

Computational Text Analysis for Content Analysis

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

POLS 7346 Resilient Cities

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-pols7346>



What is 'Computation 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.



Why Computational Text Analysis?

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

From collections of texts, researchers can discover keywords that serve as a proxy for major trends in societies, cultures, and policies. For example, computational tools can reveal patterns on how public officials communicate policies, which issues are of concern, which phrases leaders regularly employ, and much more.



[illegible]

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.



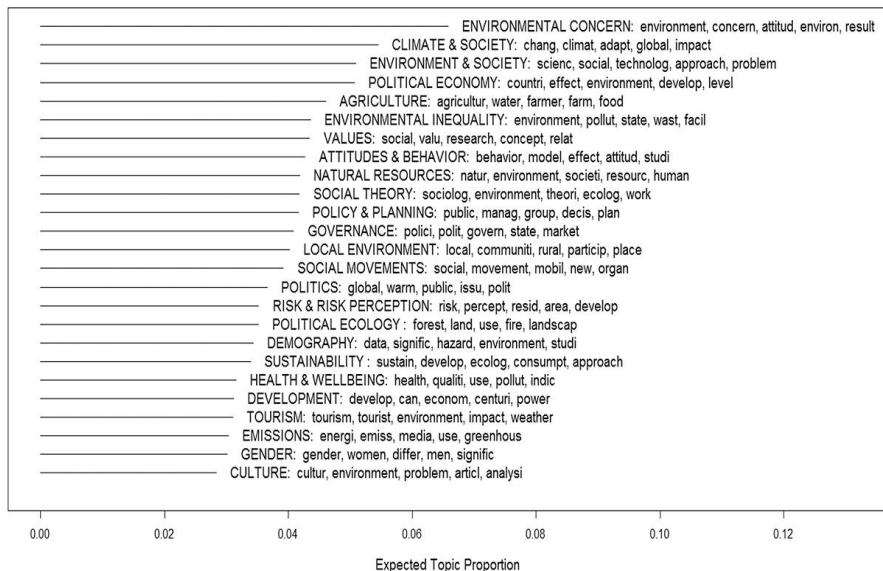
Examples from Practice



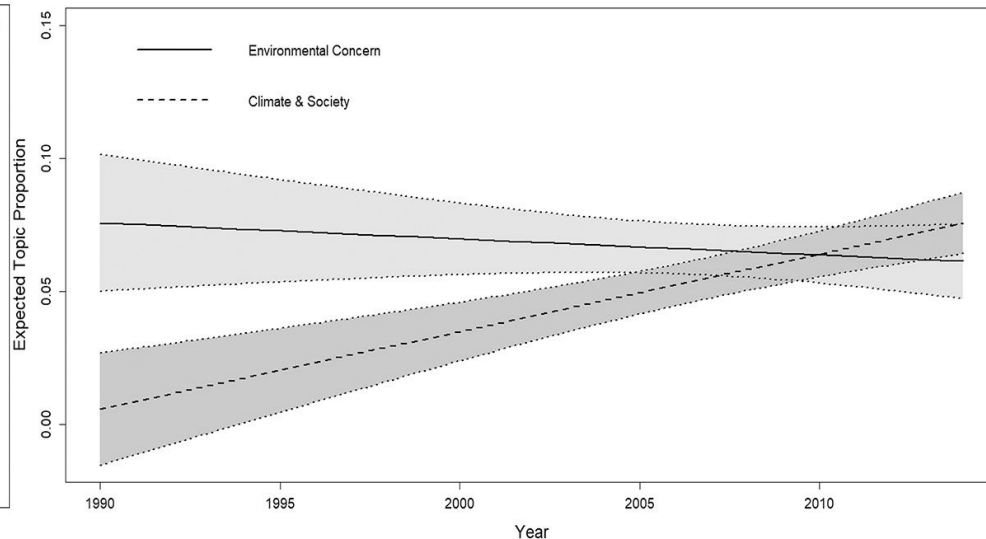
Northeastern University
NULab for Texts, Maps, and Networks

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

Key Topics in environmental sociology, 1990–2014: results from a computational text analysis



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](https://doi.org/10.1080/23251042.2017.1393863)



