Computational Text Analysis for Content Analysis

By Dipa Desai and Hunter Moskowitz
Digital Integration Teaching Initiative (DITI)

POLS 2395 Environmental Politics

Daniel Aldrich

Spring 2024

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 (CTA).
- Discussion of CTA's applications and uses in research.
- Introduction to web-based text analysis tools.
  - Word Counter, Word Trees, Voyant, Lexos

Slides, handouts, and data available at:

https://bit.ly/sp24-aldrich-pols2395

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

What is Computational Text Analysis?

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 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!

Language used in Climate News

Word Cloud of TV News on "Global warming." Terms like "believe" and "threat" appear frequently with "global warming" in TV news coverage since 2009.

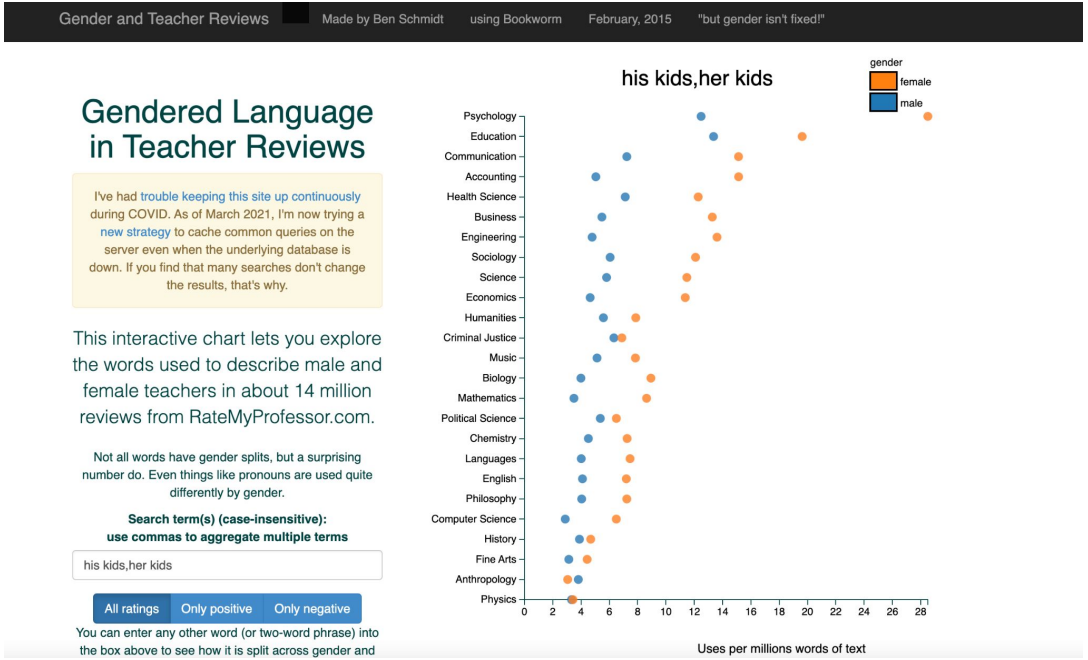
Go to the Television Explorer. Search "global warming," "climate crisis," "greenhouse effect."

- What do you notice about the TV coverage of these terms over time? What is surprising?
- How do you think political values affects climate language?
- How might this language shape policies?

Northeastern University
NULab for Texts, Maps, and Networks

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

—How do you think Schmidt determined gender for this tool?

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

Key Terms

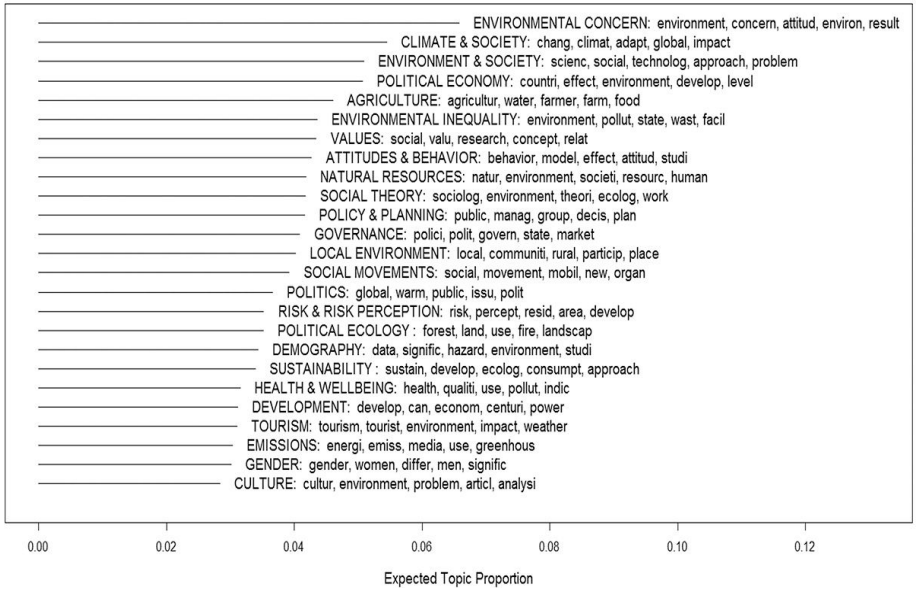
- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stop words include: a, the, she, he, 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.'
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.

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

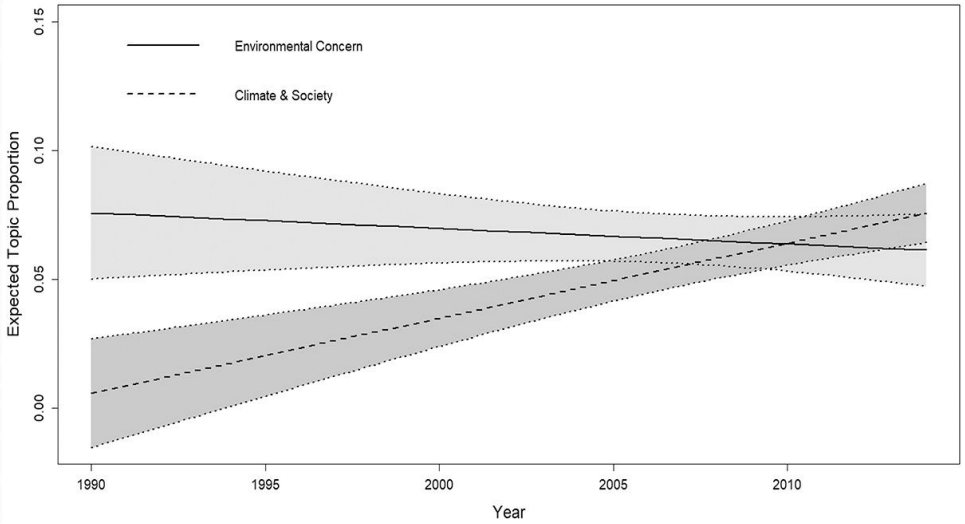
Examples from Practice

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

Key Topics in Environmental Sociology



25 topics ranked from most to least prevalent in the corpus of 815 environmental sociology articles, including the top five associated word stems. The *x*-axis represents the proportion of each topic within the overall corpus.



