Computational Text Analysis for Content Analysis

Taught By Ayah Aboelela and Sara Morrell
Digital Integration Teaching Initiative (DITI)

POLS 3482 East Asian Politics
Daniel Aldrich
Spring 2025

Northeastern University
NULab for Digital Humanities and Computational Social Science

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/sp25-aldrich-pols3482-text-analysis

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

What is Computational Text Analysis?

Northeastern University
NULab for Digital Humanities and Computational Social Science

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!

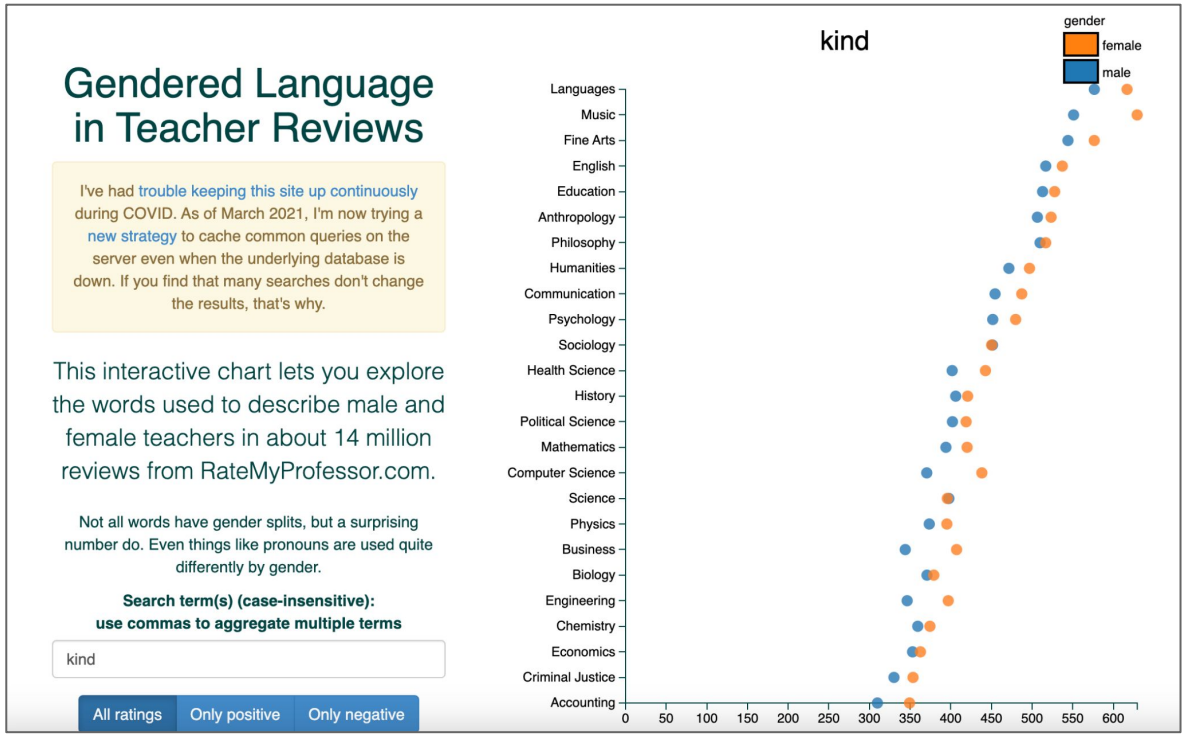
Why Computational Text Analysis?

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!

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?

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Language used in U.S. News

Word Cloud of U.S. TV News on "Japan". Terms like "Alliance" and "Security" appear frequently with "Japan" in U.S. TV news coverage since 2009.

Go to the Television Explorer. Search "Japan", "Korea", and "China".

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Key Terms (1/2)

- **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.

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

Key Terms (2/2)

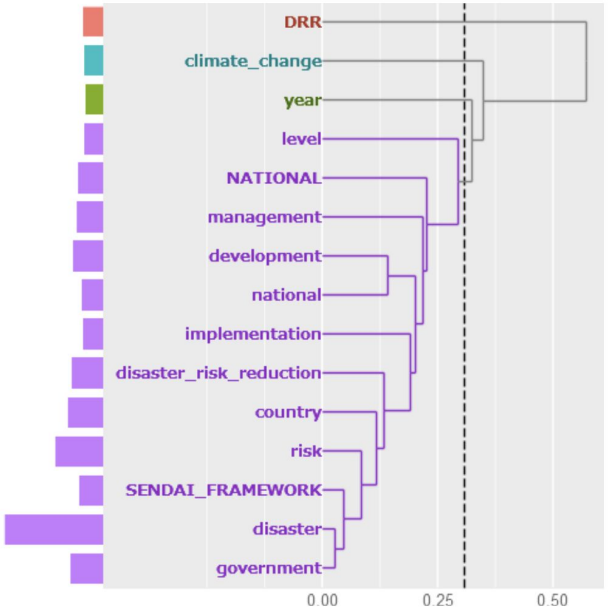
- **nGram:** A continuous sequence of *n* items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular 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!

Word Frequency and Clusters



The bars on the left-hand side indicate the TF of each word [21].

Fig. 1. Result of the hierarchical cluster analysis.

Sasaki, D. (2019). Analysis of the attitude within asia-pacific countries towards disaster risk reduction: Text mining of the official statements of 2018 Asian ministerial conference on disaster risk reduction. Journal of Disaster Research, 14(8), 1024-1029.

