Computational Text Analysis for Content Analysis

Jonathan Sullivan & Colleen Nugent
International Relations
Carl Cilke
Fall 2020

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Defining Computational Text Analysis
- Demonstration of web-based text analysis tools
- Your turn!

Slides, handouts, and data available at **https://bit.ly/diti-fall2020-cilke**

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

Workshop Objectives

- Understand best practices for collecting and storing textual data when performing basic computational text analysis
- Understanding how web-based computational text analysis programs work, such as in their behind-the-scenes data preparation
- Understanding how to interpret the results from your text analysis

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

Computational Text Analysis

Computational text analysis refers to an array of methods that can be used to “read” texts with a computer. This form of analysis can range from basic word frequency counts to more advanced techniques like machine learning.

Text analysis is often used on a **corpus**, or a collection of multiple texts, and provides a glimpse into patterns across the texts. Some people also perform text analysis on larger individual documents, like novels.

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

Why Computational Text Analysis?

Computational text analysis can help us analyze a **ton** of data and discover **patterns** in texts.

Particular disciplines care **deeply** about the language used and how this language may reach intended audiences. Text analysis provides another method for approaching these discourses.

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

Creating a Corpus

1. Choose the texts you want to include in your corpus
2. Create a folder on your computer titled “awd_corpus” or something even more specific
3. Copy and paste your texts into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
    a. Mac users, you will need to make your Text Edit into a plain text editor. Open Text Edit, go to Preferences, and make sure “plain text” is selected
4. Save each text as a different plain text file (with a .txt extension). Name your files so you know what is in them!

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

Our Sample Corpus

Our corpus collects several State of the Union addresses from presidents over the years—spanning George H. W. Bush to Donald Trump. Our files are a series of plain text (.txt) files.

Download this corpus from the email that Professor Cilke forwarded, or from: **https://bit.ly/diti-fall2020-cilke**

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

Lexos

Northeastern University
NULab for Texts, Maps, and Networks

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

Lexos: http://lexos.wheatoncollege.edu/upload

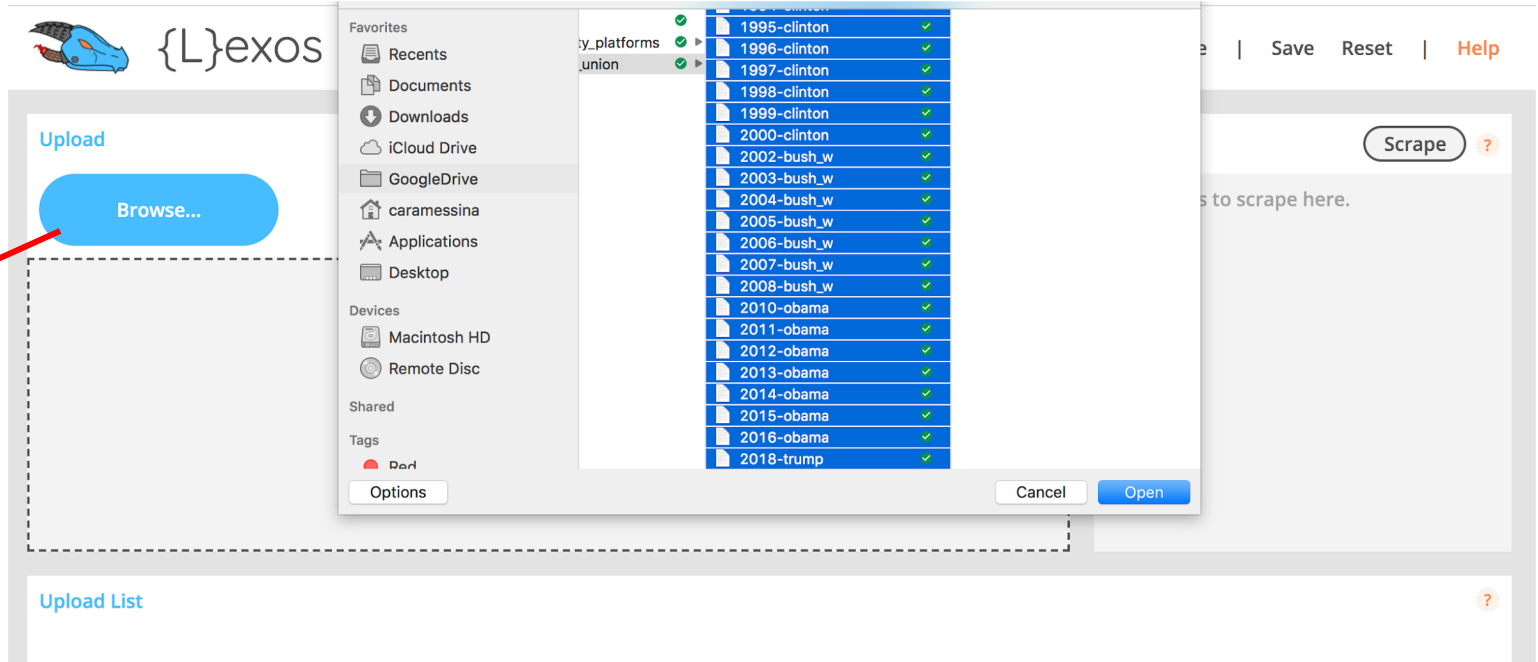
Lexos provides a step-by-step guide for corpus uploading, preparation, and analysis.

