Introduction to Computational Text Analysis

POLS 2395: Environmental Politics and Policy

Nancy Riordan
Spring 2023

Sarah Connell
Digital Integration Teaching Initiative (DITI)

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Workshop Agenda

- Introduction to key terms and concepts in computational text analysis (What)
- Discussion on its applications and uses in research (Why)
- Demonstration of web-based text analysis tools (How)
  - Word Counter, Word Trees, Voyant, Lexos

Slides, handouts, and data available at:

https://bit.ly/sp23-riordan

Feel free to ask questions at any point during the presentation!

What is Computational Text Analysis?

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Computational Text Analysis

Computational text analysis refers to the **array of methods used to "read" texts with a computer.** It is similar to statistical analysis, but the data is texts (words) instead of numbers.

Text analysis:

- Involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- Includes methods such as word count frequency, keywords in context, computational modeling (with machine learning), and sentiment analysis.
- Is conducted using web-based tools or coding languages like Python and R

Feel free to ask questions at any point during the presentation!

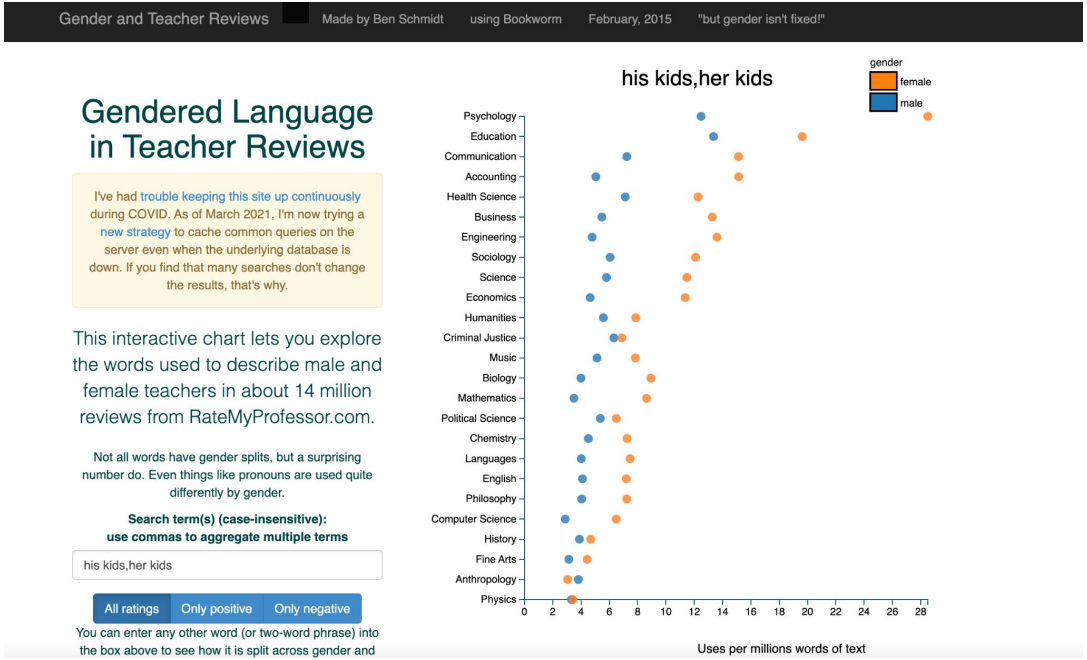
Why Computational Text Analysis?

Computational text analysis can help us **analyze very large amounts of data, identify keywords,** and **discover patterns** in texts. Using text analysis, researchers may **find surprising results** that they would not have discovered from close reading or traditional methods alone.

For example: "Gendered Language in Teacher Reviews" by Ben Schmidt shows stark differences in the ways that male and female professors are reviewed on "Rate My Professor."

Feel free to ask questions at any point during the presentation!

Gendered Language



Go to bit.ly/schmidt-gender and try a few queries.

For example:

- Smart
- Ditzy
- Unprofessional
- Nice
- What else did you try?
- How do you think Schmidt determined professor gender?

Feel free to ask questions at any point during the presentation!

Key Terms

- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stop words include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **nGram:** A continuous sequence of *n* items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'

Feel free to ask questions at any point during the presentation!