Topical prevalence of 'Environmental Concern' and 'Climate & Society' over time, 1990 – 2014 (with 95% confidence intervals).

Jeremiah Bohr & Riley E. Dunlap (2018) Key Topics in environmental sociology, 1990–2014: results from a computational text analysis, Environmental Sociology, 4:2, 181-195, DOI: 10.1080/23251042.2017.1393863

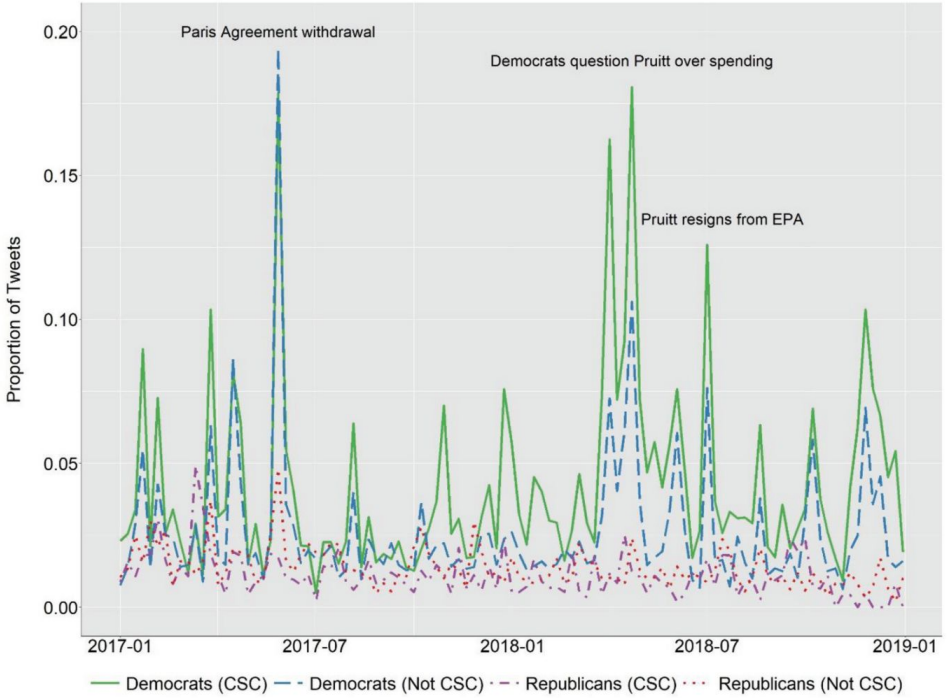
Northeastern University
NULab for Texts, Maps, and Networks

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

U.S. Environmental Politics

To what extent do politicians publicly discuss environmental issues in line with public opinion and the economic characteristics of their constituents?

- Nominally pro-environment Republicans representing more moderate constituents fail to oppose their partisan colleagues, particularly during the Trump administration's withdrawal from the Paris Agreement. At the same time, very few openly attacked climate science



Weekly proportions of tweets discussing environmental issues sent by the 115th House of Representatives.

Key events and challenges: a computational text analysis of the 115th house of representatives on Twitter - Jeremiah Bohr in Environmental Politics (2021), 30 (3): 399-422

Additional Examples

- National interests and coalition positions on climate change: A text-based analysis - Paula Castro in *International Political Science Review* (2020) ,42 (1): 95-113
- The Meaning of Action: Linking Goal Orientations, Tactics, and Strategies in the Environmental Movement - Laura K. Nelson and Brayden G King in *Mobilization: An International Quarterly* (2020) 25 (3): 315–338.

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

Text Preparation

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?

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

Preparing Your Text

1. Choose the texts or text selections that you would like to include.
2. Create a folder on your computer or cloud storage where you will store your corpus. Give it a clearly descriptive name, without spaces or special characters.
3. If you are using a text that isn't already plain text, then copy and paste your text into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
   a. Mac users, you may need to make your Text Edit into a 'plain text'. Open Text Edit, go to Preferences, and make sure "plain text" is selected
4. Save the text as a plain text file (with a .txt extension). Always make sure to name your files so you know what is in them!
5. Repeat steps above for each text in the corpus.

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

Our Text

Our text is plain text (.txt file) of President Joe Biden's speech at COP27 Climate Summit 2022 in Egypt. The primary objective is to explore this text using web-based computational text analysis tools.

We will also use the speeches of climate activist Leah Namugerwa and Kausea Natano, Prime Minister of Tuvalu to see how a corpus can be analyzed. The primary objective is to compare and contrast the three speeches.

(Tuvalu: An island nation in Oceania)

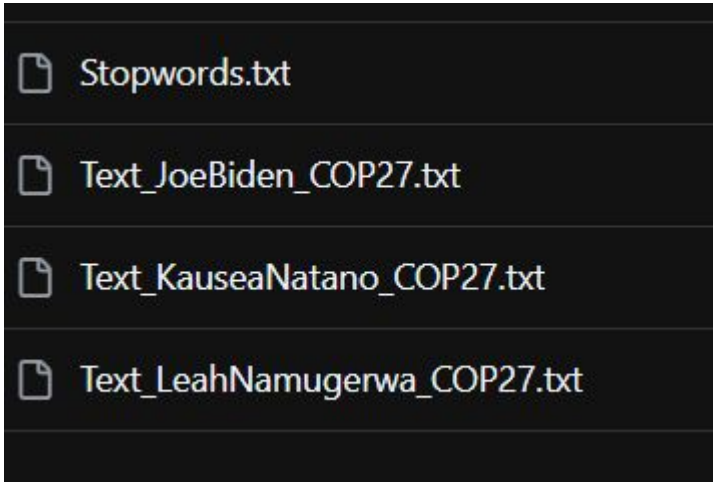
Feel free to ask questions at any point during the presentation!

Sample Corpus

The following .txt files are available on:

http://bit.ly/sp24-aldrich-pols2395

- For each file, click “Raw” in the top right corner.
- Right-click (PC) or Ctrl-click (Macs) on the text and choose “Save As.”
- Save as a .txt file on your computer.