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

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

Lexos: Upload

Click Browse and select your entire corpus (or drag and drop)



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

Lexos: Manage

Make sure all the documents in your corpus you want to use are selected (blue = selected, gray = not selected)

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

Lexos: Prepare (scrub)

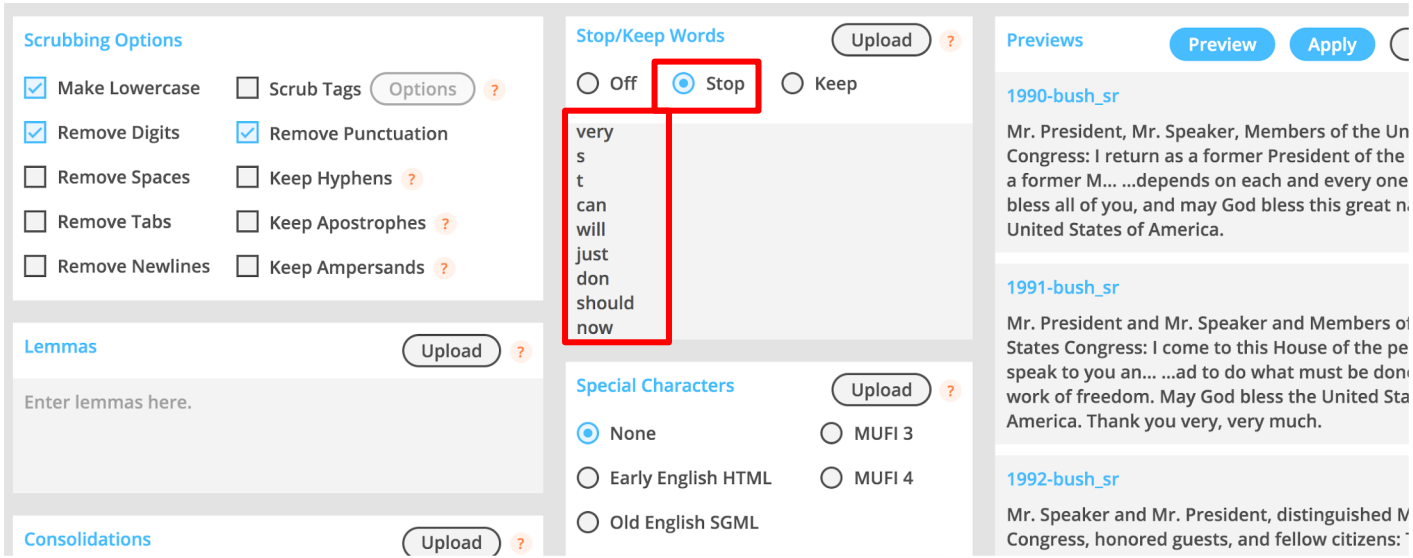
Lexos demonstrates the different options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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