Corpus Building

Questions to consider as you begin your research:

- What are my research questions and why am I creating a corpus?
- What am I asking my corpus to do?
- What text(s) should form my corpus to answer my research questions?
- How should I organize my corpus to streamline my research processes and save time?
- For more on building a corpus, see **this handout**.

Feel free to ask questions at any point during the presentation!

Our Corpus

For our corpus, we will work with a set of State of the Union addresses from 1990 to 2019.

You can download these files here: bit.ly/corpus-pols2395

The easiest way to work with these files is to choose "Download all" and open them with a plain-text editor (TextEdit on Mac, Notepad on Windows).

Feel free to ask questions at any point during the presentation!

Initial Corpus Analysis

Open any one of the texts from the sample corpus:

What can you observe about the text? How long is it? What kinds of language does it use? What kinds of analysis might you do with a text like this?

Scan through a few more: do they seem largely similar? What do you think might be different?

Feel free to ask questions at any point during the presentation!

Exploratory Tools: Word Counter and Word Trees

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

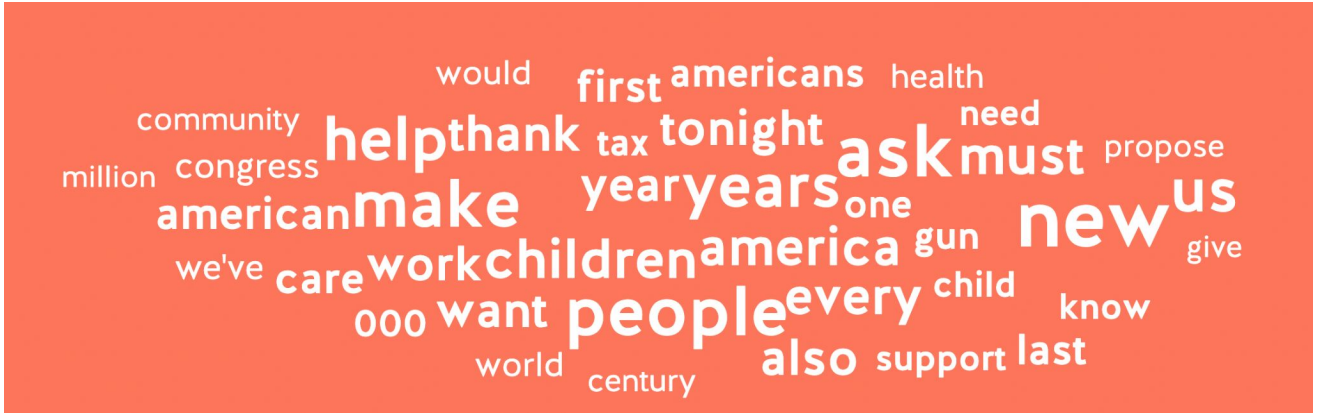
Word Counter

- https://databasic.io/en/wordcounter/
- A user-friendly **basic word counting tool**
- Allows you to count words, bigrams, and trigrams in plain text files and to download spreadsheets with your results
- The max file upload is 10MB
- The default is to lowercase all words and apply stopwords
- It can be run with and without stopwords and lowercasing

Feel free to ask questions at any point during the presentation!

Word Counter Example

This is a **word cloud**. It is helpful to get a sense of the **most used words in a document**. Words used more often are bigger, and ones used less often are smaller.



What seems significant in the most frequent terms from Clinton's 2000 SotU address?