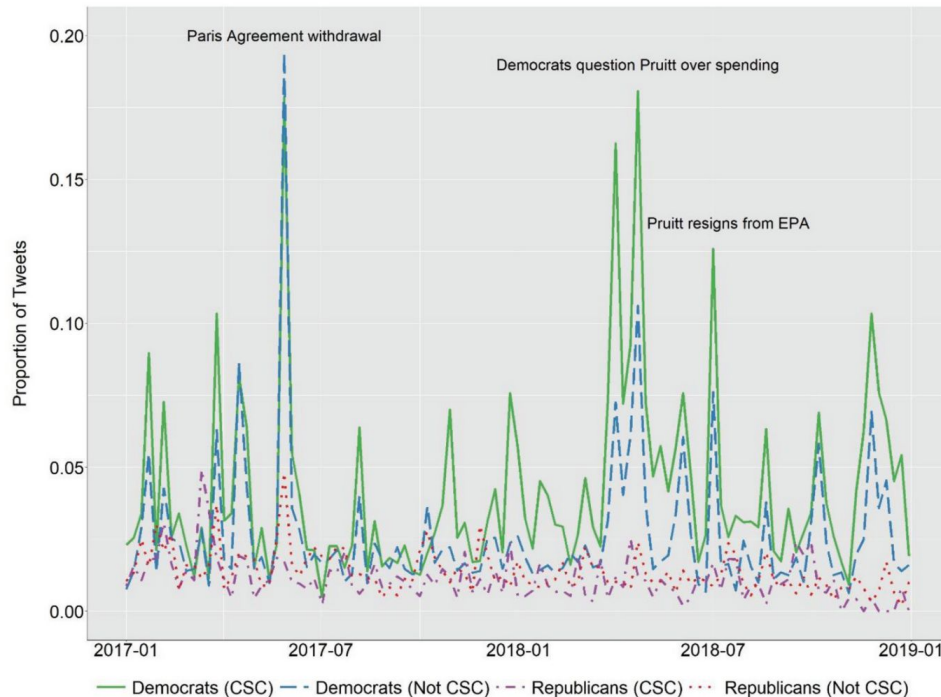
Northeastern University
NULab for Texts, Maps, and Networks

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

U.S. Environmental Politics

To what extent politicians publicly discuss environmental issues in line with public opinion and 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.

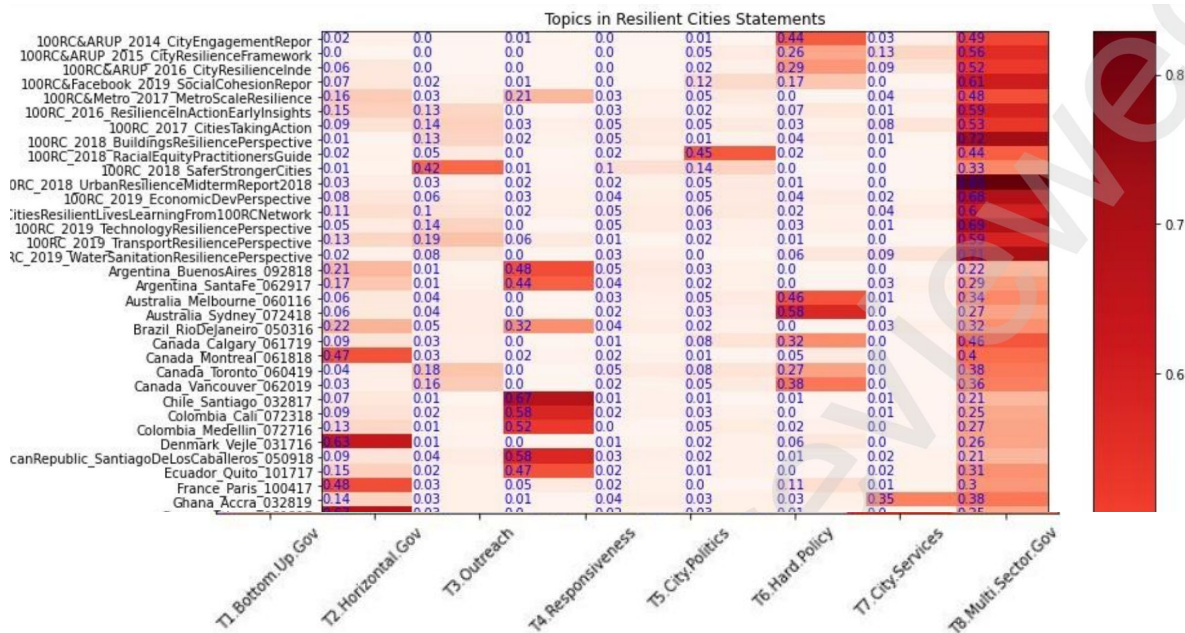


Weekly proportions of tweets discussing environmental issues sent by the 115th House of Representatives.

[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



Resilient Cities Statements Computational Text Analysis



Computational text analysis of resilience strategy language shows different place-based priorities, as well as gaps and overlaps in 100 cities' resilience strategies.

DITI and POLS 7346 class alum Garrett Morrow applied computational text analysis to model and identify [topics in resilient cities' strategy documents](#).



Northeastern University
NULab for Texts, Maps, and Networks

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

Boston Area Research Initiative (BARI) Data

You can use the [Boston Data Portal](#) and [BARI GIS map](#) to access Boston-specific information. This dataset includes 311 data, 911 calls, Airbnb listings, Craigslist housing ads, and more.

For example, BARI researchers looked at 911 calls and weather data [to identify heat islands and related illnesses around Boston](#), and consider how policies may be improved to reduce heat islands in Boston.



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.



Text Preparation



Northeastern University
NULab for Texts, Maps, and Networks

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

Corpus Building

Questions to consider before you begin:

- 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 information, see our [Corpus Building Handout](#).



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. Copy and paste the 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.



Our Text

Our text is plain text (.txt file) of the 2019 [Boston Climate Action Plan](#). The primary objective is to explore this text using web-based computational text analysis tools.