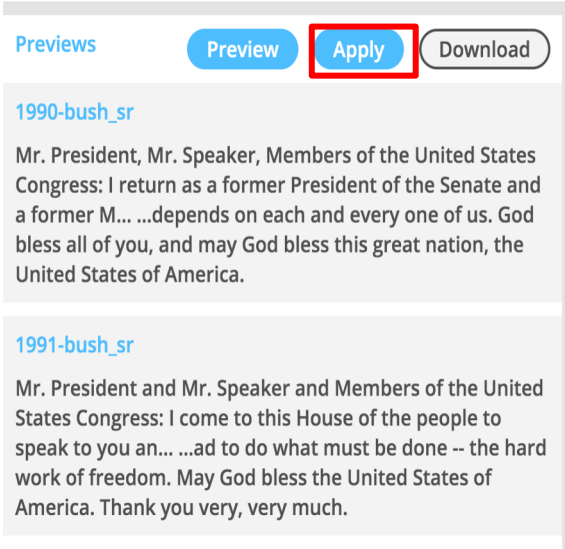
Get a list of English stopwords here: https://gist.github.com/sebleier/554280 (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the "Stop/Keep Words" box then select "Stop"



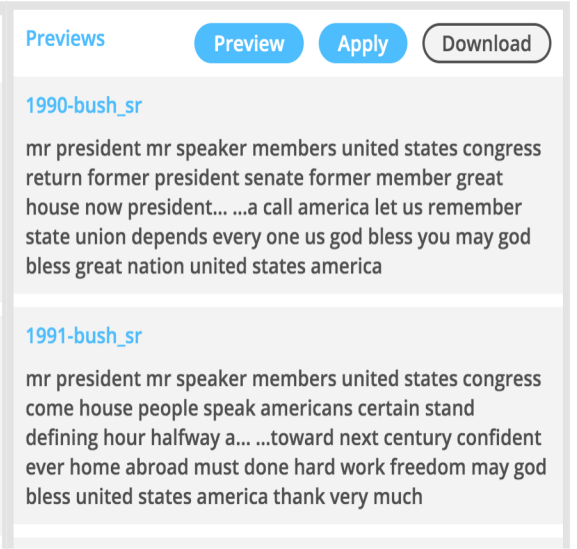
Lexos: Applying your Preparations

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.

BEFORE PREP



AFTER PREP



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

Lexos: Visualize

Word Cloud: visualize a wordcloud across the entire corpus.

Multi Cloud: visualize wordclouds for each individual document/text

Northeastern University
NULab for Texts, Maps, and Networks

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

Lexos: Rolling Window

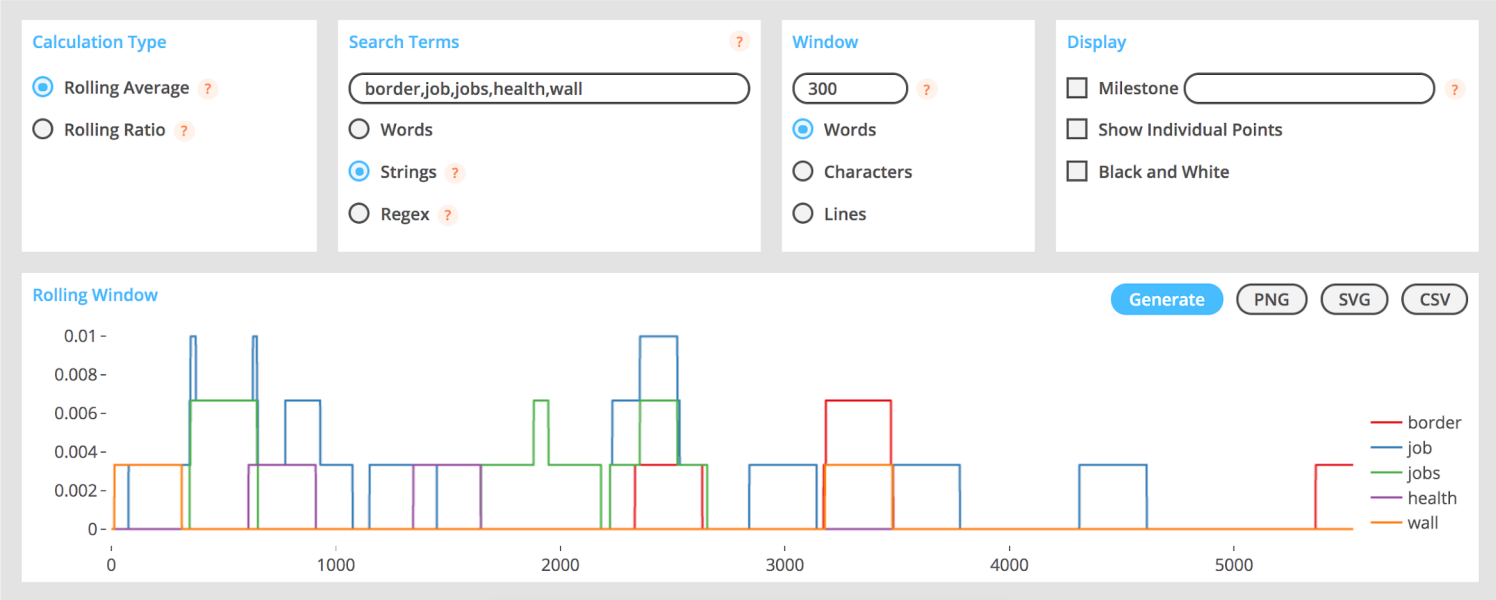
Rolling windows allow you to look at word trends across **one** document. To use a rolling window:

1. Go to “Manage” and right click one blue dot. Click “Deactivate all”
2. Choose the one **document** you want to analyze with the rolling window
3. Go back to “Visualize-> Rolling Window” and type in a search term you want to visualize. You can also search multiple terms by clicking “String” and separating words with a comma (president,health,republican)
4. Choose a Window size (the number of words each “window” contains). For shorter documents, it’s good to have a number like 300/500. For larger documents, you may want to make your window larger. Play around with the window size until you get a visualization that makes sense.
5. Click “Generate”

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

Lexos: Rolling Window Results

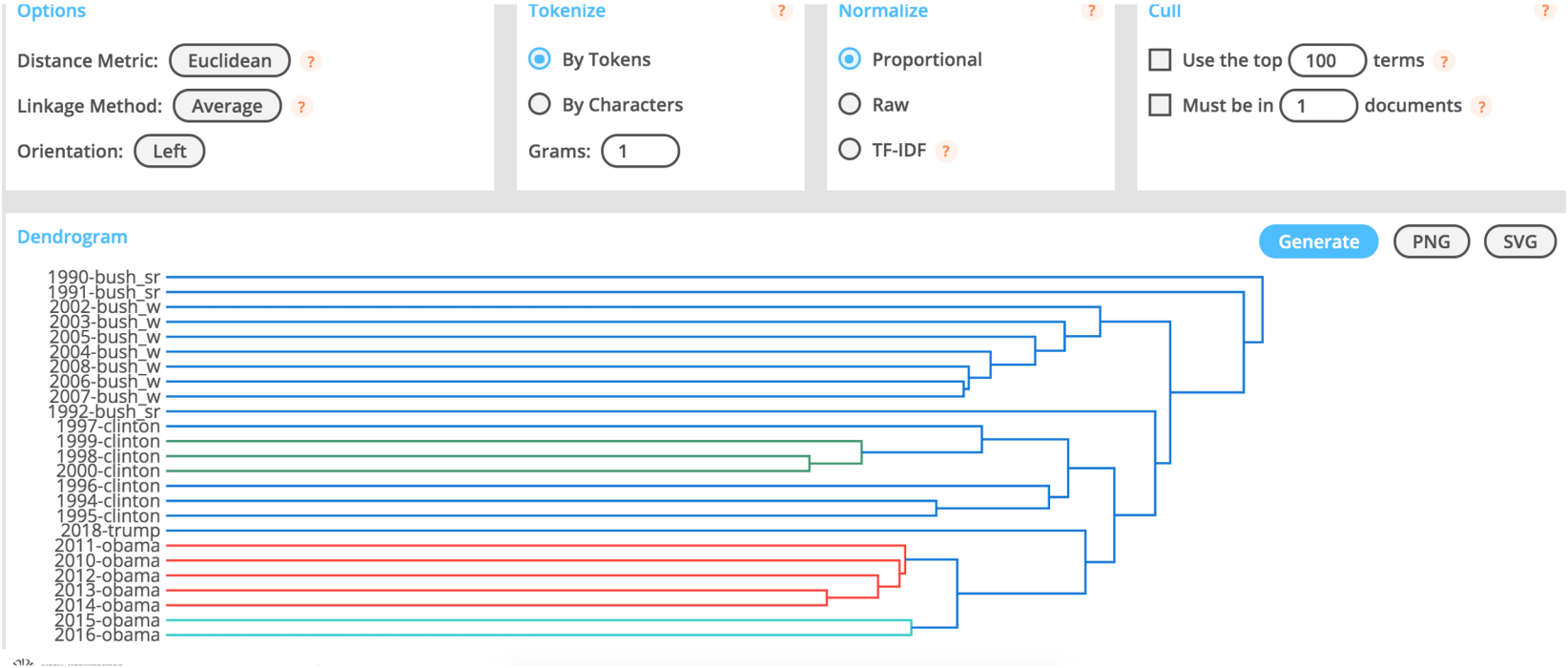
Using Trump's State of the Union from 2017, and searching for the strings "border,job,jobs" with a window of 300, we can get an idea of how different terms work together in Trump's speech. You may also be interested in **contrasting** terms to see how they're used across a text.