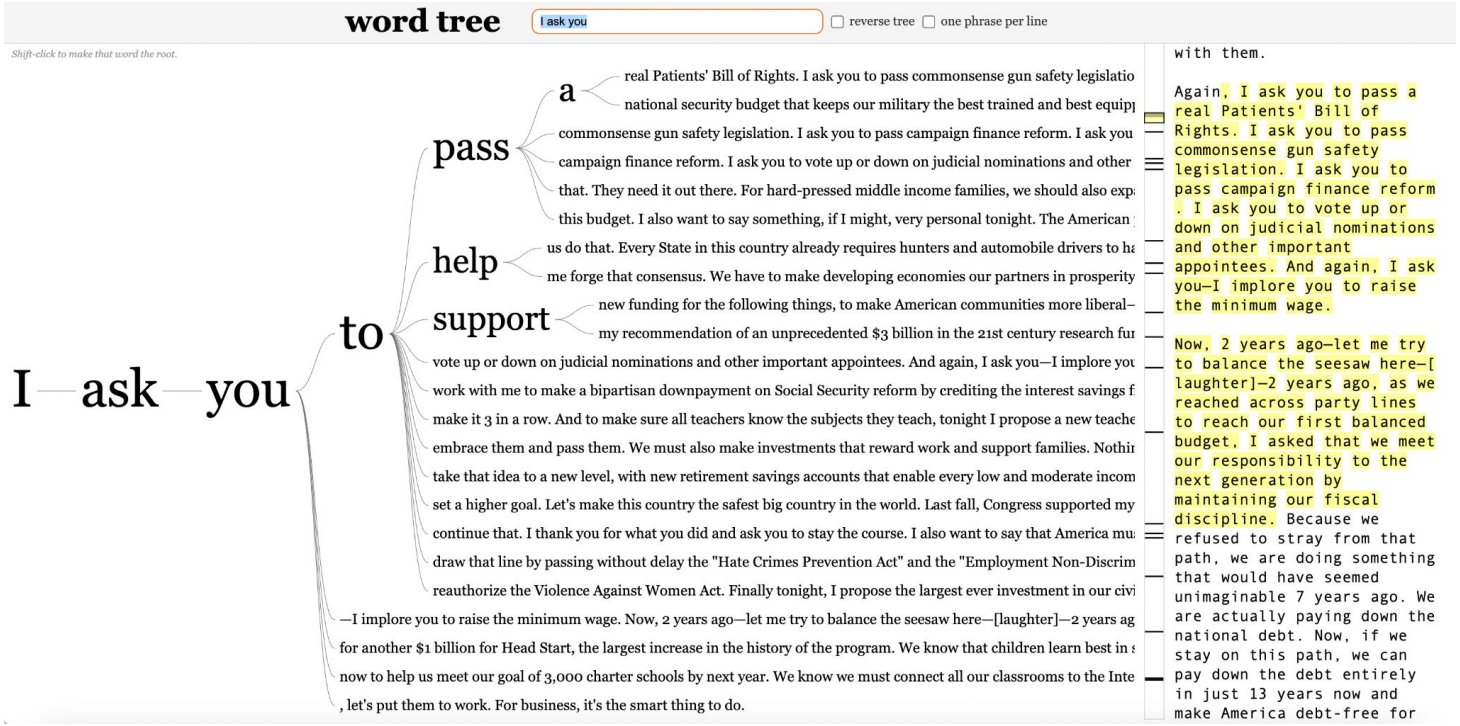
Are there any things here that surprise you?

Word Trees

- https://www.jasondavies.com/wordtree/
- A word tree **depicts multiple parallel sequences of words**
- This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest.
- There are some restrictions in size with this tool: fewer than 1 million words should work

Feel free to ask questions at any point during the presentation!

Word Tree Example



Your Turn!

Use the sample text or texts of your choice and begin practicing web-browser text analysis. **Explore Word Counter and Word Trees!**

Discussion Prompts

- What limitations are you observing? What functionalities do you wish these tools might offer?
- Even with these limitations, how can you apply these simple tools in your research and exploration?

Feel free to ask questions at any point during the presentation!

Tools for corpus exploration: Voyant and Lexos

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Voyant

Voyant makes it possible to **perform analyses on one or multiple files in many ways**, including word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances.

https://voyant-tools.org/

Feel free to ask questions at any point during the presentation!

VOYANT

see through your text

Click on Upload and navigate to the folder with the text documents you wish to analyze.

Alternatively, insert URLs or full text into the textbox.