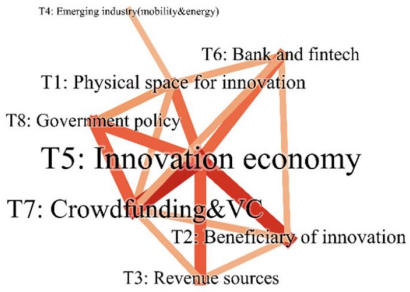
Northeastern University
NULab for Digital Humanities and Computational Social Science

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

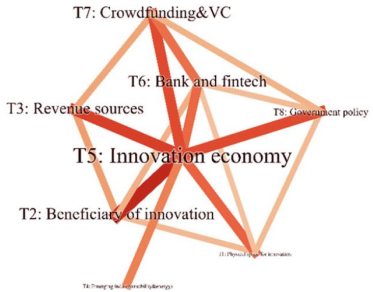
Comparative Topic Modeling

Table 11. Inter-topic linkages.

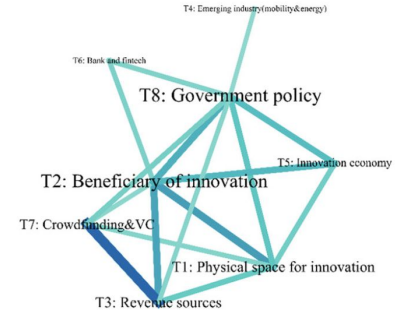
East Asia 2000s



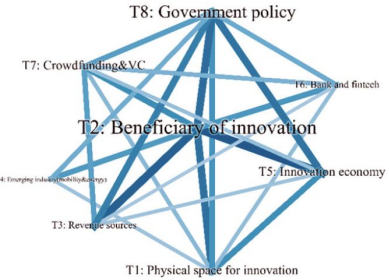
East Asia 2010s



North America 2000s



North America 2010s



Re Lee, K., Hyun Kim, J., Jang, J., Yoon, J., Nan, D., Kim, Y., & Kim, B. (2023). News big data analysis of international start-up innovation discourses through topic modelling and network analysis: comparing East Asia and North America. Asian Journal of Technology Innovation, 31(3), 581-603.

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

We will use a set of plain text news articles from the Japan Times pertaining to the environment published between 10/1/24 and 11/30/24:

- U.N. nature summit ends in limbo as countries spar over funding
- A blueprint for youth inclusion in climate diplomacy
- Japan studying measures for more renewable energy areas
- How a DJ became an unlikely champion for green farming
- The world promised to tame methane, but emissions are still rising

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

Sample Corpus

The sample .txt files are available on:

https://bit.ly/sp25-aldrich-pols3482-text-analysis

- 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!

Word Counter and Word Tree

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

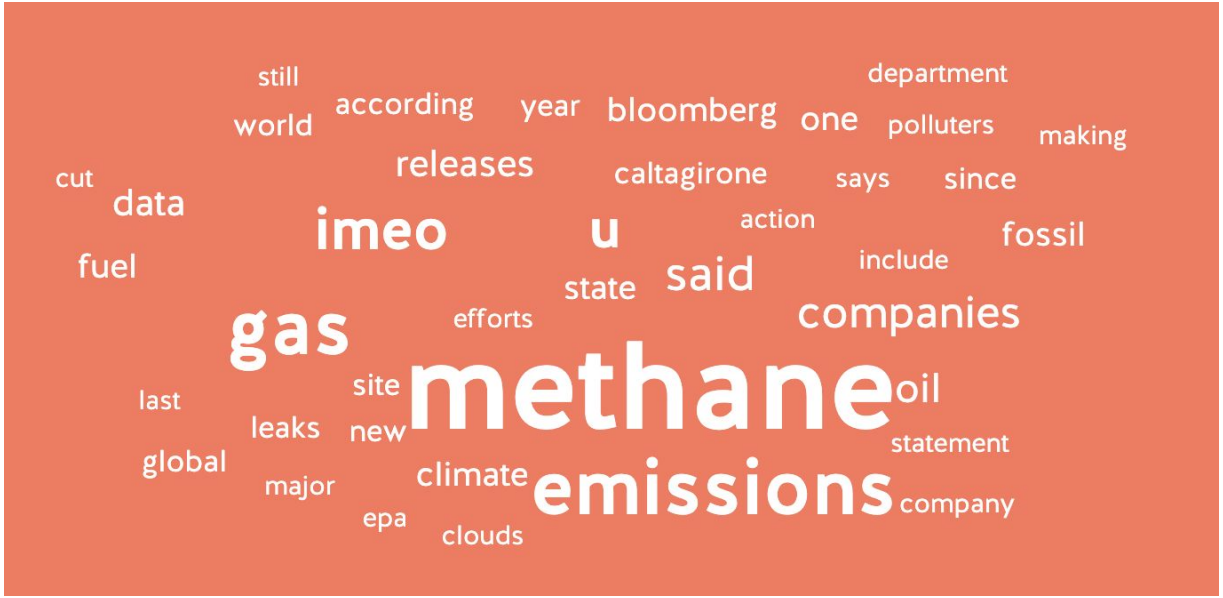
Word Counter

- https://databasic.io/en/wordcounter/
- A user-friendly **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.



[The world promised to tame methane, but emissions are still rising]

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Word Counter Examples

Shows the top words in the text.

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

TOP WORDS

Word	Frequency
methane	32
emissions	19
gas	19
imeo	13
u	12
said	10

BIGRAMS

bigram	Frequency
of the	13
in the	13
the u	9
u s	9
methane emissions	7

TRIGRAMS

trigram	Frequency
the u s	7
oil and gas	6
one of the	3
the world s	3
methane emissions from	3

[The world promised to tame methane, but emissions are still rising]

