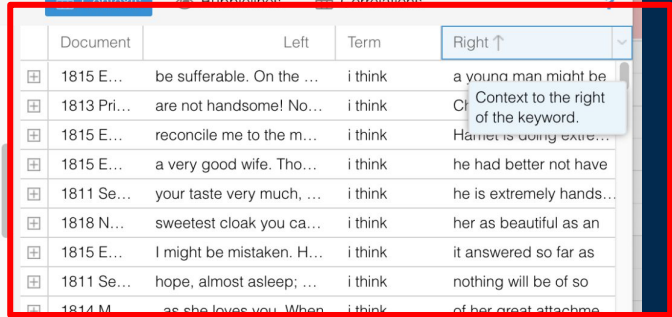
This gives you the context of words in your corpus as they appear in each document.

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?

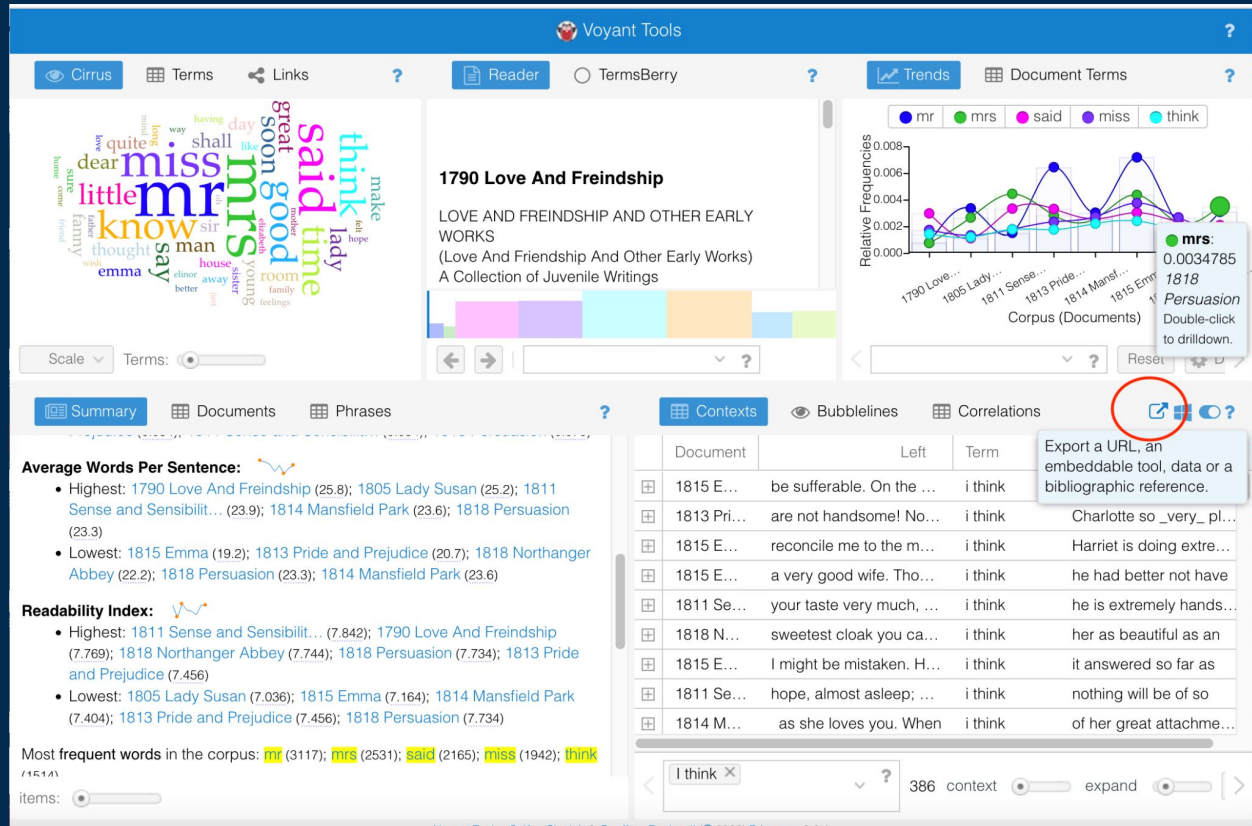


Voyant Tools