Click here for help and advanced options

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Voyant: Dashboard

Results:

After you upload your corpus, you will see the default results page with multiple panes:

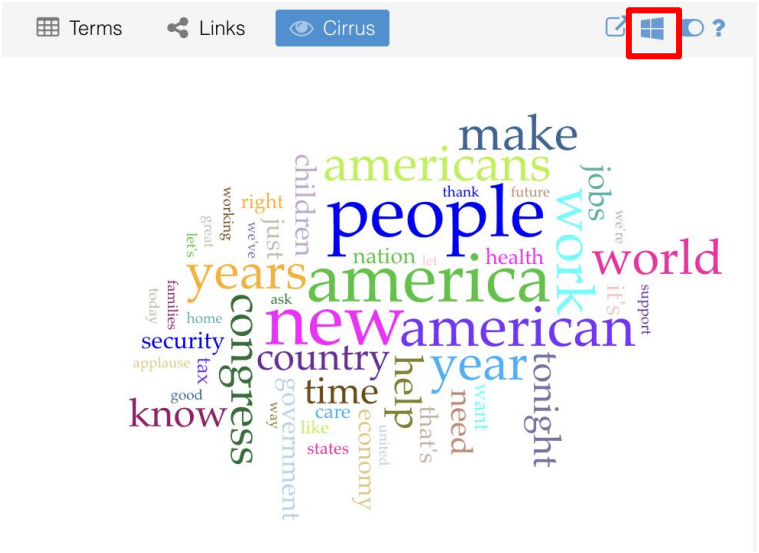
- A word cloud
- Reader section
- Trends
- Document Summary
- Word Contexts

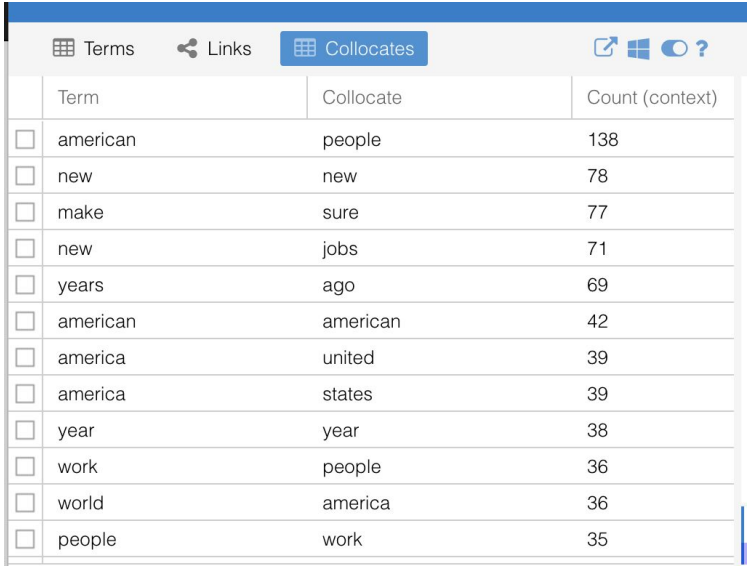
These boxes can all be changed!

Feel free to ask questions at any point during the presentation!

Voyant: Changing displayed results

Hover on the right top corner of a pane and buttons will appear. Select the panes button and choose a new option from the dropdown menu. For example, we might want to try out the "Collocates" tool instead of the word cloud. Click on the **?** to learn more about how the tool works.





Term	Collocate	Count (context)
american	people	138
new	new	78
make	sure	77
new	jobs	71
years	ago	69
american	american	42
america	united	39
america	states	39
year	year	38
work	people	36
world	america	36
people	work	35

Feel free to ask questions at any point during the presentation!

Lexos

Lexos provides a step-by-step guide for text uploading, preparation, and analysis.

- **Upload**: upload your .txt file
- **Manage**: select the files you want to prepare and analyze
- **Prepare**: prepare your text for analysis
- **Visualize**: create visualizations of patterns across your corpus or in single texts
- **Analyze**: analyze your text