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

Lexos: Analyze, Dendrogram

The dendrogram demonstrates similarity between the different documents.



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 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!

Voyant

Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant: https://voyant-tools.org/

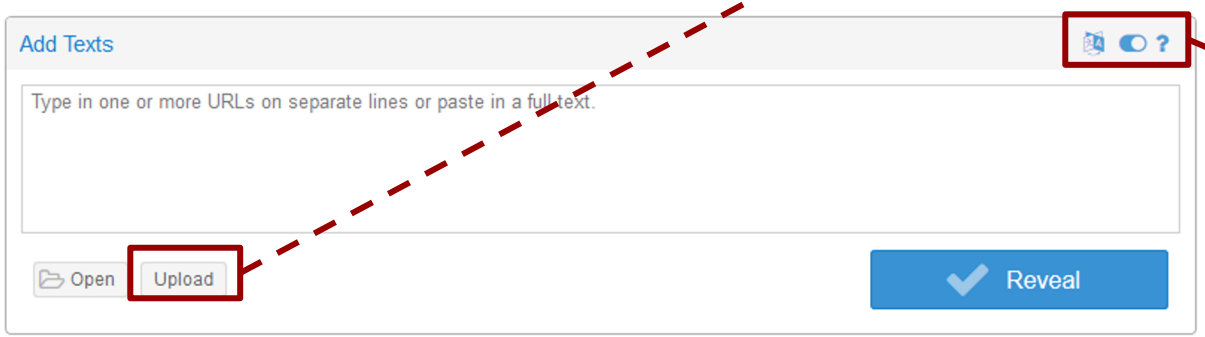
Voyant makes it possible to perform analyses on one or multiple files in many ways, including word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances. It also makes nice visualizations!

Click “Upload” and choose all the texts you want to analyze.