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

Exploratory Tools: Word Counter and Word Tree

Northeastern University
NULab for Texts, Maps, and Networks

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

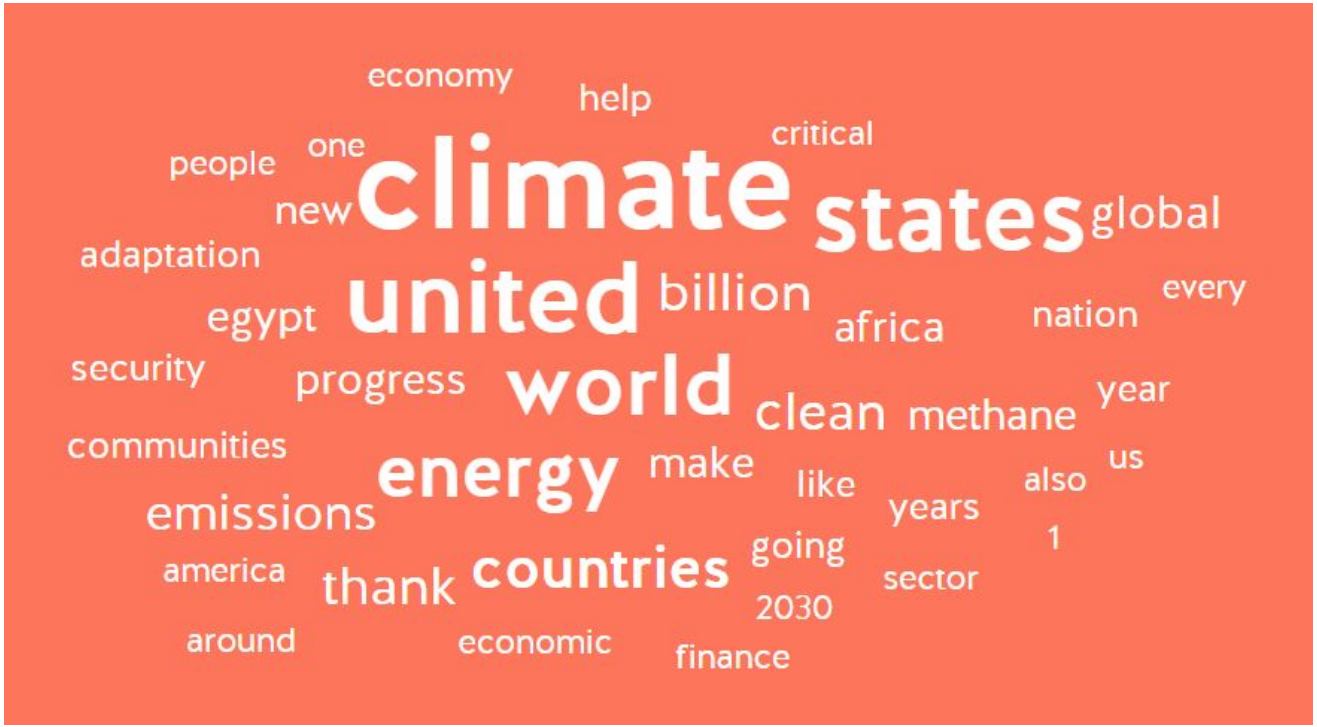
Word Counter

- https://databasic.io/en/wordcounter/
- A user-friendly **basic word counting tool**
- It 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 remove stopwords, but you can control these options

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

Word Counter Examples

Word Counter will show you a word cloud, which can give you a sense of the **most used words in a document**. Words used more often are bigger, and ones used less often are smaller.



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

Word Counter Examples

TOP WORDS

Word	Frequency
climate	35
united	27
states	27
world	22
energy	19
countries	14
thank	12
clean	12
billion	12
emissions	11
global	10

Shows the top words in the text.

Stopwords aren't removed for the bigrams and trigrams because they need context.

BIGRAMS

bigram	Frequency
the united	27
united states	27
we re	20
the world	19
it s	14
of the	14
in the	13
and the	12

TRIGRAMS

trigram	Frequency
the united states	27
in the united	5
around the world	5
that s why	5
thank you thank	4
you thank you	4
the climate crisis	4
united states is	4

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

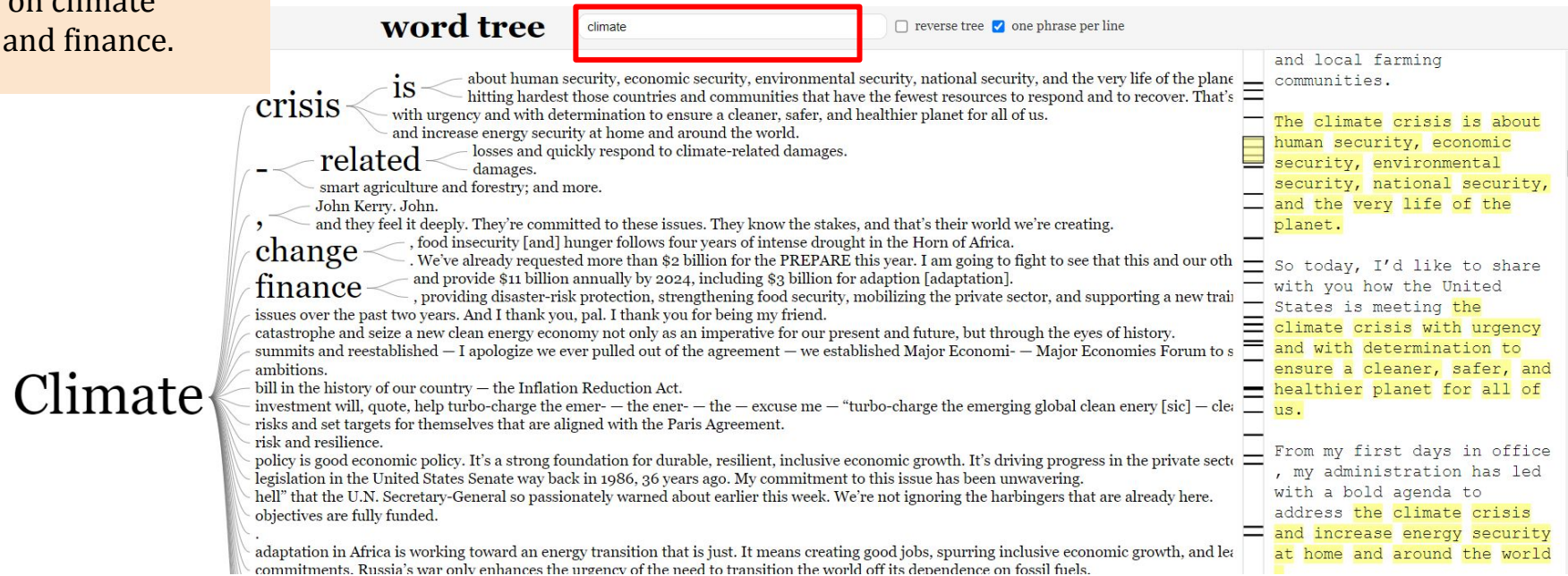
Word Tree

- 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.
- Upload your text, enter a keyword or phrase to search, then try reversing the tree.
- It's often useful to search frequent terms identified by WordCounter

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

Word Tree Example

Reflects the focus of the speech on climate change and finance.



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

Word Tree: Reverse Trees

It is worth reversing the tree to see the words that often precede it. To do this click “reverse tree” next to the search bar.

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 Word Counter and Word Tree!**

Discussion Prompts

- What limitations are you observing?
- Even with these limitations, how can you apply these tools in your research of environmental issues?
- What types of text would be interesting to explore with these tools?

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

Powerful Platform: Voyant

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. It also makes nice visualizations!

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 document you wish to analyze.

Alternatively, insert URLs or full text into the textbox.