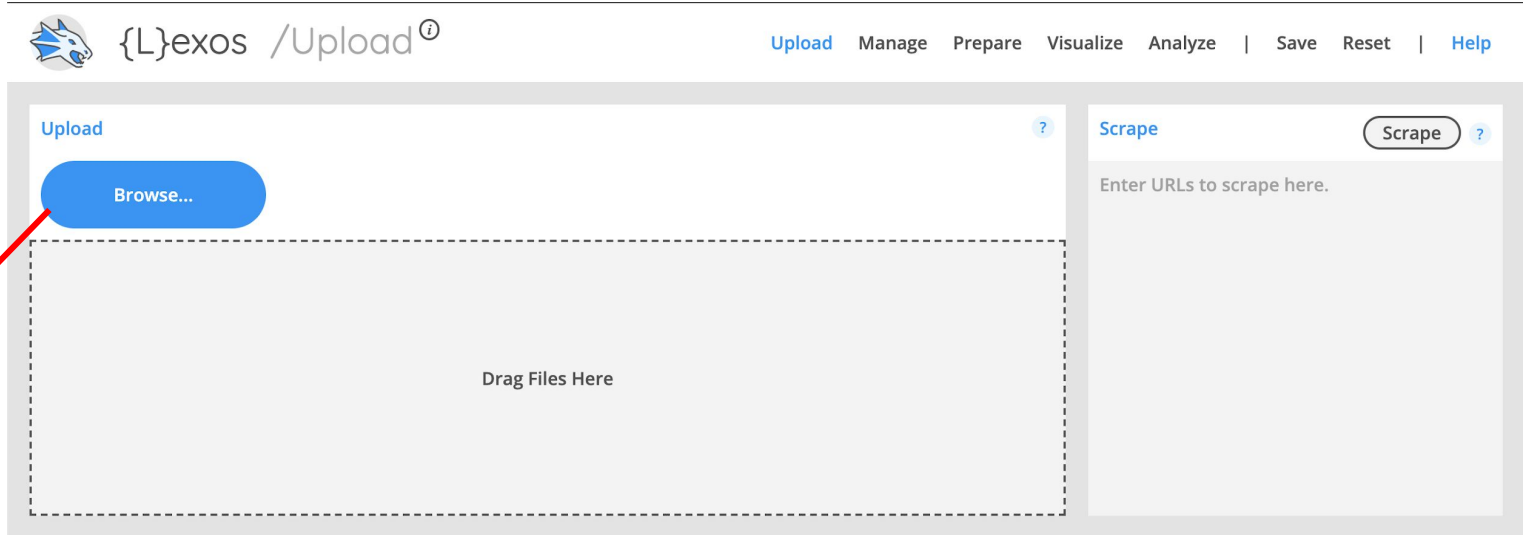
http://lexos.wheatoncollege.edu/upload

Feel free to ask questions at any point during the presentation!

Lexos: Upload

Click Browse and select your entire text (or drag file into the “Drag Files Here” area)

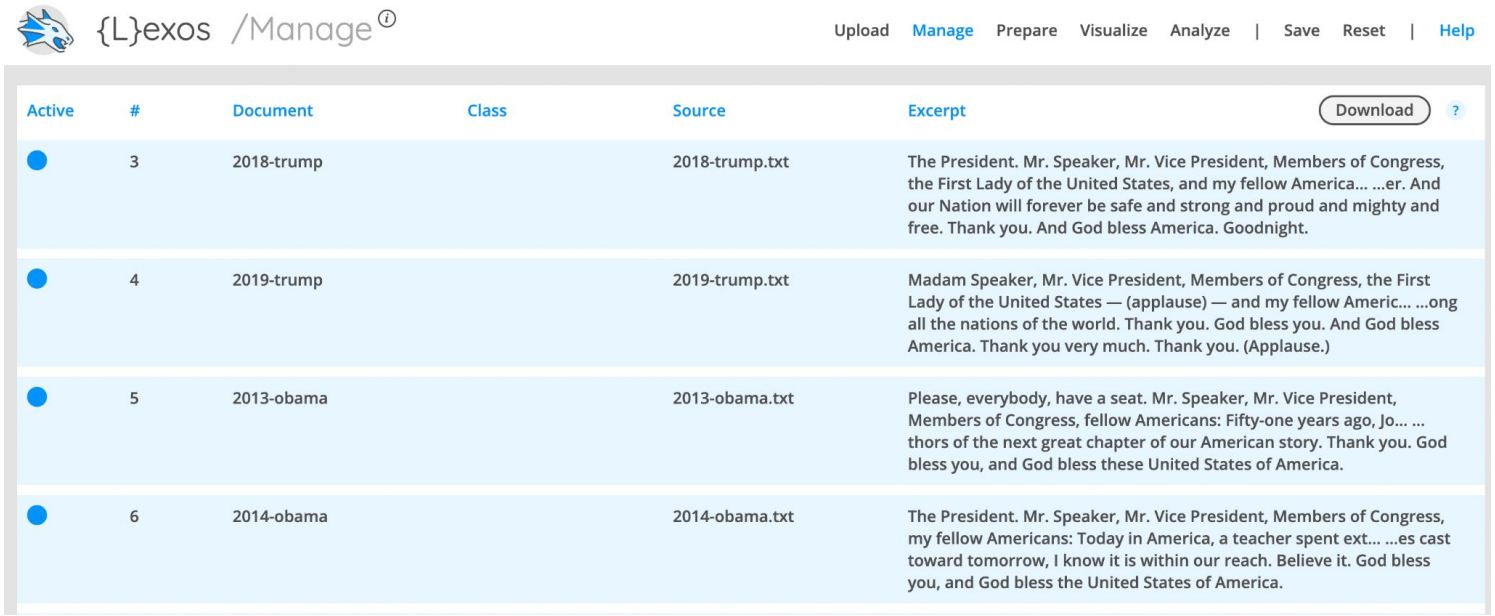
It can be easy to miss when the upload is done—click “Manage” to double check that the text file is there.