We will also use climate change plans of [New York City](#), [Chicago](#), [New Orleans](#), and [Phoenix](#) to see how a corpus can be analyzed. The primary objective is to compare the climate change plans of these cities with Boston.

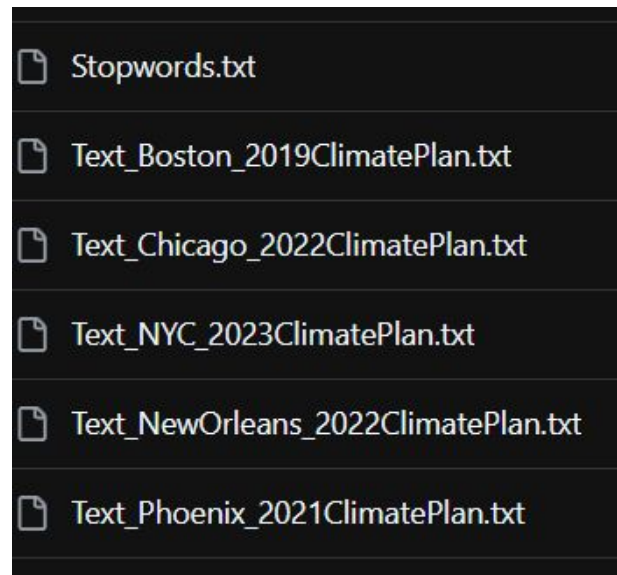


Sample Corpus

The following .txt files are available on:

<http://bit.ly/sp24-aldrich-pols7346>

- 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.



Exploratory Tool: Word Counter



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



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.



Word Counter Examples

TOP WORDS ⬇

Word	Frequency
boston	551
city	299
energy	279
carbon	196
2019	153
climate	145
buildings	143
emissions	141
building	135
update	119
transportation	109

Shows the top words in the text.

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

BIGRAMS ⬇

bigram	Frequency
the city	182
of boston	97
2019 boston	87
cap update	87
boston cap	86
boston s	84
city of	76
in the	75
of the	64
carbon emissions	54

TRIGRAMS ⬇

trigram	Frequency
2019 boston cap	86
boston cap update	86
city of boston	70
the city of	61
the city s	28
the city will	27
of boston s	23
of boston will	22
zero net carbon	21
cap update buildings	21
go boston 2030	20
climate action plan	19

It is interesting that though 'energy' is third most used term, it does not appear in top few trigrams.

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



Exploratory Tool: Word Tree



Northeastern University
NULab for Texts, Maps, and Networks

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

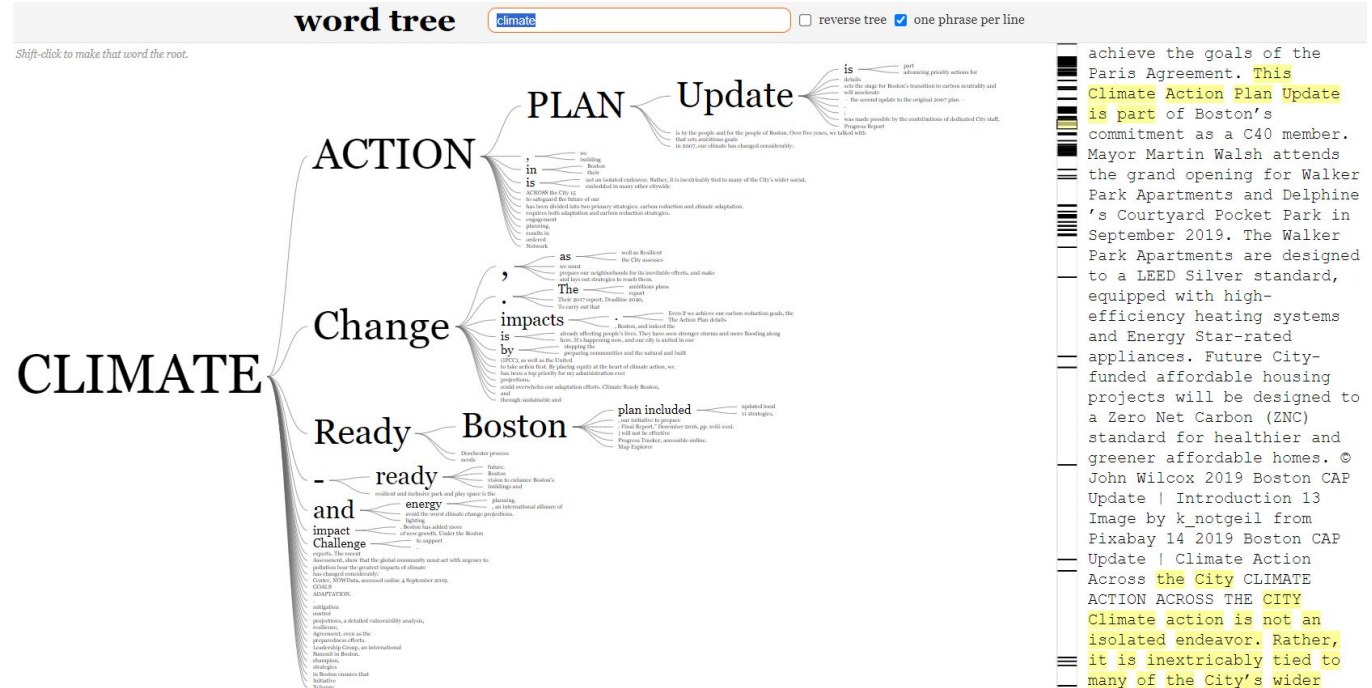
Word Tree

- <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: **fewer than 1 million words** should work