Click here for help and advanced options

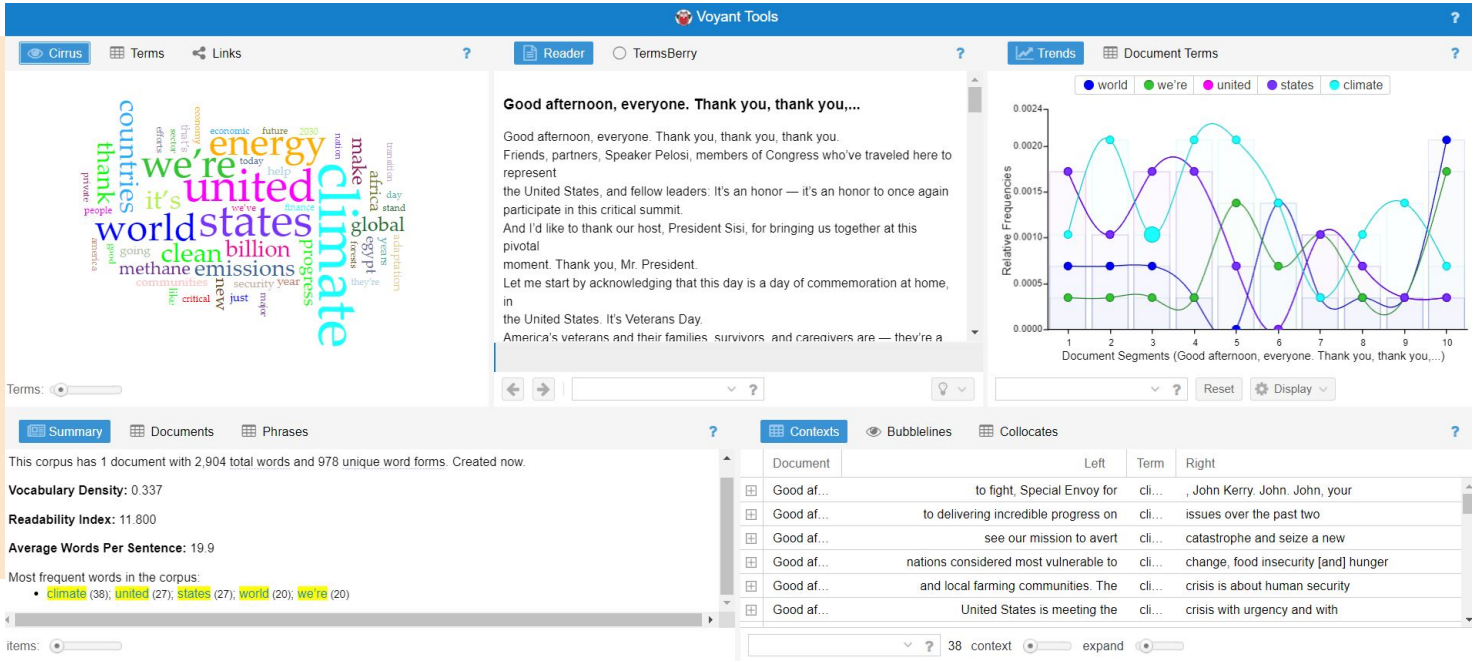
Feel free to ask questions at any point during the presentation!

Voyant: Basic Dashboard

You can see the default results page with multiple panes:

- A word cloud
- Reader section
- Trends
- Document summary
- Word contexts

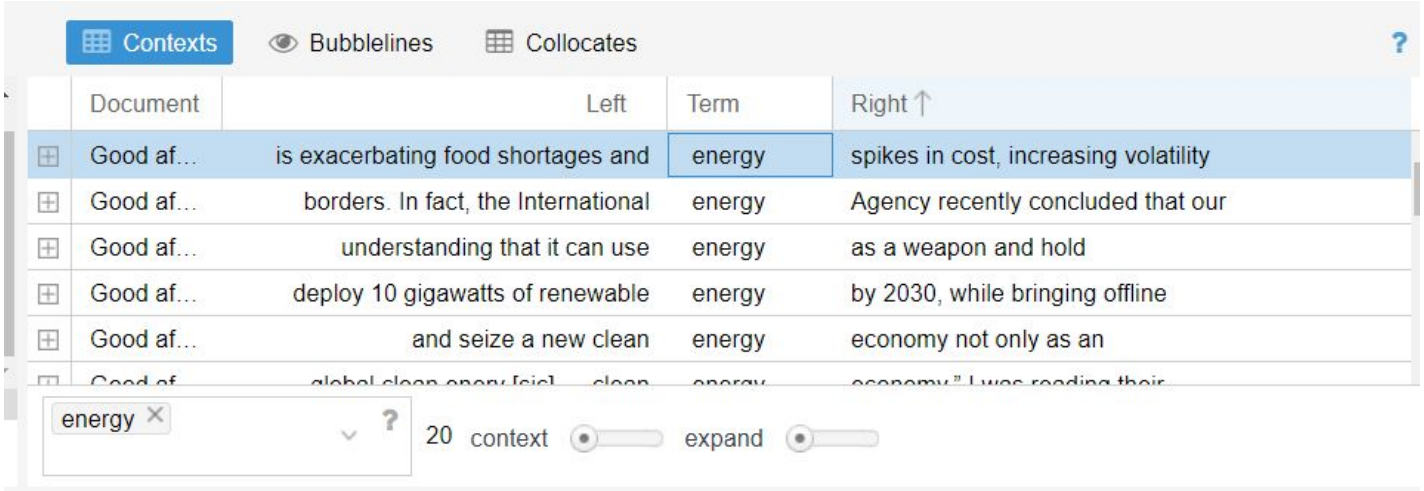
These boxes can all be changed!



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 “energy” appears in the text and the contexts in which it appears.



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

Voyant: Changing displayed results

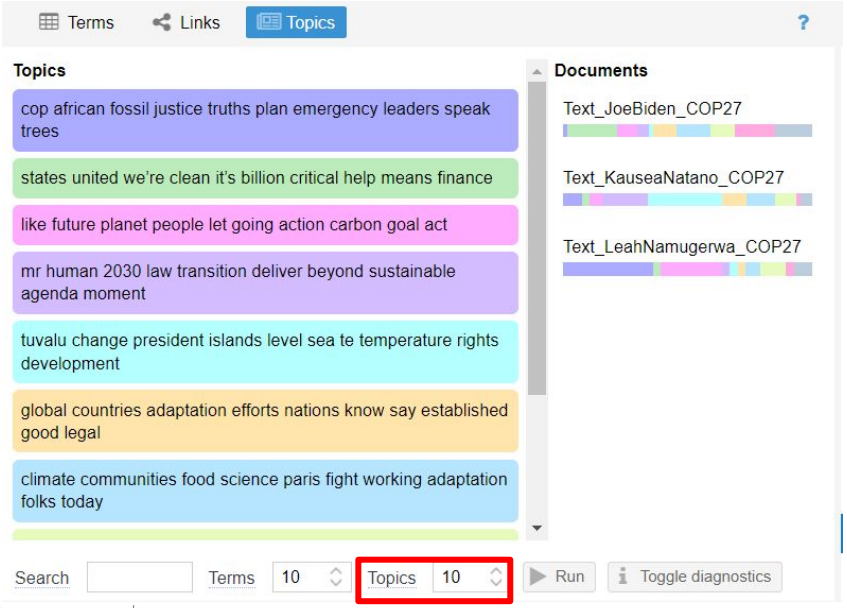
Hover on the right top corner of a pane and buttons will appear. Select the panes button and choose a new option from the dropdown menu

For 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!