Feel free to ask questions at any point during the presentation!

Lexos: Manage

Make sure the document you want to use is selected (blue = selected, gray = not selected)



Feel free to ask questions at any point during the presentation!

Lexos: Prepare (scrub)

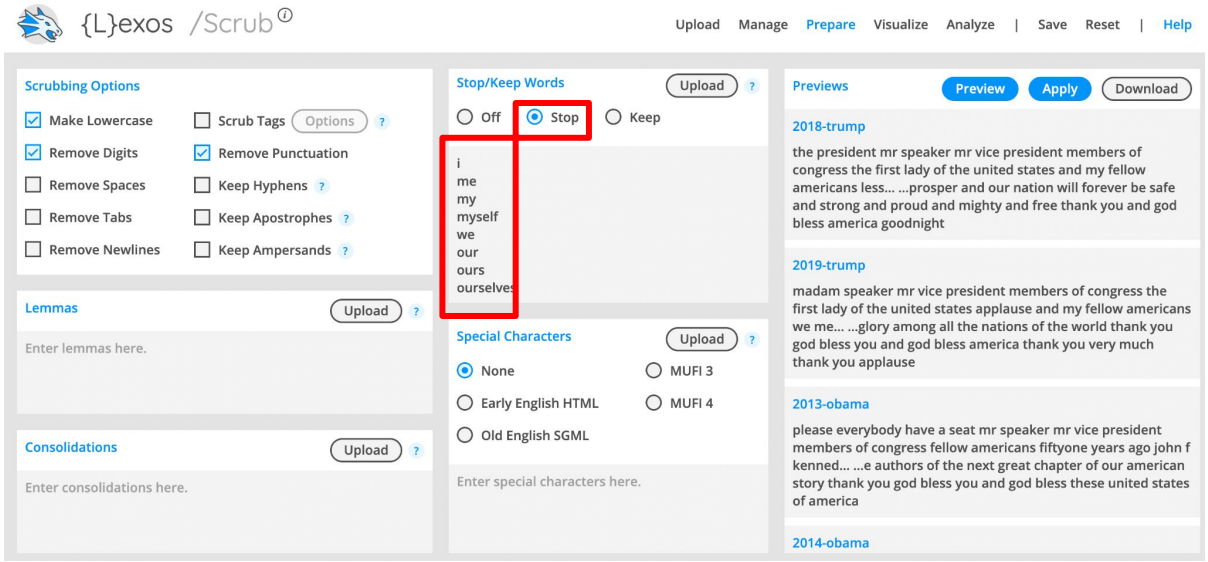
Lexos demonstrates some more advanced options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **Make Lowercase**: make all your letters lowercase. Even though you know "A" and "a" are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation**: remove punctuation, which may influence your results.
- **Stop/Keep Words**: remove a list of words. Usually these would be **stopwords**, or the most common words in a language (English: the, a she, her, it, him, they, etc).
- **Lemmas**: standardize to the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to "talk"

Feel free to ask questions at any point during the presentation!