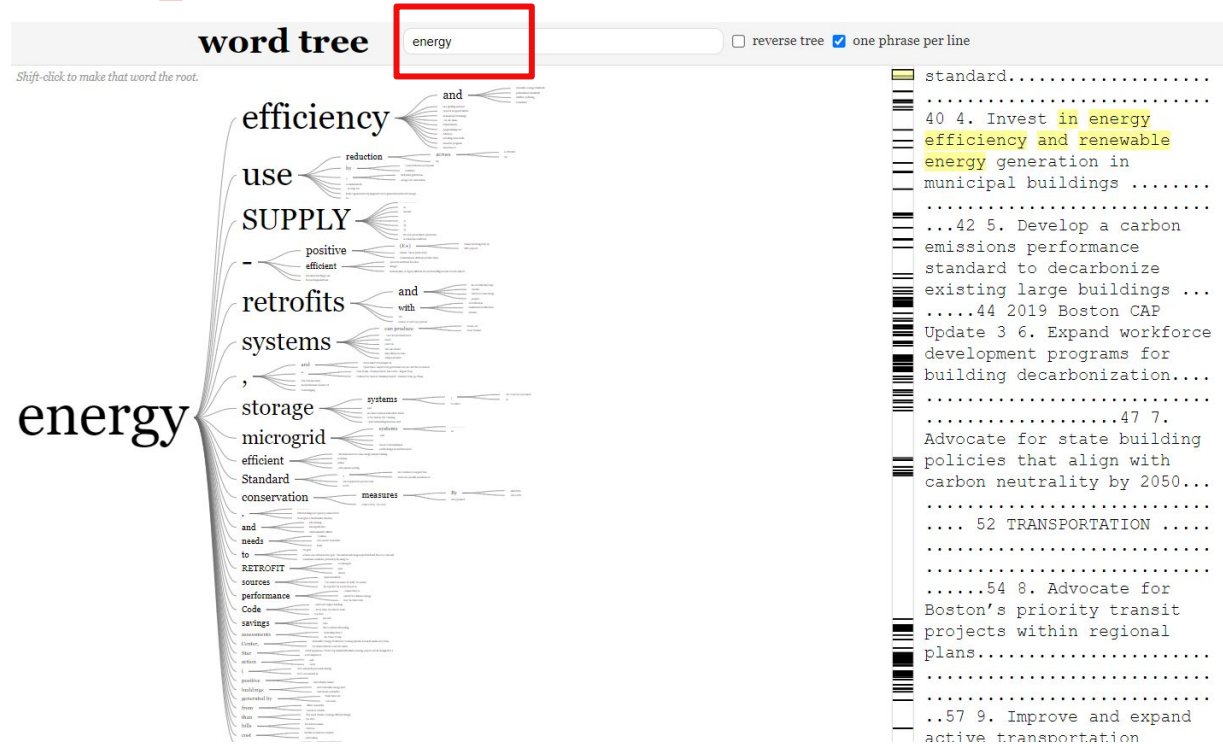
Word Tree Example

Reflects the focus of the report on climate ready boston, climate change, resilience initiatives, net carbon strategies, etc.