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

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

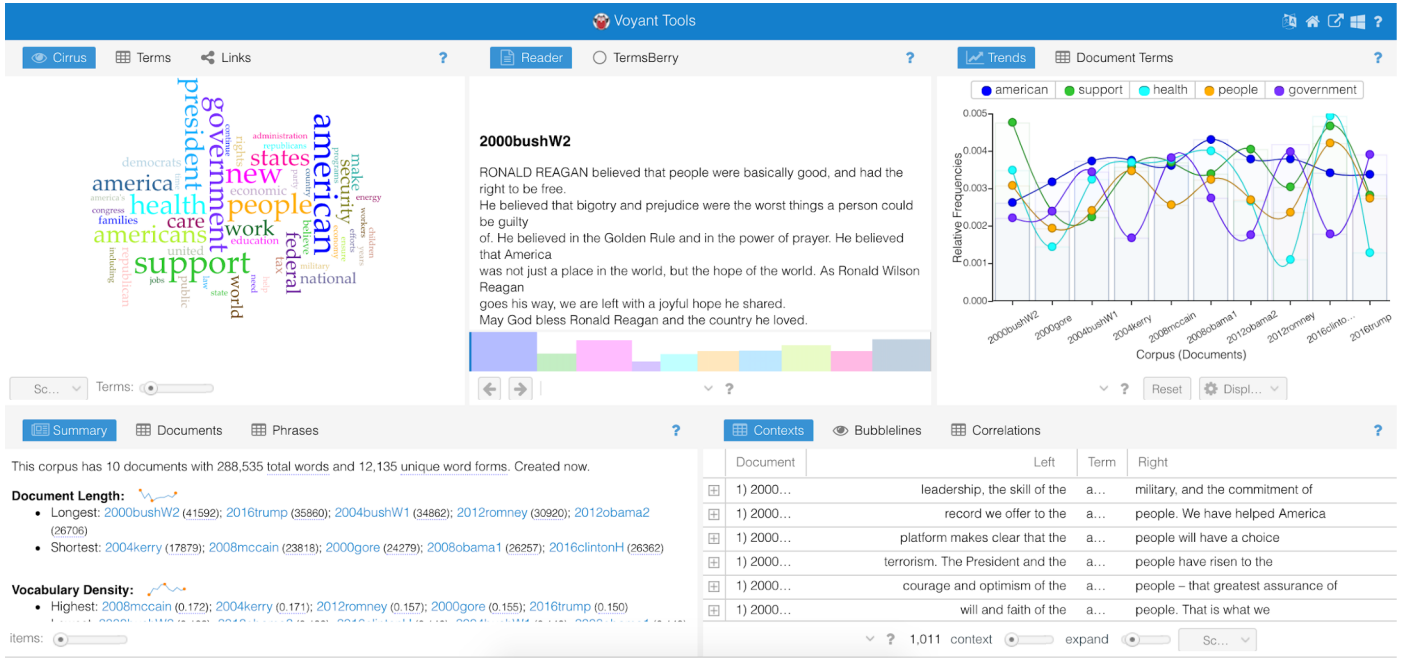
Voyant: Understanding the Dashboard

Results:

From a corpus of political party platforms you can see the default results page with multiple panes:

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

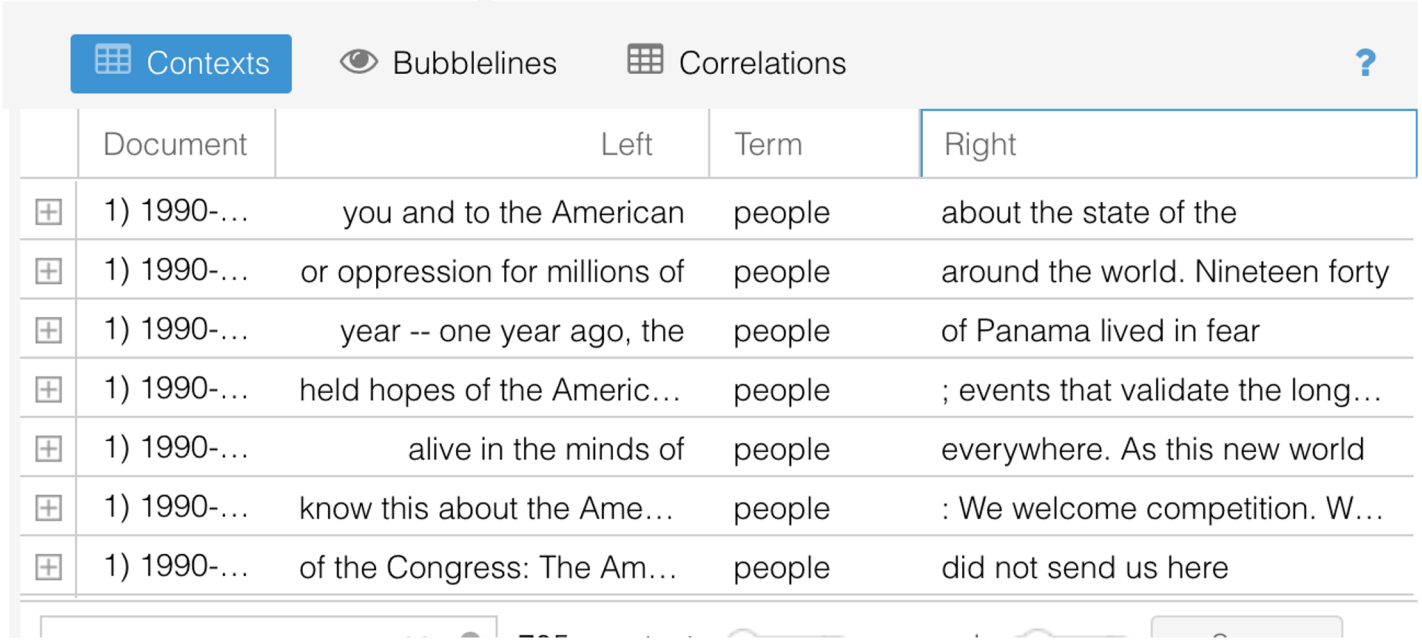
These boxes can all be changed!



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

Voyant: Contexts (concordances)

Contexts, or concordances, show the different contexts around particular search terms. For example, you can see all the times the word “president” appears in the corpus, the documents it appears in, and the contexts in which it appears.