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

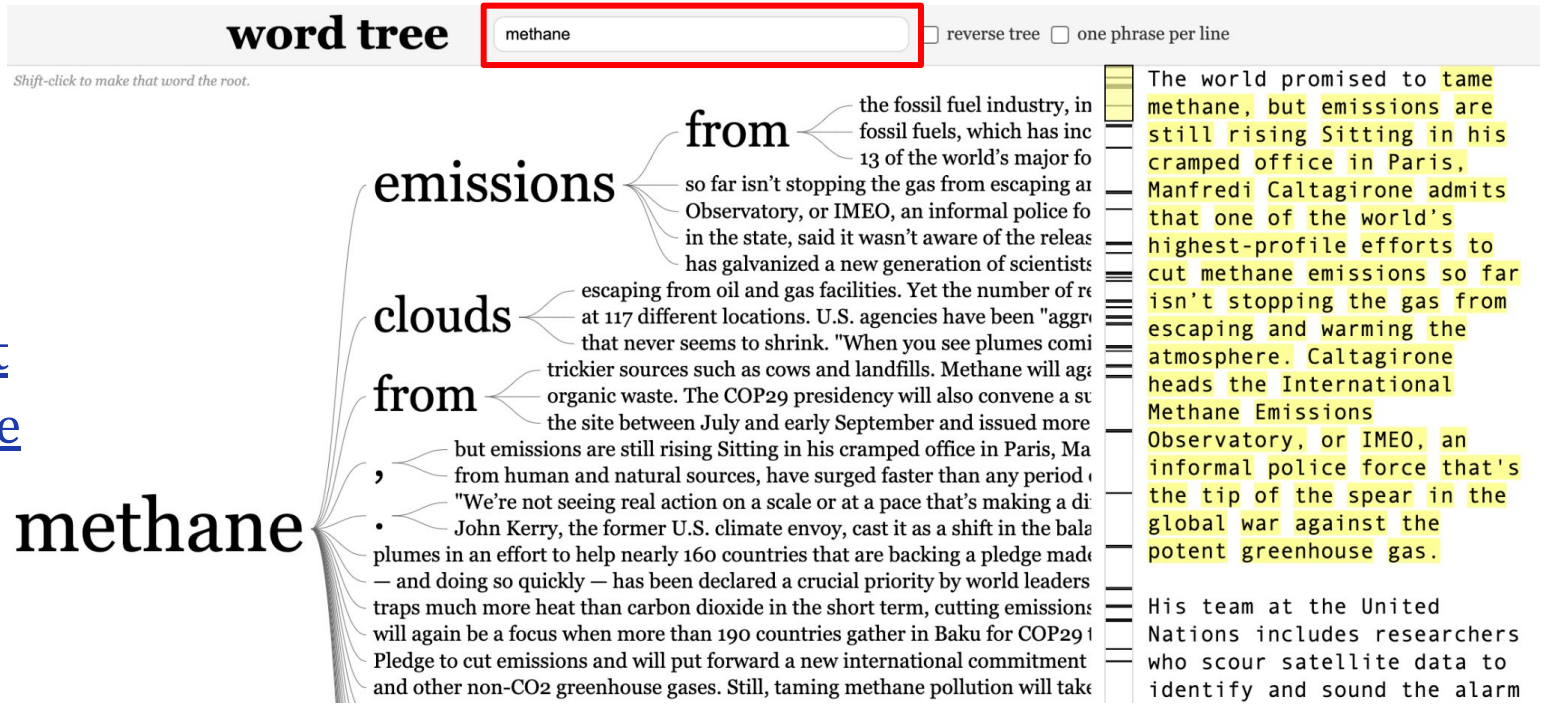
Word Tree

- 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

Text source: [The world promised to tame methane, but emissions are still rising](#)



Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

Northeastern University
NULab for Digital Humanities and Computational Social Science

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 East Asian politics?
- What types of text would be interesting to explore with these tools?

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

Voyant

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Voyant Introduction

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/

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Voyant: Upload

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.

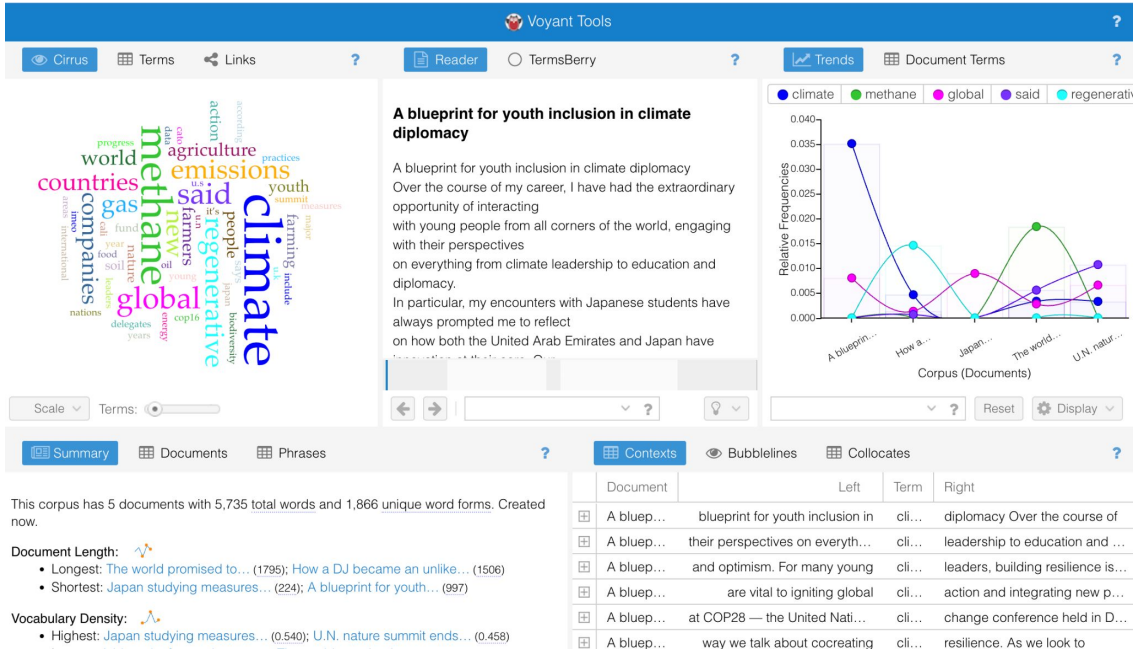
Click here for help and advanced options