Word Tree Examples

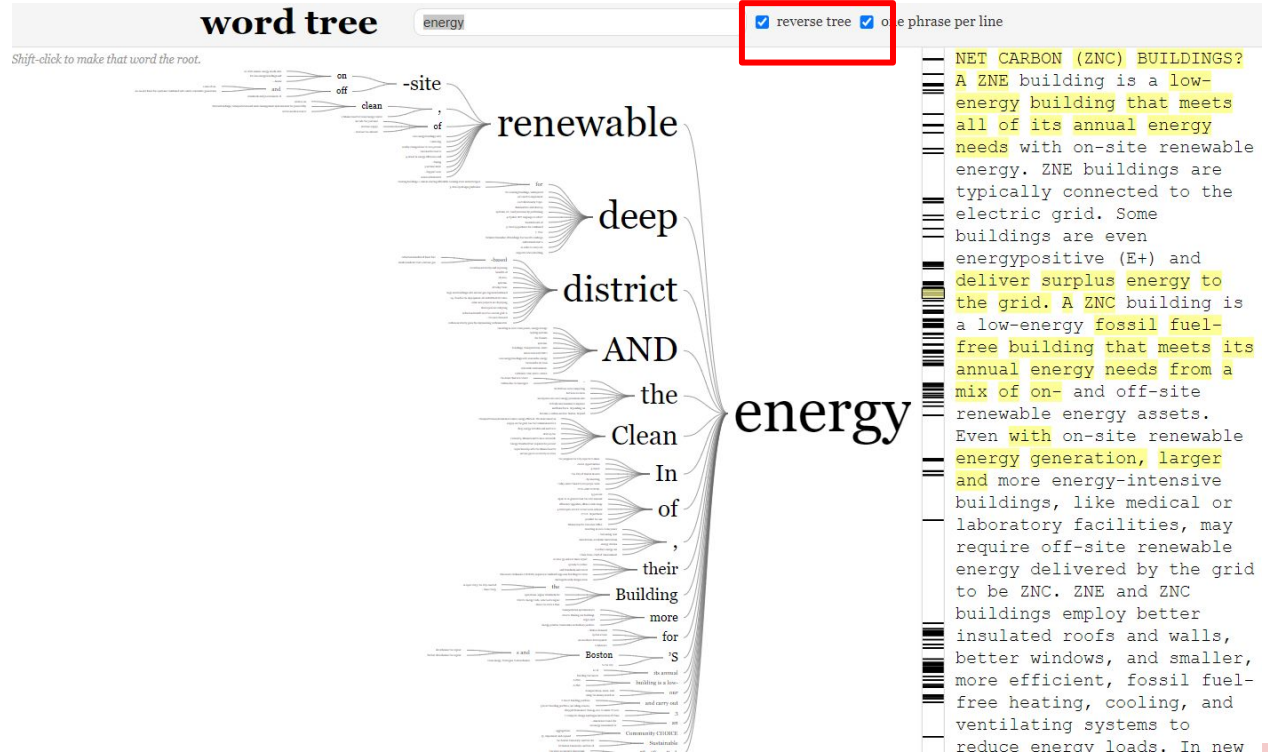
‘Energy’ is the third most used word in the text. It is followed by various patterns: efficiency, use, supply, systems, retrofits, storage, etc.



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.

Here renewable, deep, district, and clean are common words preceding the word “energy.”



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 the limitations, how do you think you can use these tools in your research?
- What types of text would be interesting to explore with these tools?



Powerful Platform: Voyant



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. It also makes nice visualizations!

<https://voyant-tools.org/>



VOYANT

see through your text

Add Texts

Type in one or more URLs on separate lines or paste in a full text.

Open Upload Reveal

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

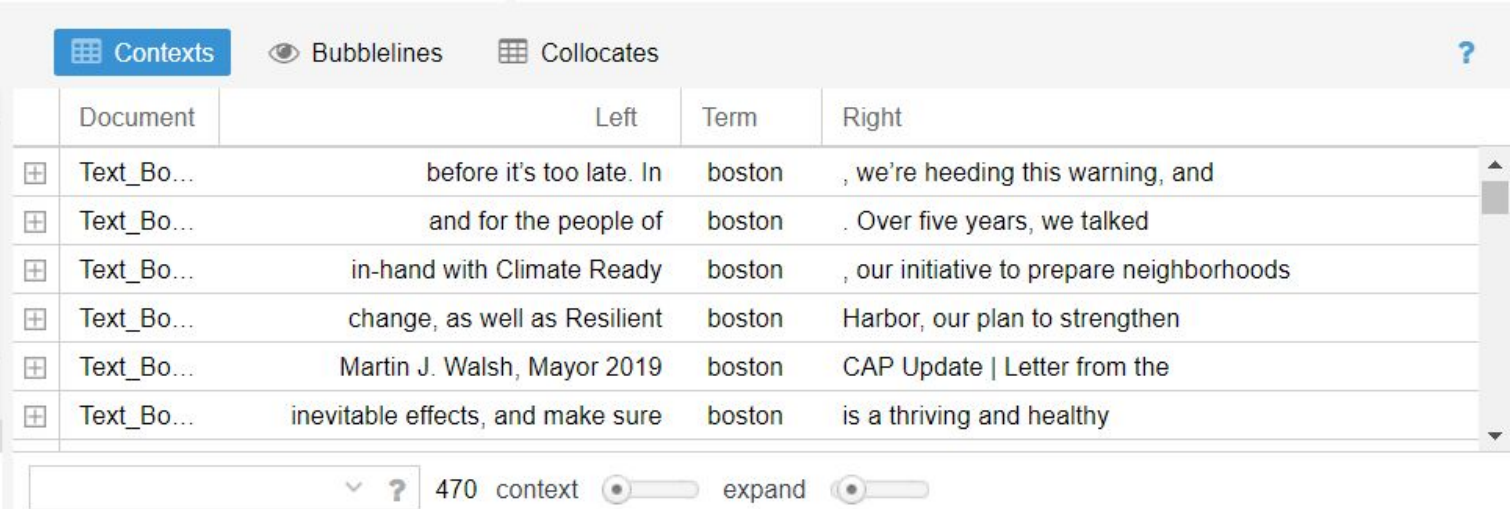
Alternatively, insert URLs or full text into the textbox.