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



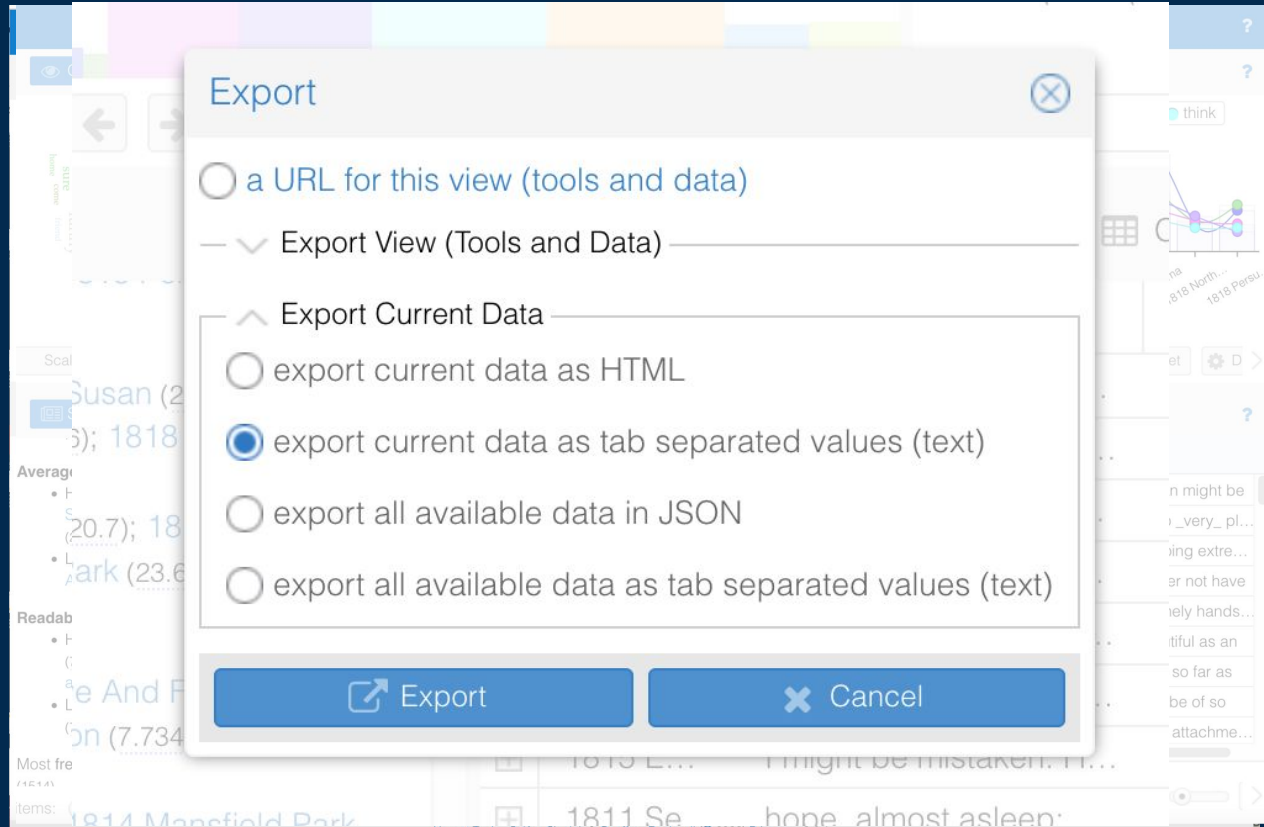
Hover over the upper right corner of the Concordance view and click on the arrow and box export view. You should see a pop up menu.