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

Voyant: Corpus 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



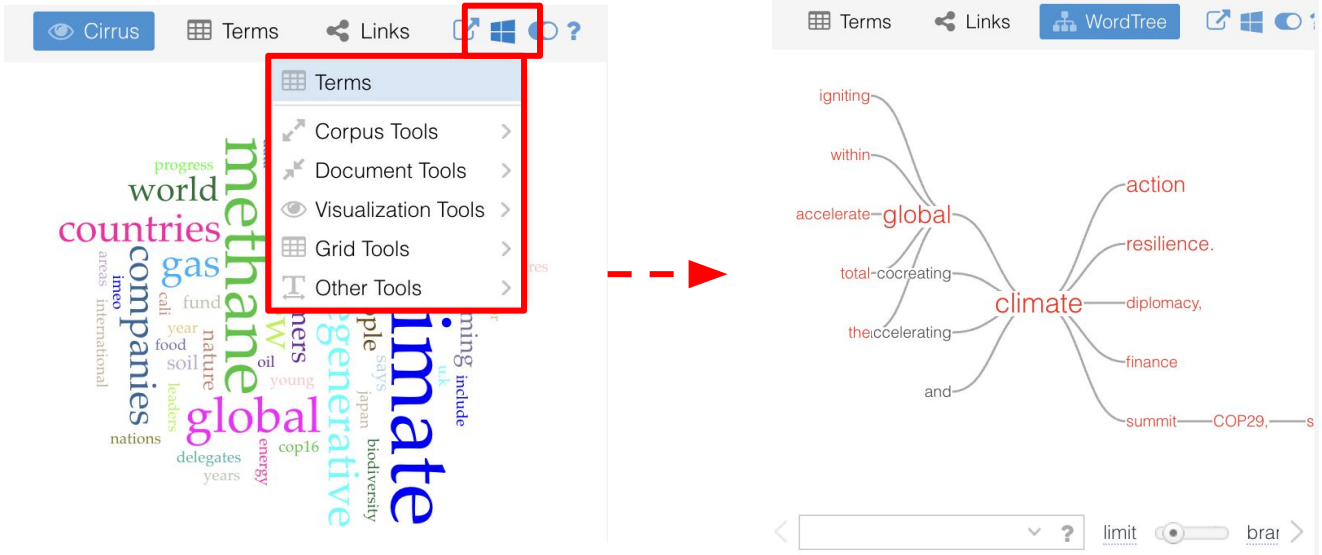
These boxes can all be changed!

Northeastern University
NULab for Digital Humanities and Computational Social Science

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



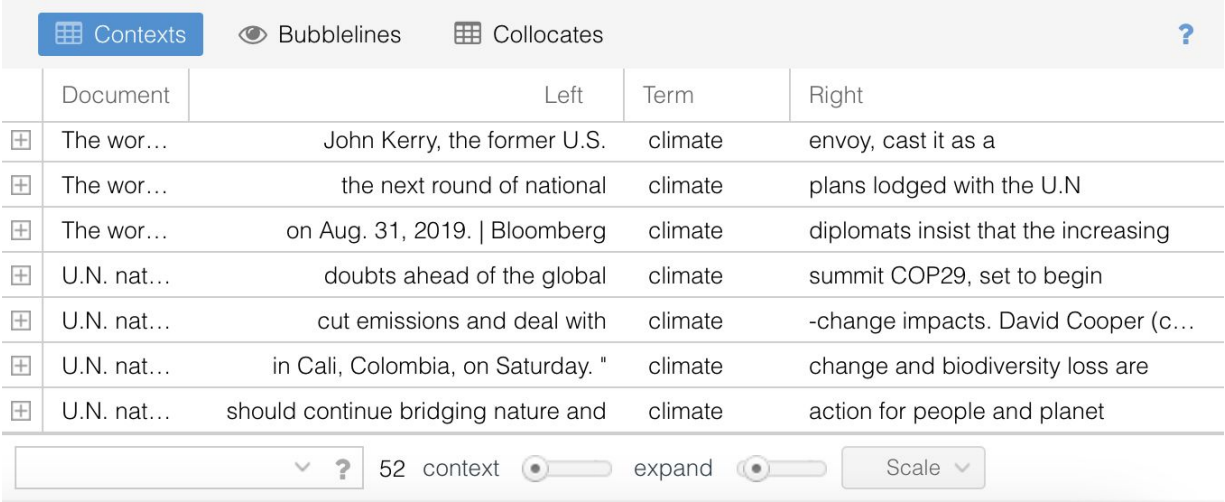
For our new pane option, we have chosen the WordTree visualization from the ‘visualization tools’ dropdown sub-menu. You can select the number of “branches” by dragging the scroll button at the bottom.

Northeastern University
NULab for Digital Humanities and Computational Social Science