Click here for help and advanced options



Voyant: Contexts (concordances)

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



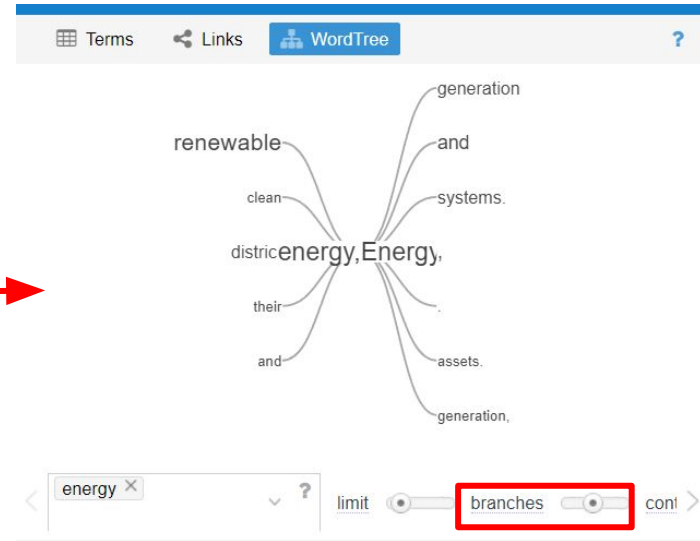
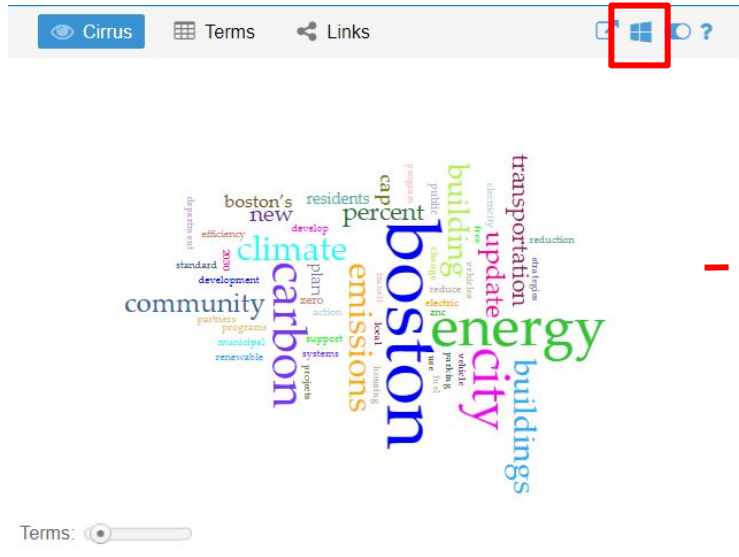
	Document	Left	Term	Right
+	Text_Bo...	before it's too late. In	boston	, we're heeding this warning, and
+	Text_Bo...	and for the people of	boston	. Over five years, we talked
+	Text_Bo...	in-hand with Climate Ready	boston	, our initiative to prepare neighborhoods
+	Text_Bo...	change, as well as Resilient	boston	Harbor, our plan to strengthen
+	Text_Bo...	Martin J. Walsh, Mayor 2019	boston	CAP Update Letter from the
+	Text_Bo...	inevitable effects, and make sure	boston	is a thriving and healthy

470 context expand



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.



Results page of the corpus containing climate reports of 5 cities.

- [illegible]

Your Turn!

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

Discussion Prompts

- What interesting or surprising results came up?
- How might you interpret those results based on what you know about current climate plans?
- What other kinds of documents would be useful to compare in Voyant to research trends in climate and disaster planning?



Powerful Platform: Lexos



Northeastern University
NULab for Texts, Maps, and Networks

*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

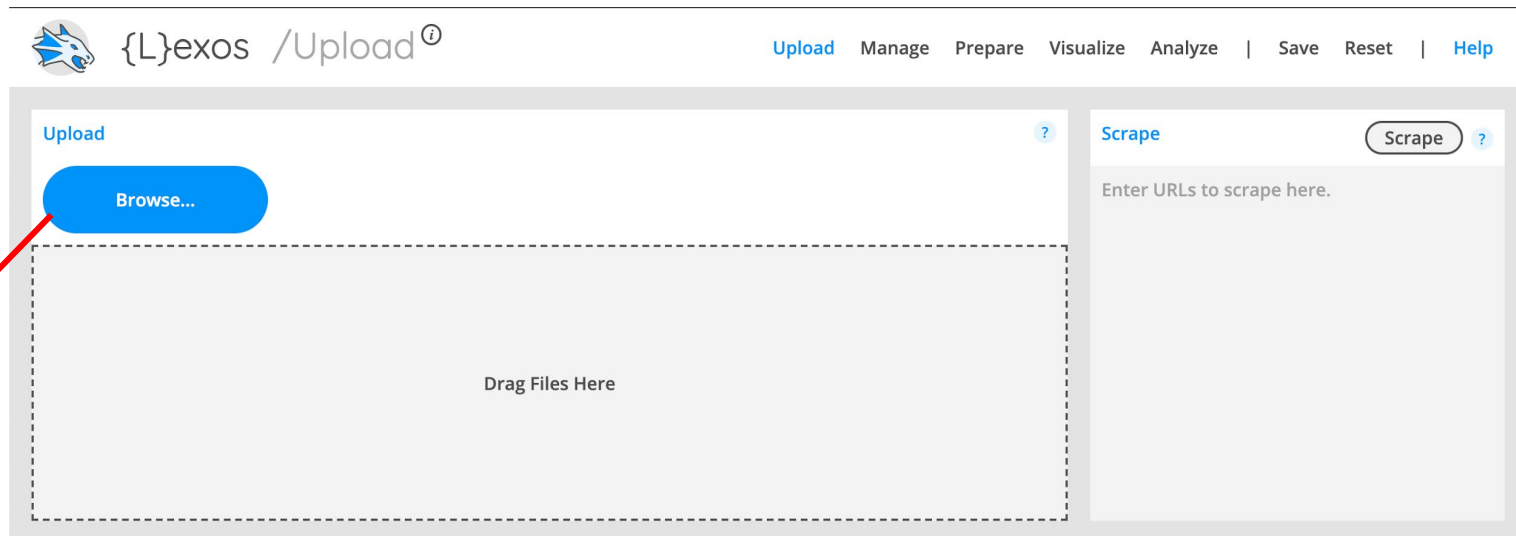
<http://lexos.wheatoncollege.edu/upload>