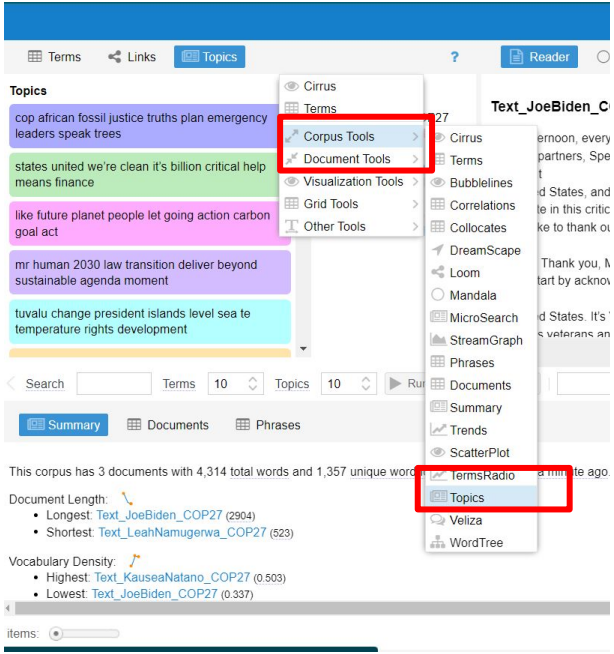
Voyant: Topics tool

You can view major topics across the corpus or individual documents by hovering over the windows icon and choosing the Topics tool under Corpus or Document tools.



From the output we can see that the topic with words like "justice, truths, emergency" is in the speeches of President Natano and activist Leah Namugerwa.

Try changing the number of topics to see how this changes the results.

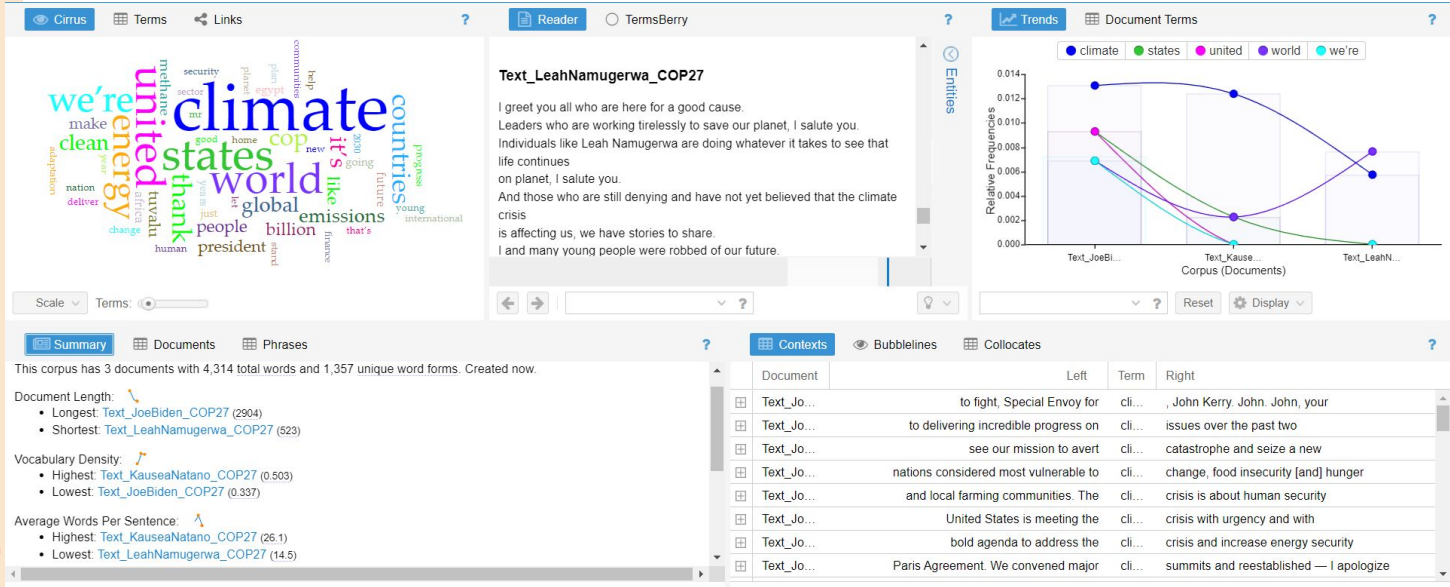


Feel free to ask questions at any point during the presentation!

Voyant: Corpus Dashboard

Results page of the corpus containing three speeches

- A word cloud: combining all texts
- Reader section: scroll down all texts
- **Trends: relative frequency of terms across all texts - good for comparison**
- **Document Summary- good for comparison**
- Word Contexts: separate for all texts



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 Voyant's features!**

Discussion Prompts

- What interesting or surprising results came up?
- How do you interpret those results based on what you know about current climate and energy talks?
- If you wanted to study an issue like drinking water pollution in the US, what kinds of documents and texts would be useful to compare?