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



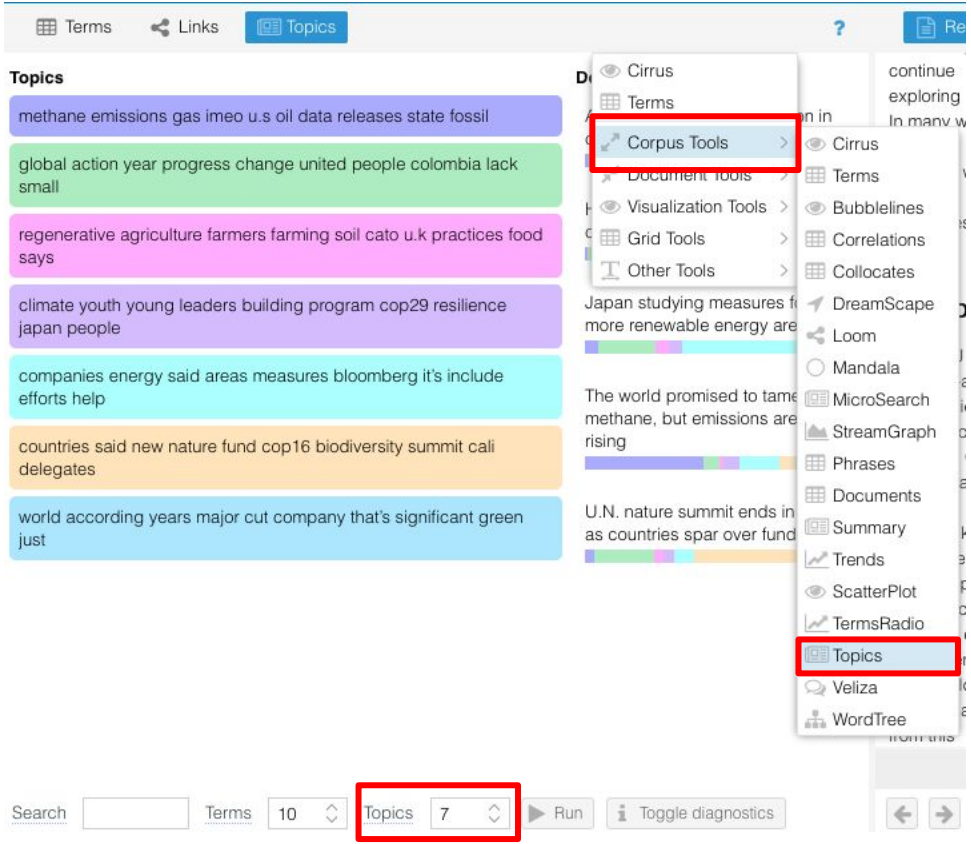
Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

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: 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 your prior knowledge?
- If you wanted to study an issue like climate diplomacy in East Asia, what kinds of documents and texts would be useful to compare?

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

Lexos

Northeastern University
NULab for Digital Humanities and
Computational Social Science