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.





Lexos: Manage

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

 {L}exos /Manageⁱ

Upload [Manage](#) Prepare Visualize Analyze | Save Reset | [Help](#)

Active	#	Document	Class	Source	Excerpt	Download
<input checked="" type="radio"/>	1	Text_Phoenix_2021ClimatePlan		Text_Phoenix_2021ClimatePlan.txt	CLIMATE ACTION PLAN 2021 EDITION SEPTEMBER 27, 2021 2 LETTER FROM THE MAYOR..... ..will be expanded from pilot to ongoing program by increasing the number of inspections from nine to 40. ACTIONS MATRIX - WATER	
<input type="radio"/>	2	Text_NYC_2023ClimatePlan		Text_NYC_2023ClimatePlan.txt	The City of New York Mayor Eric Adams April 2023 Getting Sustainability Done PlaNYC Letter from the Mayor Introduction Our Vision... ..nyc analysis by WSP USA Graphical analysis by WXY Design by Nowhere Office Policy support by Thornton Tomasetti nyc.gov/climate	
<input checked="" type="radio"/>	3	Text_Boston_2019ClimatePlan		Text_Boston_2019ClimatePlan.txt	1 October 2019 MAYOR MARTIN J. WALSH CITY OF BOSTON CLIMATE ACTION PLAN 2019 UPDATE 2 2019 Boston CAP Update CONTENTS INTRODUCT... ..e. 86 2019 Boston CAP Update 2019 Boston CAP Update 87 2019 CLIMATE ACTION PLAN UPDATE City of Boston MAYOR MARTIN J. WALSH	
<input checked="" type="radio"/>	4	Text_NewOrleans_2022ClimatePlan		Text_NewOrleans_2022ClimatePlan.txt	A Priority List for Climate Action in New Orleans Net Zero by 2050: December 2022 Introduction The Intergovernmental Panel on Climate Change and editing this plan. 30 City of New Orleans Office of Resilience & Sustainability 1300 Perdido St. 8E08 New Orleans, LA 70112	
<input checked="" type="radio"/>	5	Text_Chicago_2022ClimatePlan		Text_Chicago_2022ClimatePlan.txt	CHICAGO MAYOR LORI E. LIGHTFOOT 2022 CAP CHICAGO CLIMATE	

 Lexos v4.0 © 2019 Wheaton Lexomics

Active Documents : 4



Lexos: Prepare (scrub)

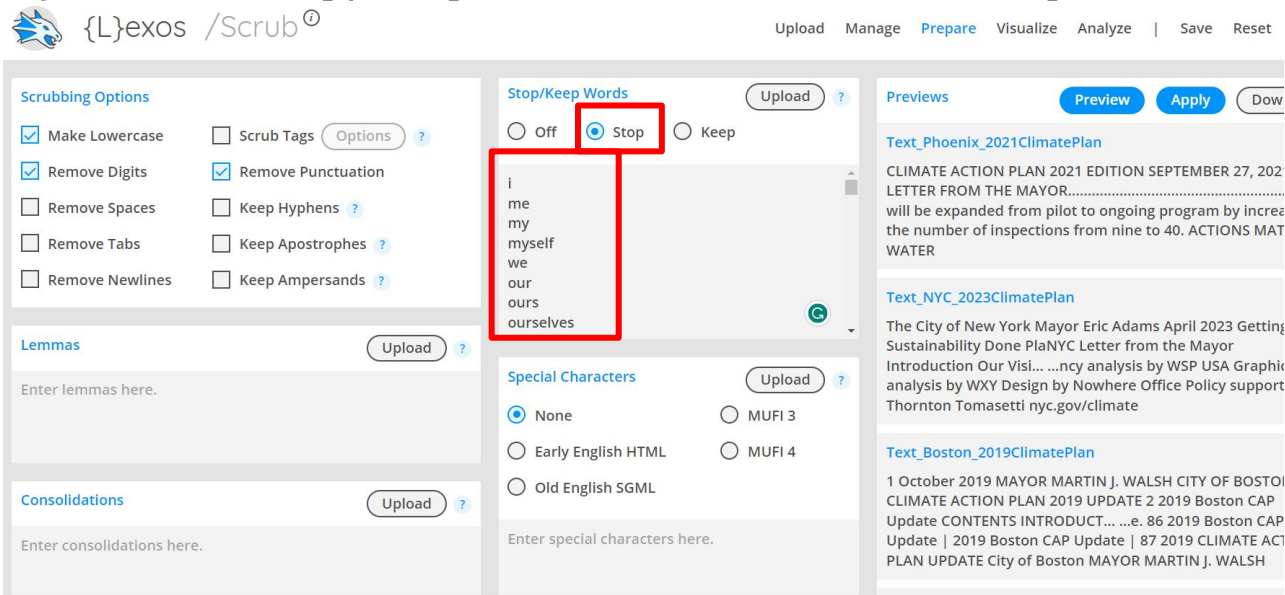
Lexos demonstrates some more advanced options you have for preparing your corpus. By “scrubbing,” you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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”