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

Powerful Platform: Lexos

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

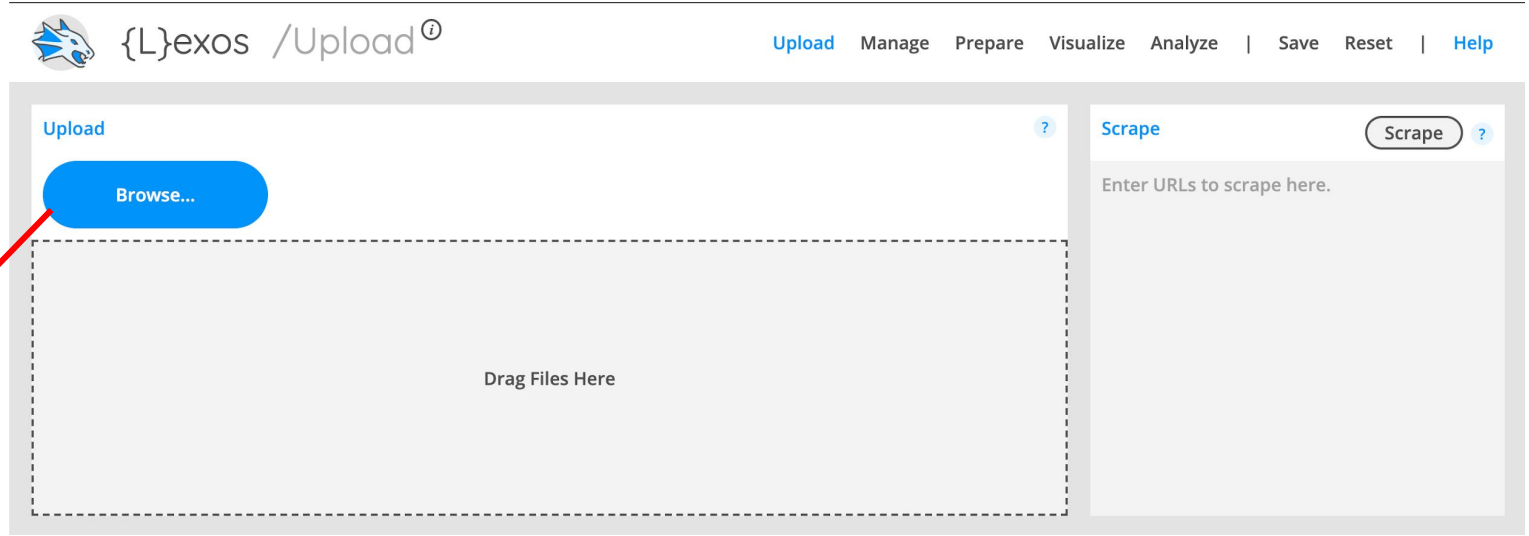
http://lexos.wheatoncollege.edu/upload

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)

You will not get a super visible notification 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)

{L}exos /Manage

Upload | Manage | Prepare | Visualize | Analyze | Save | Reset | Help

Active	#	Document	Class	Source	Excerpt	Download ?
	1	Text_JoeBiden_COP27		Text_JoeBiden_COP27.txt	Good afternoon, everyone. Thank you, thank you, thank you. Friends, partners, Speaker Pelosi, members of Congress who've trave... ...e're working toward. And we can do it together. I am confident. Thank you, thank you, thank you. And may God bless you all.	
	2	Text_KauseaNatano_COP27		Text_KauseaNatano_COP27.txt	Mr. President/Chairman, Distinguished delegates, let me thank you for your astute leadership in guiding our 27th Conference of... ...erbates many other development challenges. I thank you for this opportunity and wish COP 27 all the success. Tuvalu mo te Atua.	
	3	Text_LeahNamugerwa_COP27		Text_LeahNamugerwa_COP27.txt	I greet you all who are here for a good cause. Leaders who are working tirelessly to save our planet, I salute you. Individuals... ...ure. The time to force action is right now and right here at the African COP I believe that we can do this TOGETHER. Thank you.	

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

Lexos: Prepare (scrub)

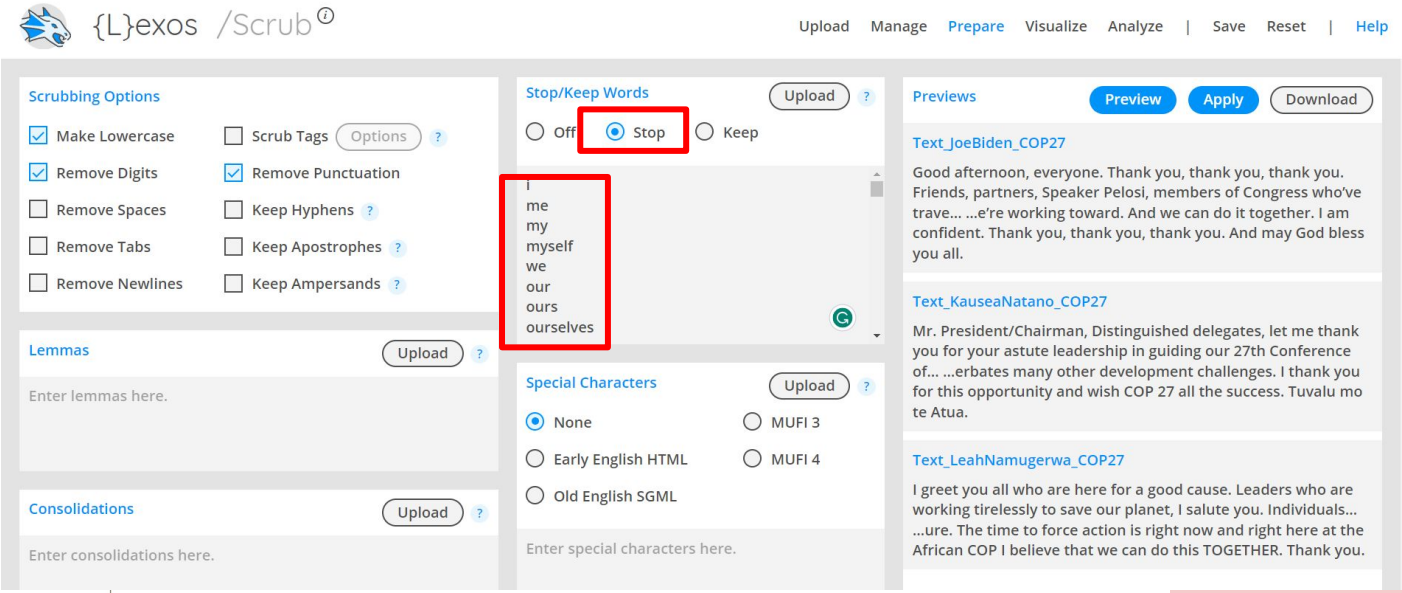
Lexos demonstrates some more advanced options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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 (or keep only words from a list). Usually you would remove **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

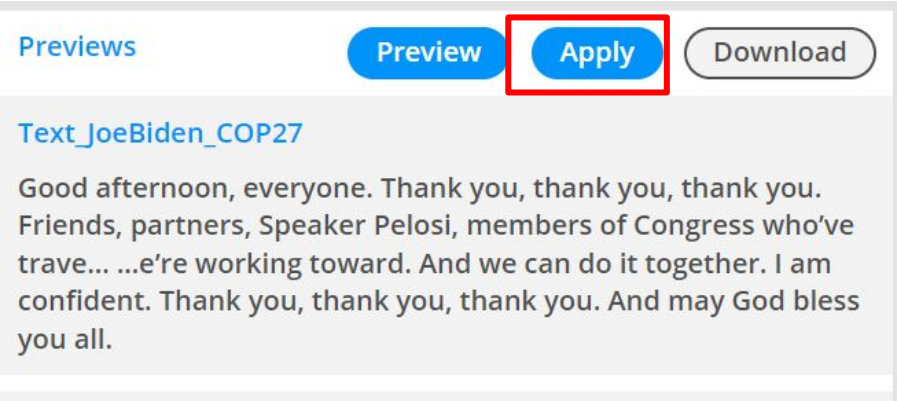
Get a list of English stopwords here: https://gist.github.com/sebleier/554280 (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the "Stop/Keep Words" box then select "Stop"



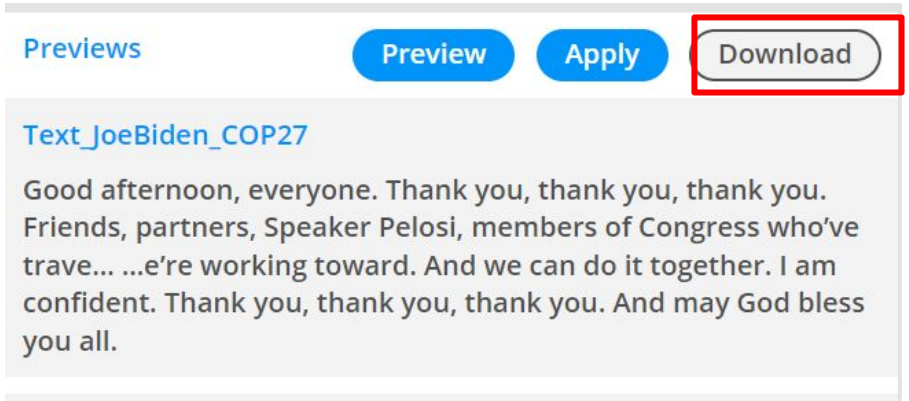
Feel free to ask questions at any point during the presentation!