Lexos: Removing Stopwords

Get a list of English stopwords here: https://gist.github.com/sebleier/554280. Copy and paste the stopwords (hit "raw", then select all and copy) into the “Stop/Keep Words” box then select “Stop”



Feel free to ask questions at any point during the presentation!

Lexos: Applying your Preparations

BEFORE PREP

2013-obama

Please, everybody, have a seat. Mr. Speaker, Mr. Vice President, Members of Congress, fellow Americans: Fifty-one years ago, Jo... ...thors of the next great chapter of our American story. Thank you. God bless you, and God bless these United States of America.

2014-obama

The President. Mr. Speaker, Mr. Vice President, Members of Congress, my fellow Americans: Today in America, a teacher spent ext... ...es cast toward tomorrow, I know it is within our reach. Believe it. God bless you, and God bless the United States of America.

AFTER PREP

2013-obama

please everybody have a seat mr speaker mr vice president members of congress fellow americans fiftyone years ago john f kenned... ...e authors of the next great chapter of our american story thank you god bless you and god bless these united states of america

2014-obama

the president mr speaker mr vice president members of congress my fellow americans today in america a teacher spent extra time... ...ur eyes cast toward tomorrow i know it is within our reach believe it god bless you and god bless the united states of america

Once you have made decisions about your preparations, click "**Apply**" and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your new corpus.

Feel free to ask questions at any point during the presentation!

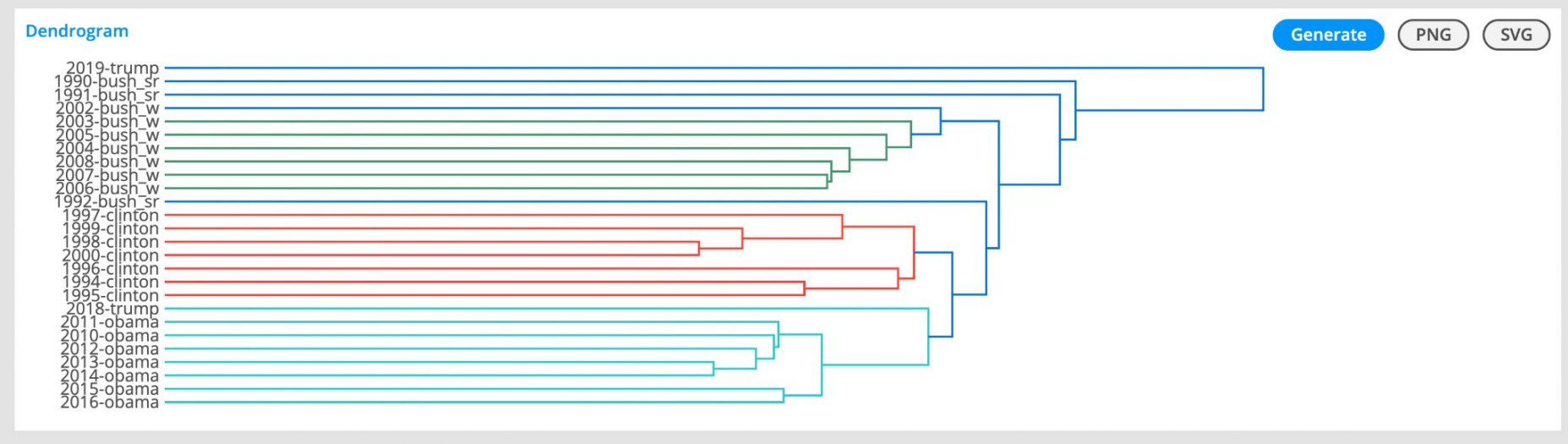
Lexos: Analyze > Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms are able to show the hierarchy between objects. Dendrograms show:

- Similarities between texts
  - The greater the distance between texts, the less similar they are
  - The smaller the distance between texts, the more similar they are