[illegible]

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



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

Voyant Tools

1790 Love And Freindship
LOVE AND FREINDSHIP AND OTHER EARLY WORKS
(Love And Friendship And Other Early Works)
A Collection of Juvenile Writings

Export Data

Copy data below, they can be pasted into a spreadsheet or text file.

5	you are a great dreamer,	i think	?" Emma was out of hearing
4	have been here a month,	i think	?" said he. "No; not quite

OK

Sentence: Love And Freindship (25.8); Sensibilit... (23.9); 1814 Mansfield Park (23.6)

Emma (19.2); 1813 Pride and Prejudice (20.7); 1818 Northanger 1818 Persuasion (23.3); 1814 Mansfield Park (23.6)

Sense and Sensibilit... (7.842); 1790 Love And Freindship Northanger Abbey (7.744); 1818 Persuasion (7.734); 1813 Pride (7.456)

Lady Susan (7.036); 1815 Emma (7.164); 1814 Mansfield Park Pride and Prejudice (7.456); 1818 Persuasion (7.734)

in the corpus: **mr** (3117); **mrs** (2531); **said** (2165); **miss** (1942); **think**

Left	Term	Right ↑	
...	i think	a young r	
No...	i think	Charlotte	
1815 E...	reconcile me to the m...	i think	Harriet is
1815 E...	a very good wife. Tho...	i think	he had b
1811 Se...	your taste very much, ...	i think	he is extr
1818 N...	sweetest cloak you ca...	i think	her as be
1815 E...	I might be mistaken. H...	i think	it answer
1811 Se...	hope, almost asleep; ...	i think	nothing w
1814 M...	as she loves you. When	i think	of her gre

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:

The screenshot displays the Topic Modeling tool interface. At the top, there are controls for 'Run 50 iterations' and 'Iterations: 0'. On the right, there is a 'Train with' slider set to '25 topics' and buttons for 'Vocabulary' and 'Downloads'. The main content area is divided into two columns. The left column lists seven topics, each with a number in brackets and a list of keywords: [0] government american nation peace america country any program war united; [1] government nation federal help country america national war security peace; [2] government federal peace america states american work such security economic; [3] government nation war federal american nations peace states made work; [4] government national america war work federal states american security economic; [5] government american war federal america states peace work nations economic; [6] government country federal nation american america any economy work national; [7] government american states federal war national nations united nation country. The right column shows document snippets sorted by their proportion of the currently selected topic, biased to prefer longer documents. It includes a 'Topic Documents' tab, a 'Stoplist' section with 'Browse...' and 'Upload' buttons, and several document snippets with their corresponding topic proportions in brackets. The snippets include text about legislation for workers in private industry, government responsibilities, surplus capacity of crude oil, the Second War Powers Act, and skepticism about intentions in Washington.

Topic Modeling

The screenshot shows a web-based interface for topic modeling. At the top, there's a header with "Run 50 iterations" and "Iterations: 0" on the left, and "Train with" followed by a slider and "25 topics" on the right. Below the header, there are three tabs: "Topic Documents", "Topic Correlations", and "Time Series". On the far right, there are buttons for "Vocabulary" and "Downloads".

The main content area is divided into three columns:

- Left Column (Topics):** A list of seven topics, each with a number in brackets and a list of words. For example, "[0] government american nation peace america country any program war united".
- Center Column (Documents):** A list of document snippets, each starting with a date and percentage in brackets, followed by a short text excerpt. For example, "[1956-120/6.2%] The legislation I have recommended for workers in private industry should be accompanied by a parallel effort for the welfare of Government employees. We have accomplished much in this field, including a contributory life insurance program; equitable pay increases and a fringe benefits program....".
- Right Column (Control Panel):** A panel titled "Use a different collection:" with sections for "Documents" (Browse... No file selected.), "Stoplist" (Browse... No file selected.), and "Upload" (Upload).