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

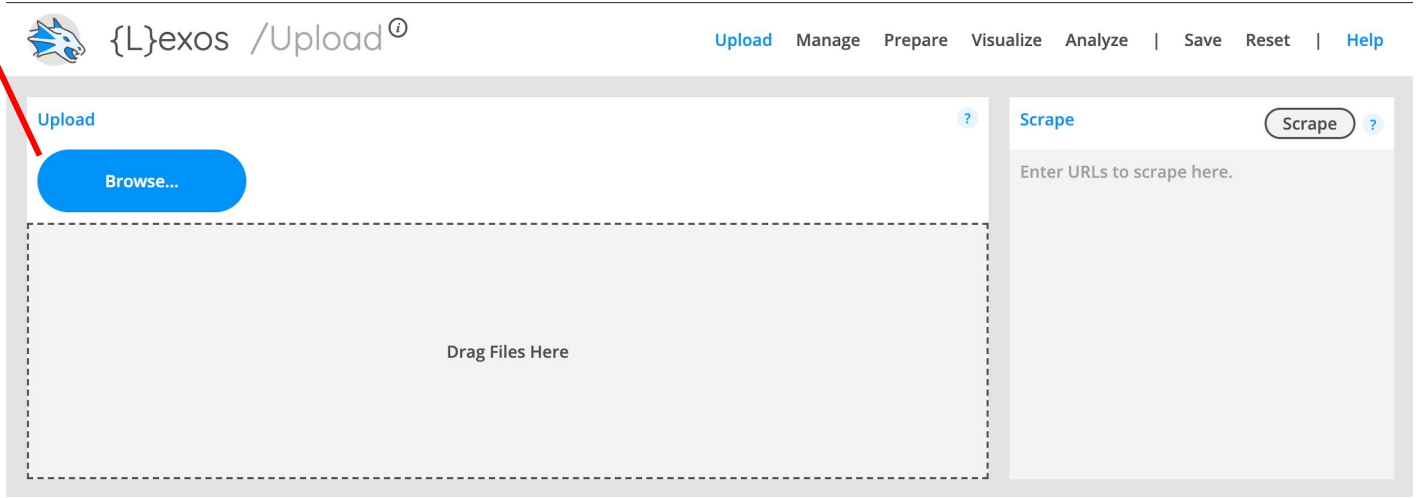
http://lexos.wheatoncollege.edu/upload

Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

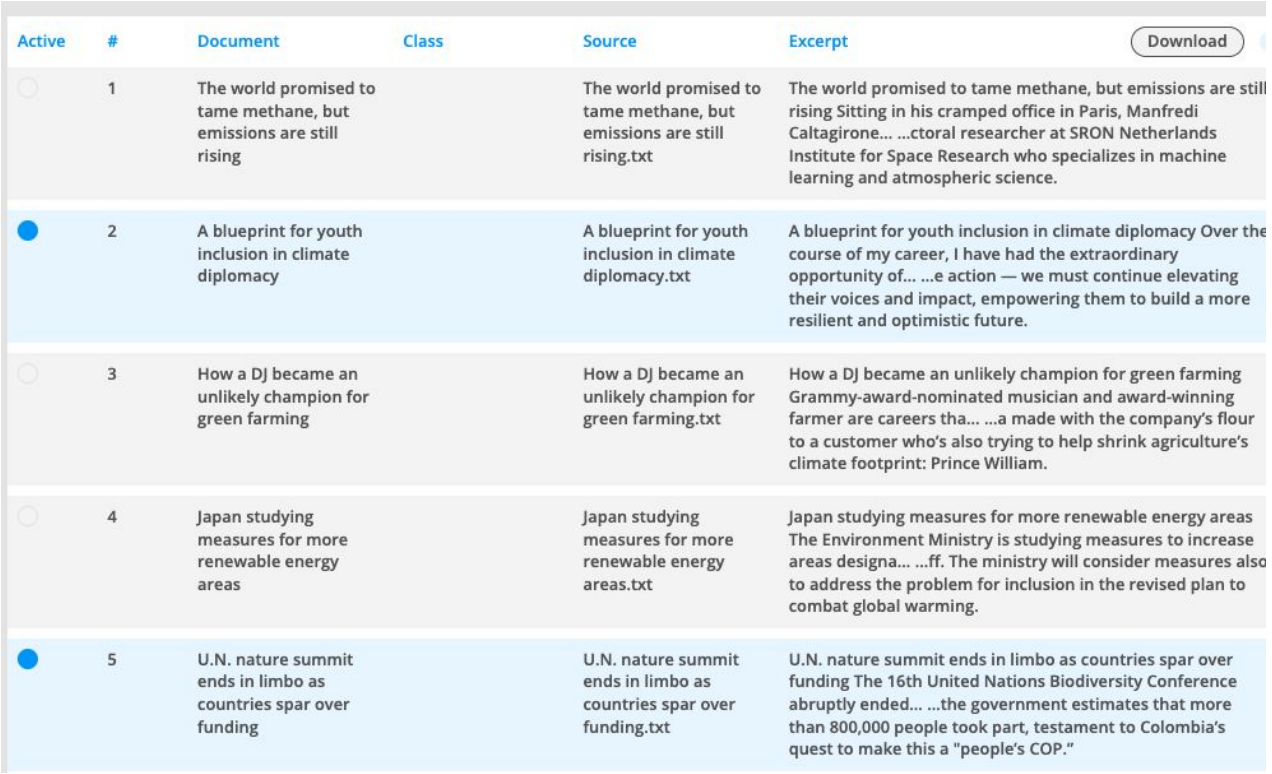


Northeastern University
NULab for Digital Humanities and Computational Social Science

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)



Active	#	Document	Class	Source	Excerpt	Download
	1	The world promised to tame methane, but emissions are still rising		The world promised to tame methane, but emissions are still rising.txt	The world promised to tame methane, but emissions are still rising Sitting in his cramped office in Paris, Manfredi Caltagirone... ...ctoral researcher at SRON Netherlands Institute for Space Research who specializes in machine learning and atmospheric science.	
●	2	A blueprint for youth inclusion in climate diplomacy		A blueprint for youth inclusion in climate diplomacy.txt	A blueprint for youth inclusion in climate diplomacy Over the course of my career, I have had the extraordinary opportunity of... ...e action — we must continue elevating their voices and impact, empowering them to build a more resilient and optimistic future.	
	3	How a DJ became an unlikely champion for green farming		How a DJ became an unlikely champion for green farming.txt	How a DJ became an unlikely champion for green farming Grammy-award-nominated musician and award-winning farmer are careers tha... ...a made with the company's flour to a customer who's also trying to help shrink agriculture's climate footprint: Prince William.	
	4	Japan studying measures for more renewable energy areas		Japan studying measures for more renewable energy areas.txt	Japan studying measures for more renewable energy areas The Environment Ministry is studying measures to increase areas designa... ...ff. The ministry will consider measures also to address the problem for inclusion in the revised plan to combat global warming.	
●	5	U.N. nature summit ends in limbo as countries spar over funding		U.N. nature summit ends in limbo as countries spar over funding.txt	U.N. nature summit ends in limbo as countries spar over funding The 16th United Nations Biodiversity Conference abruptly ended... ...the government estimates that more than 800,000 people took part, testament to Colombia's quest to make this a "people's COP."	

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

Lexos: Prepare (Scrub Case and Punctuation)

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.

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

Lexos: Prepare (Scrub Words)

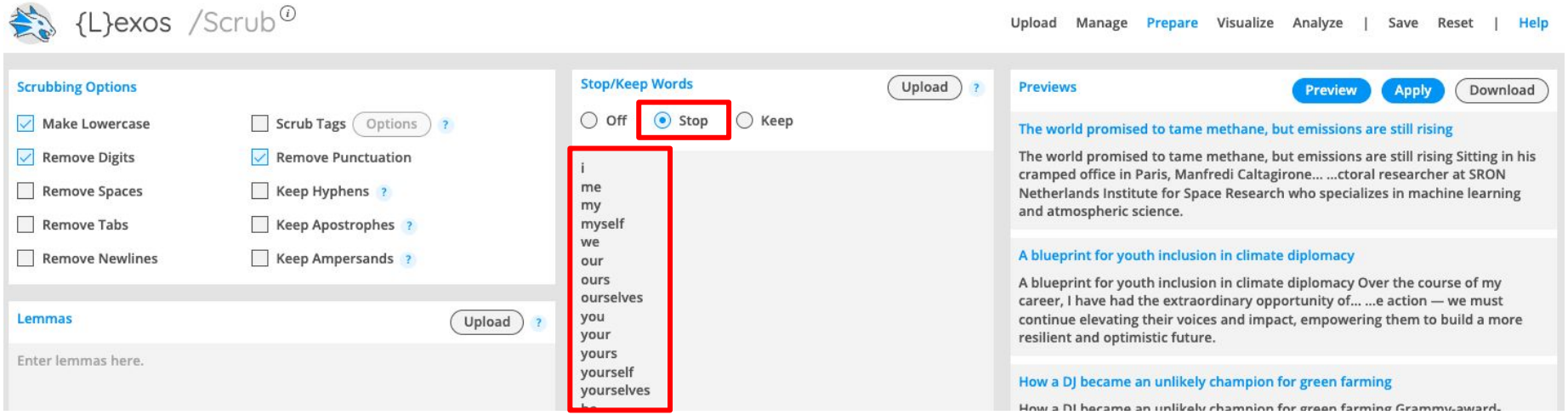
You can also stem words and remove certain words. Here are some possibilities:

- **Stop/Keep Words**: remove a list of words. Usually these would be **stopwords**. With WordCounter, you had to use the stopwords list the tool provided—now, you can choose your own.
- **Lemmas**: standardize to the *stem* of word. For example, you can stem all forms of the verb talk: talking, talked, talks, etc. to "talk"