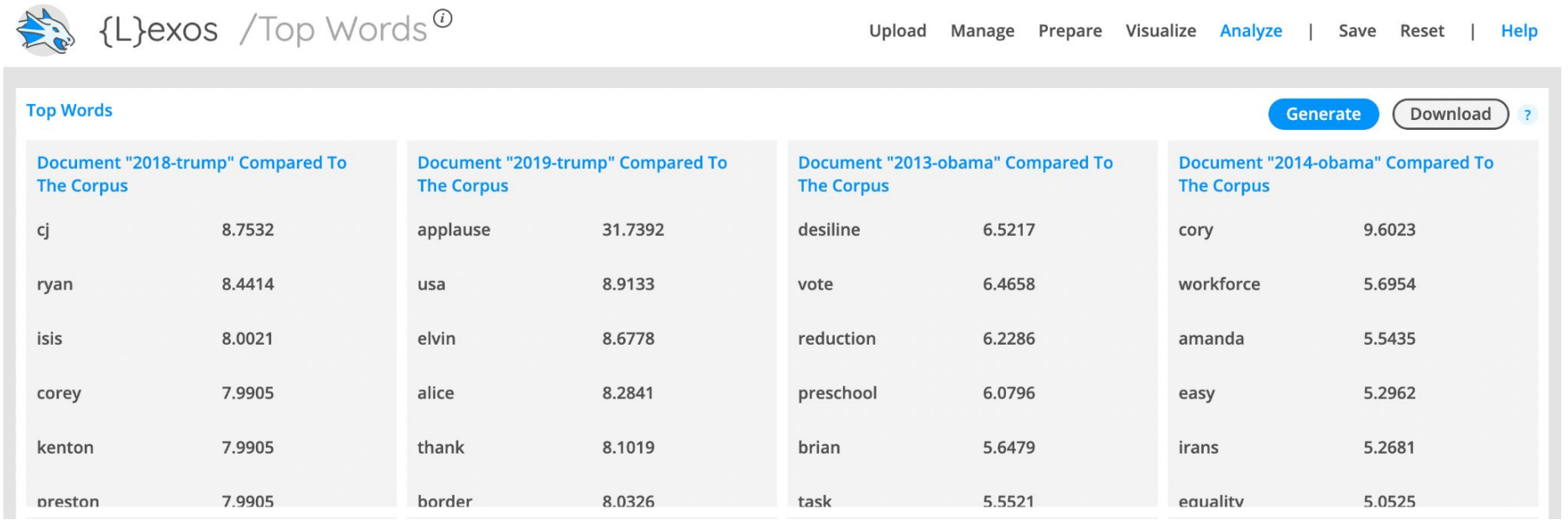
Feel free to ask questions at any point during the presentation!

Lexos: Analyze > Dendrogram



Feel free to ask questions at any point during the presentation!

Lexos: Analyze > Top words



Feel free to ask questions at any point during the presentation!

Lexos: Save or Reset Your Results

Lexos allows you to **save** your results as a Lexos file. If you do this, you can re-upload the Lexos file any time to access your cleaned-up corpus as well as the different analyses you've done. You can also download modified text files from the "Manage" page, which you can use with other tools if you would like.

You can also save individual visualizations as images (PNGs).

Finally, if you want to start over, you can "Reset" your Lexos dashboard.

Feel free to ask questions at any point during the presentation!

Your Turn!

Use the sample text or texts of your choice and begin practicing web-browser text analysis. **Explore Lexos and Voyant!**

Discussion Prompts

- What difference did you notice between Voyant and Lexos?
- Which tool do you prefer and why?
- How would you want to use these tools in this class and future?

Feel free to ask questions at any point during the presentation!

Learn more

Handouts (download these to use embedded links):

- Building a corpus
- More links and resources for text analysis
- Lexos
- Voyant
- WordCounter

Feel free to ask questions at any point during the presentation!

Thank you!

If you have any questions, contact us at nulab.info@gmail.com

Sign up for our office hours at: https://calendly.com/diti-nu

We'd love your feedback! Please fill out a short survey here: https://bit.ly/diti-feedback

Slides, handouts, and data is available at: https://bit.ly/sp23-riordan

Feel free to ask questions at any point during the presentation!