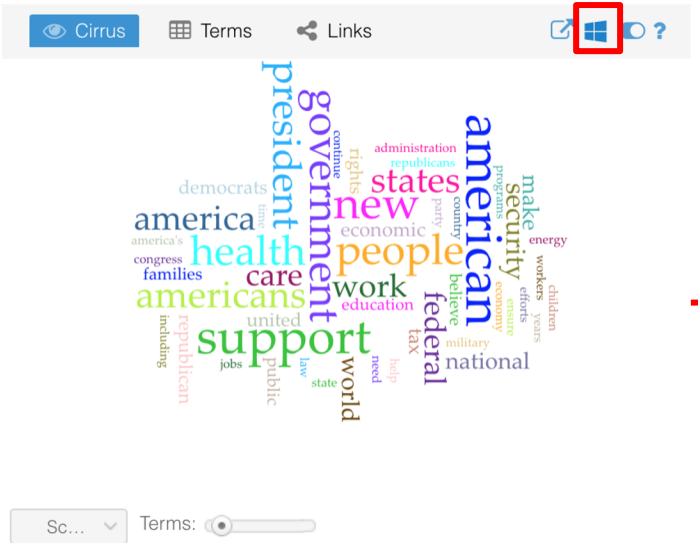


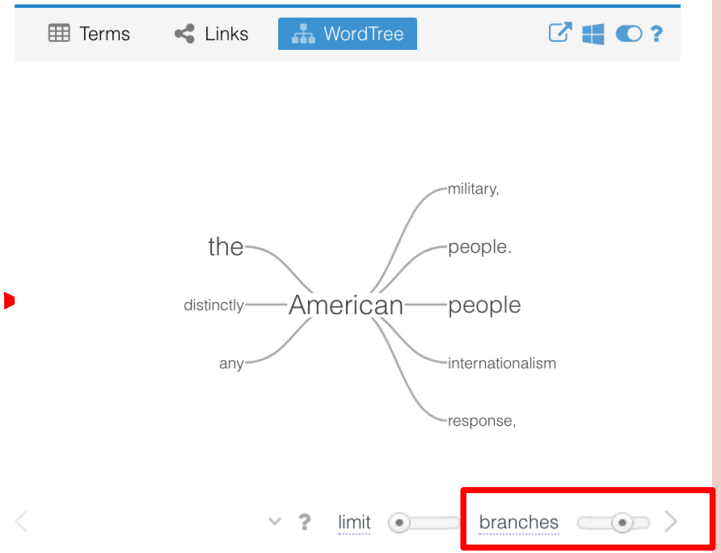
Document	Left	Term	Right
1) 1990-…	you and to the American	people	about the state of the
1) 1990-…	or oppression for millions of	people	around the world. Nineteen forty
1) 1990-…	year -- one year ago, the	people	of Panama lived in fear
1) 1990-…	held hopes of the Americ…	people	; events that validate the long…
1) 1990-…	alive in the minds of	people	everywhere. As this new world
1) 1990-…	know this about the Ame…	people	: We welcome competition. W…
1) 1990-…	of the Congress: The Am…	people	did not send us here

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

Voyant: Changing displayed results

Select the panes button and choose a new option from the dropdown menu





For our new pane option, we have chosen the WordTree visualization from the 'visualization tools' dropdown sub-menu. You can select the number of "branches" by dragging the scroll button at the bottom.

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

Your Turn!

Using the corpus you prepared for today—the three articles from your field—begin practicing web-browser text analysis

- Follow the "How to Build a Corpus" steps to create your corpus
- Prep your corpus using Lexos. Which preparation steps did you choose and why?
  - See what happens if you keep the stopwords. What are some of the most-used verbs and pronouns?
- Explore different Lexos features.
- Explore different Voyant features.

Slides, handout, and data: **https://bit.ly/diti-fall2020-cilke**

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

Post-Exploration Discussion

- What do you find challenging or exciting about these tools?
- What interesting or surprising results came up?
- How might you interpret those results based on what you know about your field?
  - What language is often used in the literature?
  - What values might be reflected by this language?

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

Thank you!

If you have any questions, contact us at:

Jonathan Sullivan
DITI Research Fellow
sullivan.jonat@northeastern.edu

Colleen Nugent
DITI Research Fellow
colleen.n@northeastern.edu

Slides, handouts, and data available at **https://bit.ly/diti-fall2020-cilke**

Schedule an appointment with us! **http://bit.ly/diti-office-hours**

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