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

Lexos: Removing Stopwords

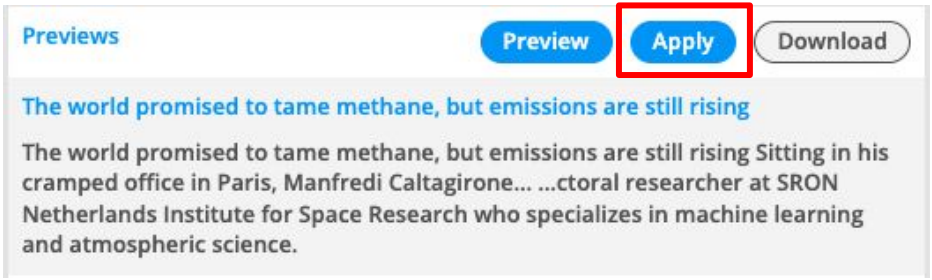
Get a list of English stopwords here: https://gist.github.com/sebleier/554280 (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the "Stop/Keep Words" box then select "Stop"



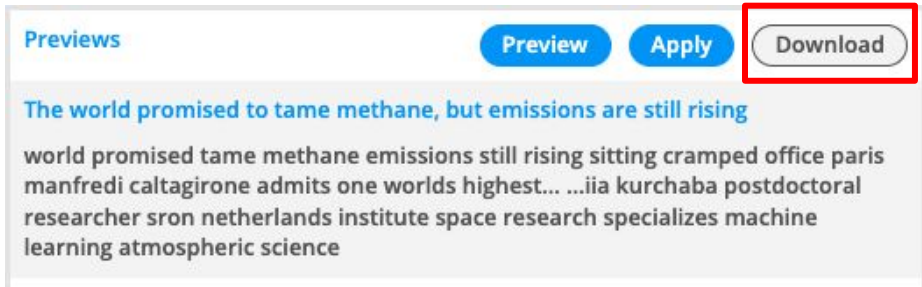
Feel free to ask questions at any point during the presentation!

Lexos: Applying your Preparations

BEFORE PREP



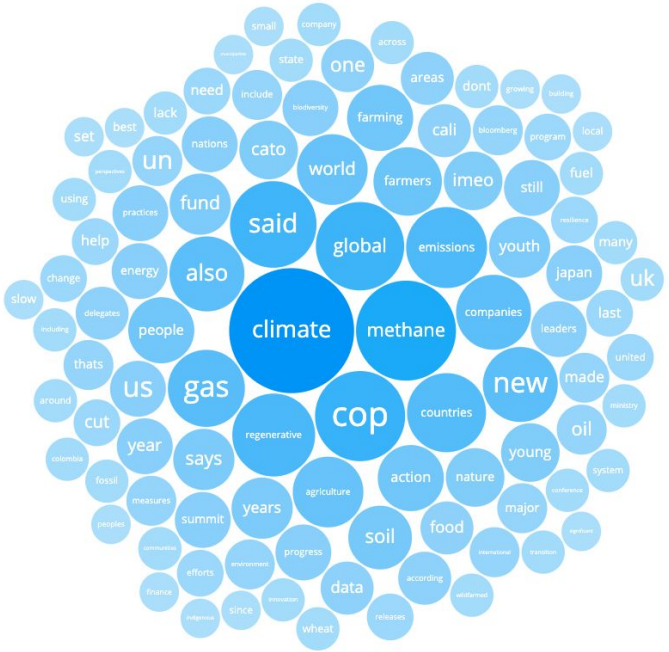
AFTER PREP



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



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

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Lexos: Visualize > Multicloud

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Voyant vs. Lexos: Word Clouds

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

Lexos Word Cloud

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Lexos: Rolling Window

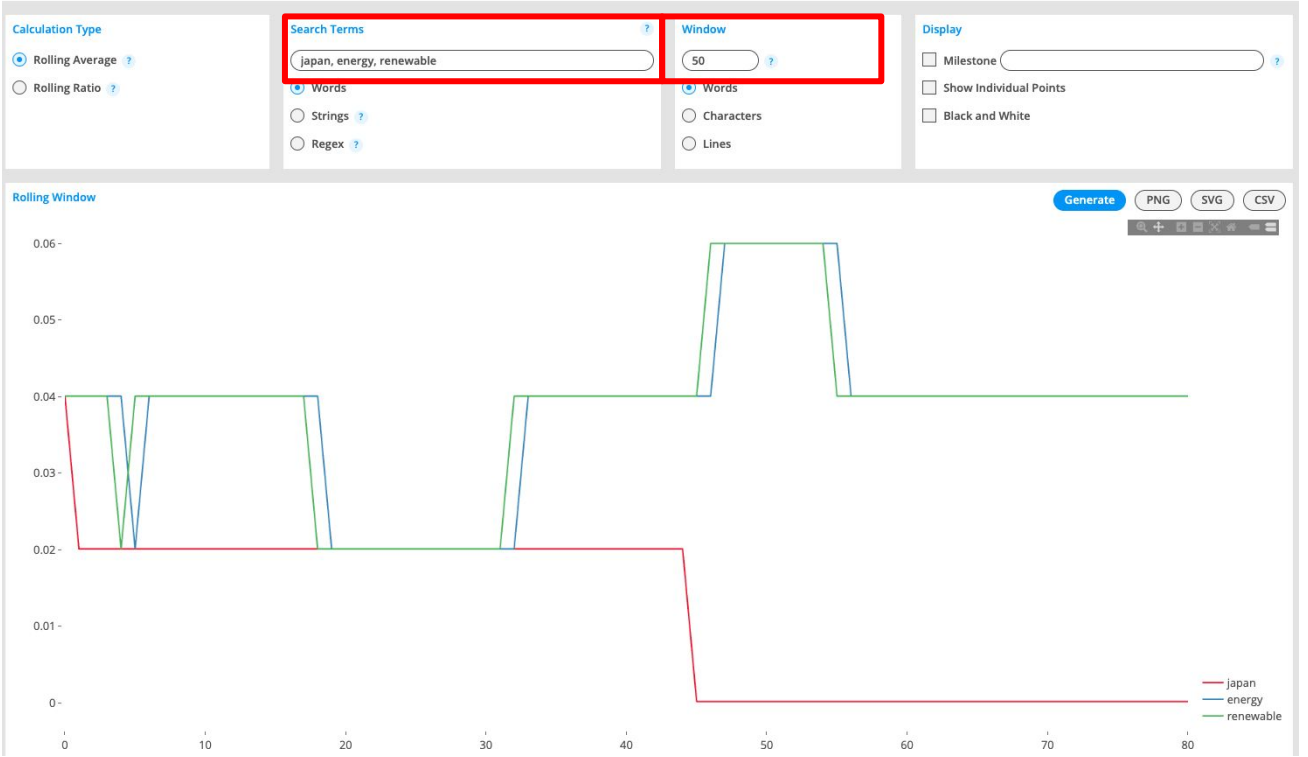
Rolling windows allow you to look at word trends across **one** document. To use a rolling window, first select a single text in the "Manage" screen, then:

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

Searching for the words ‘japan’, ‘energy’ and ‘renewable’ with a window of 50 (since this is a small document), we can get an idea of how these terms work together in the article.



Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

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

Lexos: Analyze > Top words

Top Words

Document "The world promised to tame methane, but emissions are still rising" Compared To The Corpus	
methane	4.9862
gas	3.8308
emissions	3.5576
imeo	2.9095
regenerative	-2.6199
oil	2.4796

Document " A blueprint for youth inclusion in climate diplomacy" Compared To The Corpus	
climate	6.8234
youth	4.8114
young	4.6054
japan	3.7625
leaders	3.5024
perspectives	3.397

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

Lexos: Analyze > Dendrogram

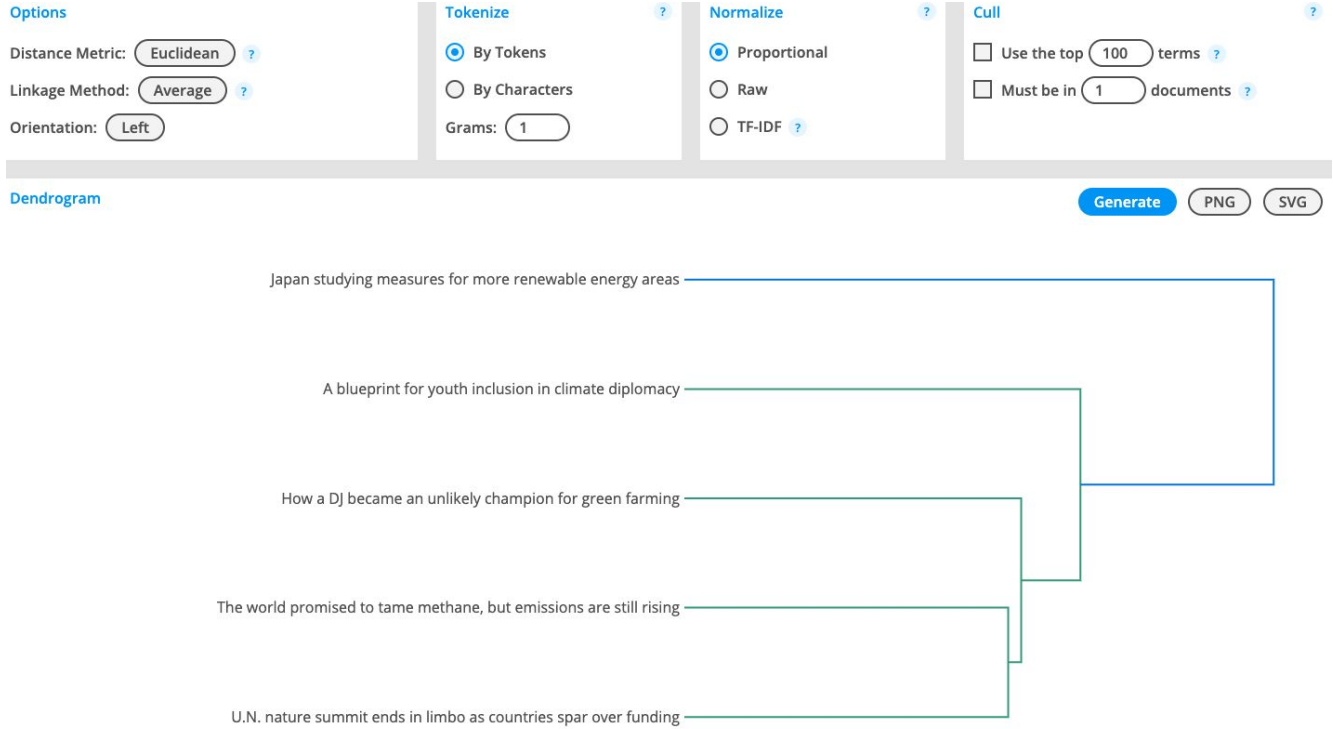
The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms show:

- 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.



Northeastern University
NULab for Digital Humanities and Computational Social Science

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!

Lexos: 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 national energy policy, 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!

—Developed by Dipa Desai, Vaishali Kushwaha, Garrett Morrow, Sara Morrell, and Ayah Aboelela

- For more information on the DITI, please see: https://bit.ly/diti-about
- Schedule an appointment with us! https://bit.ly/diti-meeting
- If you have any questions, contact us at: nulab.info@gmail.com
- We'd love your feedback! Please fill out a short survey here: https://bit.ly/diti-feedback

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