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

The screenshot shows a web-based topic modeling interface. At the top, it says "Run 50 iterations" and "Iterations: 0". On the right, there's a "Train with" slider set to "25 topics". Below this are tabs for "Topic Documents", "Topic Correlations", and "Time Series". On the far right, there are buttons for "Vocabulary" and "Downloads".

On the left, a list of topics is shown, each with a number in brackets and a list of words. Topic [0] is highlighted:

- [0] government american nation peace america country any program war united
- [1] government nation federal help country america national war security peace
- [2] government federal peace america states american work such security economic
- [3] government nation war federal american nations peace states made work
- [4] government national america war work federal states american security economic
- [5] government american war federal america states peace work nations economic
- [6] government country federal nation american america any economy work national
- [7] government american states federal war national nations united nation econu

On the right, under the "Topic Documents" tab, a text box explains: "Documents are sorted by their proportion of the currently selected topic, biased to prefer longer documents." Below this, several document excerpts are shown, each starting with a topic and percentage in brackets, followed by a snippet of text. For example, "[1956-120/6.2%] The legislation I have recommended for workers in private industry should be accompanied by a parallel effort for the welfare of Government employees. We have accomplished much in this field, including a contributory life insurance program; equitable pay increases and a fringe benefits program,..."

At the top right of the document excerpts, there's a section titled "Use a different collection:" with buttons for "Documents", "Stoplist", and "Upload". Each button has a "Browse..." link and a status "No file selected."

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

The screenshot shows a web-based interface for topic modeling. At the top, it says "Run 50 iterations" and "Iterations: 0". On the right, there's a "Train with" slider set to "25 topics" and buttons for "Vocabulary" and "Downloads". Below the top bar, there are three tabs: "Topic Documents" (selected), "Topic Correlations", and "Time Series". The main area is divided into two columns. The left column lists 7 topics, each with a list of associated words. The right column shows a list of documents sorted by their proportion of the currently selected topic, with a note that documents are sorted by their proportion of the currently selected topic, biased to prefer longer documents. A sidebar on the right contains a "Use a different collection:" section with buttons for "Documents", "Stoplist", and "Upload", each with a "Browse..." button and a "No file selected." message.

Run 50 iterations Iterations: 0 Train with 25 topics Vocabulary Downloads

Topic Documents Topic Correlations Time Series

Documents are sorted by their proportion of the currently selected topic, biased to prefer longer documents.

Use a different collection:
Documents Browse... No file selected.
Stoplist Browse... No file selected.
Upload

[0] government american nation peace america country any program war united

[1] government nation federal help country america national war security peace

[2] government federal peace america states american work such security economic

[3] government nation war federal american nations peace states made work

[4] government national america war work federal states american security economic

[5] government american war federal america states peace work nations economic

[6] government country federal nation american america any economy work national

[7] government american states federal war national nations united nation country

[1956-120/6.2%] The legislation I have recommended for workers in private industry should be accompanied by a parallel effort for the welfare of Government employees. We have accomplished much in this field, including a contributory life insurance program; equitable pay increases and a fringe benefits program,...

[1950-66/5.3%] To take full advantage of the increasing possibilities of nature we must equip ourselves with increasing knowledge. Government has a responsibility to see that our country maintains its position in the advance of science. As a step toward this end, the Congress should complete action on the mea...

[1975-37/5.3%] During the 1960's, this country had a surplus capacity of crude oil which we were able to make available to our trading partners whenever there was a disruption of supply. This surplus capacity enabled us to influence both supplies and prices of crude oil throughout the world. Our excess capacit...

[1946-215/5.3%] The Second War Powers Act has recently been extended by the Congress for six months instead of for a year. It will now expire, unless further extended, on June 30, 1946. This act is the basis for priority and inventory controls governing the use of scarce materials, as well as for other powers ...

[2009-18/5.3%] I know there are some in this chamber and watching at home who are skeptical of whether this plan will work. I understand that skepticism. Here in Washington, we've all seen how quickly good intentions can turn into broken promises and wasteful spending. And

- 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!