Lexos: Applying your Preparations

BEFORE PREP



AFTER PREP



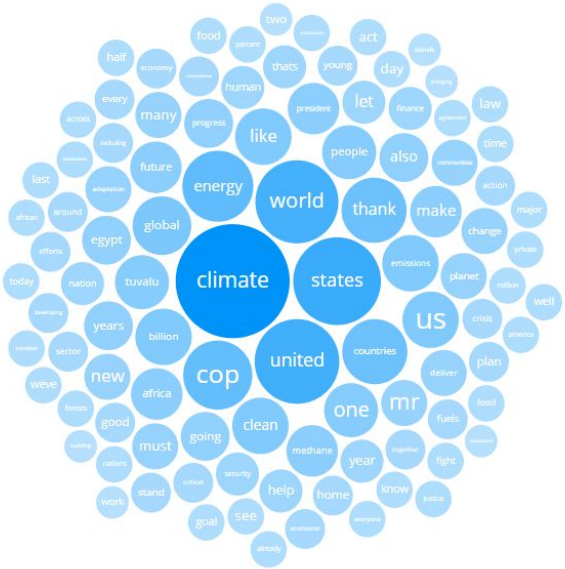
Once you have made decisions about your preparations, click "**Apply**" and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your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

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

Lexos: Visualize

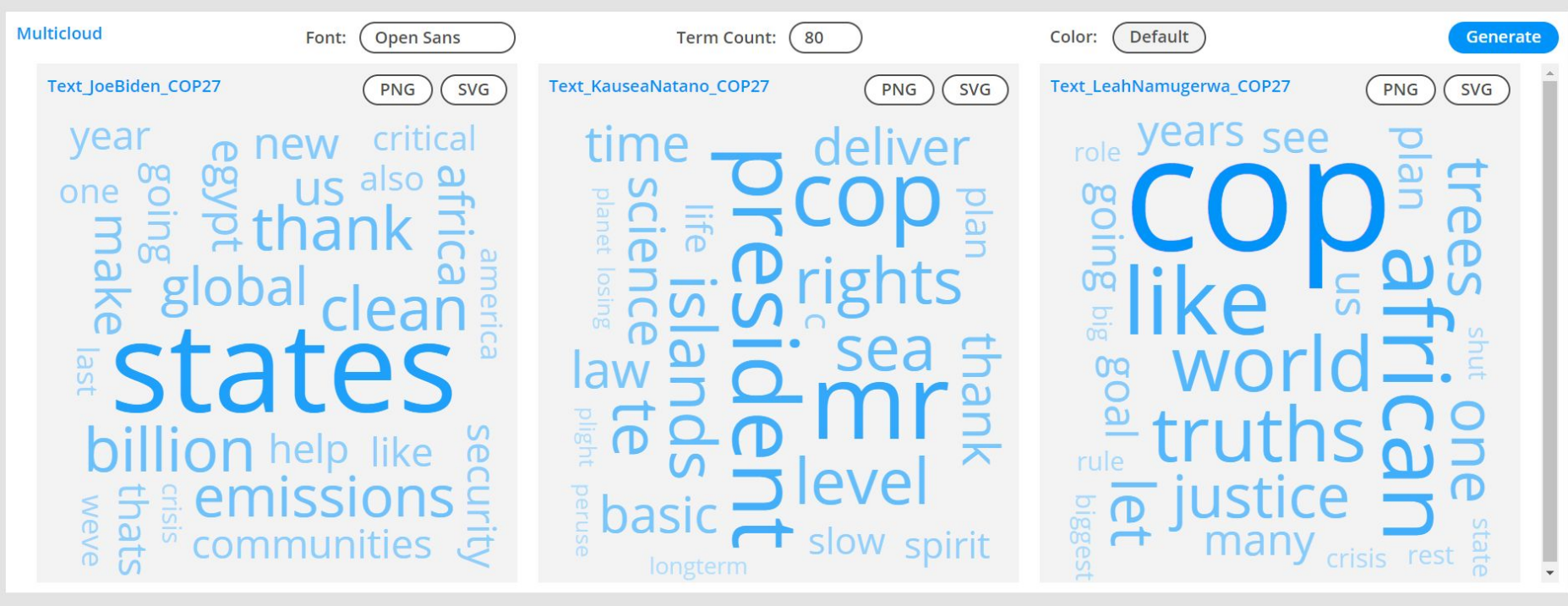
Bubbleviz: visualize word counts through bubbles across the entire text/corpus.

Word Cloud: visualize a wordcloud across the entire text/corpus.



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

Lexos: Visualize > Multicloud



Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant vs. Lexos: Wordclouds

How does the Voyant wordcloud below compare to the own made using Lexos?

Lexos Wordcloud

What could be causing this distinction? This helps demonstrate the importance of understanding what a tool is doing to the texts in the background.

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

Lexos: Rolling Window

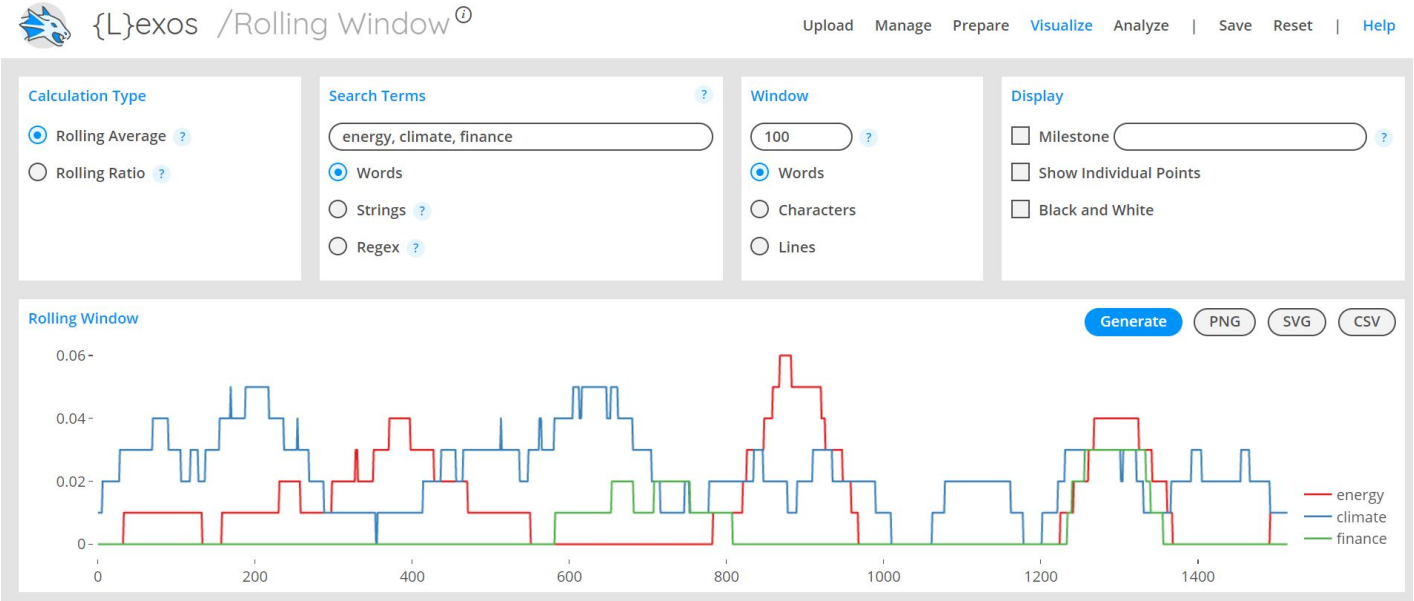
Rolling windows allow you to look at word trends across **one** document. To use a rolling window, first select a single text in the "Manage" screen, then:

1. Go 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.
2. 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.
3. Click “Generate”

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

Lexos: Rolling Window Results

Using Joe Biden's speech, and searching for the words 'energy', 'climate' and 'finance' with a window of 100 (since this is a small document), we can get an idea of how these terms work together in the report.



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

Lexos: Analyze > Top Words

The top words tool lets you compare word usage between individual documents and your corpus as a whole. If you want to make more specific comparisons, you can also assign “classes” to subsets of tools with the “Manage” screen.

- Words with high positive scores are **used more often** in each document, relative to the rest of the corpus.
- Words with high negative scores are **used less often**.

Hit the “Generate” button to see the top words for your texts.

Lexos: Analyze > Top words

Top Words

Generate Download ?

Document "Text_JoeBiden_COP27" Compared To The Corpus

tuvalu	-2.7333
cop	-2.4436

Document "Text_KauseaNatano_COP27" Compared To The Corpus

tuvalu	3.9876
mr	2.9494
united	-2.4199
islands	2.3988
level	2.3988
rights	2.3988

Document "Text_LeahNamugerwa_COP27" Compared To The Corpus

african	4.3036
cop	4.1579
truths	3.8478
speak	3.331
trees	3.331
leaders	2.9602

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

Lexos: Analyze > Dendrogram

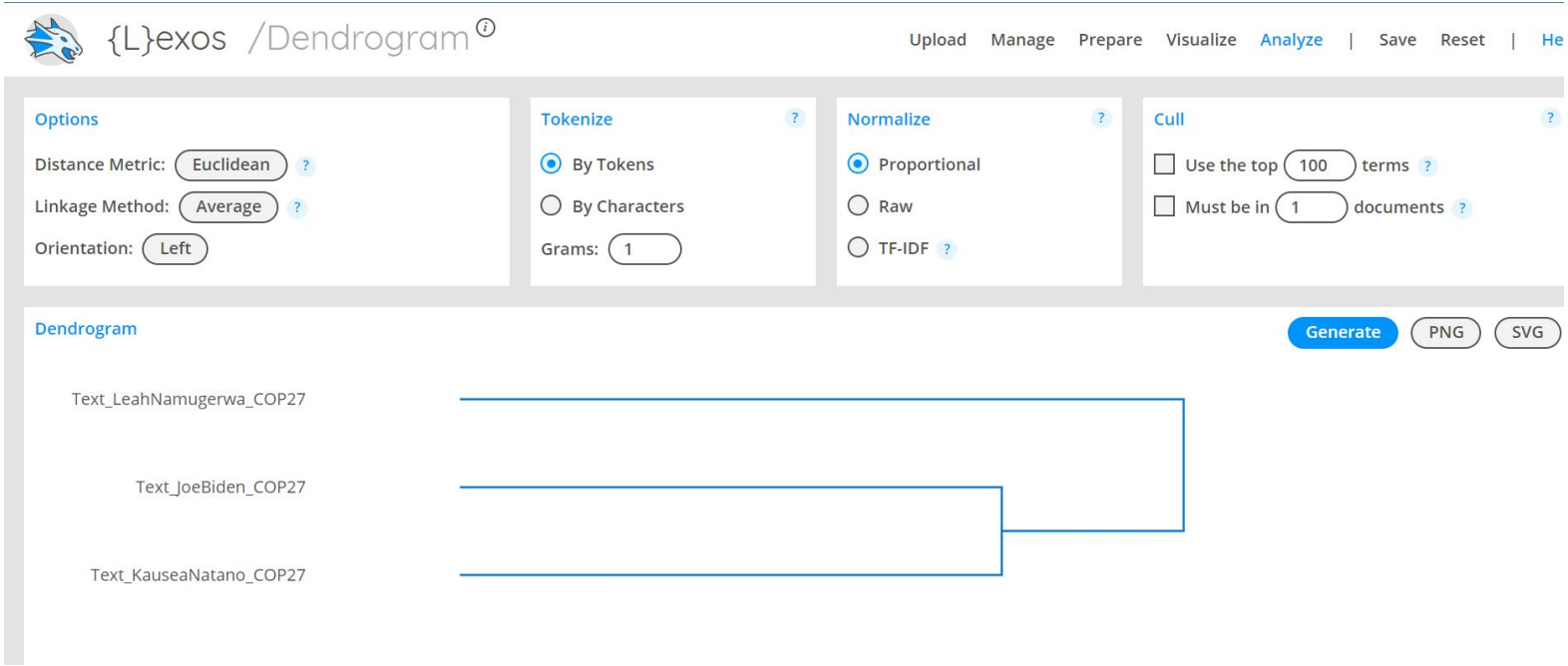
The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms are able to show the hierarchy between objects. Dendrograms show:

- 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: Dendrogram

The dendrogram demonstrates similarity between the different documents.



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.

Your Turn!

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

Discussion Prompts

- What interesting or surprising results came up?
- If you wanted to study an issue like drinking water pollution in the US, what kinds of documents and texts would be useful to compare?
- Between Voyant and Lexos, which tool did you prefer and why?
- Which features do you think will be useful in your analysis?

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

Developed by Dipa Desai, Vaishali Kushwaha, and Garrett Morrow
Delivered by Dipa Desai and Hunter Moskowitz
DITI Research Fellows
Digital Integration Teaching Initiative

- Slides, handouts, and data available at **http://bit.ly/sp24-aldrich-pols2395**
- We'd love your feedback! Please fill out a short survey here: **https://bit.ly/diti-feedback**
- Schedule an office hours appointment with us! **https://bit.ly/diti-meeting**

Northeastern University
NULab for Texts, Maps, and Networks

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