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”



The screenshot shows the Lexos v4.0 interface. The 'Stop/Keep Words' section is highlighted with a red box around the 'Stop' radio button. Below it, a list of stopwords is visible: 'i', 'me', 'my', 'myself', 'we', 'our', 'ours', and 'ourselves'. The 'Scrubbing Options' section on the left includes checkboxes for 'Make Lowercase', 'Remove Digits', 'Remove Spaces', 'Remove Tabs', 'Remove Newlines', 'Scrub Tags', 'Remove Punctuation', 'Keep Hyphens', 'Keep Apostrophes', and 'Keep Ampersands'. The 'Special Characters' section on the right has radio buttons for 'None', 'Early English HTML', and 'Old English SGML'. The 'Previews' section on the right shows three document previews: 'Text_Phoenix_2021ClimatePlan', 'Text_NYC_2023ClimatePlan', and 'Text_Boston_2019ClimatePlan'.



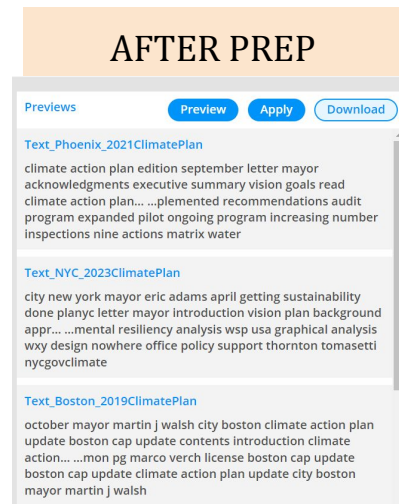
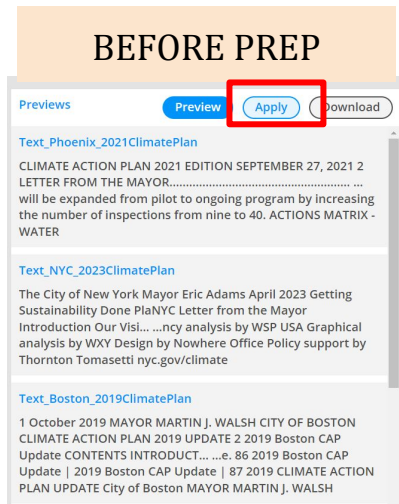
Lexos v4.0 © 2019 Wheaton Lexomics

Northeastern University
NULab for Texts, Maps, and Networks

Active Do

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

Lexos: Applying your Preparations



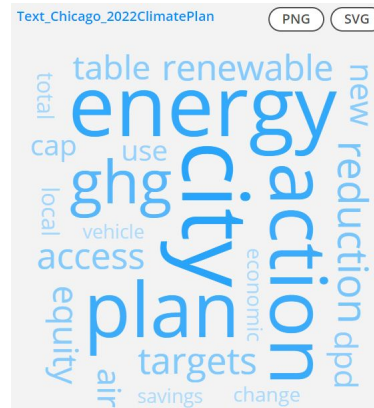
Once you have made decisions about your preparations, click “**Apply**” and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your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



The screenshot shows the Word Cloud Generator interface. At the top, there are navigation links: Upload, Manage, Prepare, Visualize, Analyze, Save, Reset, and Help. Below these, the text 'Word Cloud' is displayed. The main area shows a word cloud generated from the input text. The words are arranged in a circular pattern, with 'city' and 'climate' being the largest. Other prominent words include 'new', 'energy', 'environmental', 'support', 'work', 'water', 'office', 'actions', 'projects', 'also', 'reduce', 'lead', 'department', 'local', 'food', 'plan', 'york', 'state', 'air', 'building', 'emissions', 'waste', 'ghg', 'heat', 'carbon', 'development', 'transportation', 'buildings', 'public', 'program', 'nyc', and 'development'. The interface includes controls for Font (Open Sans), Term Count (80), Color (Default), and buttons for Generate, PNG, and SVG.

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

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



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

Voyant vs. Lexos: Wordclouds

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



Lexos Wordcloud



What could be causing this distinction?



Lexos: Rolling Window

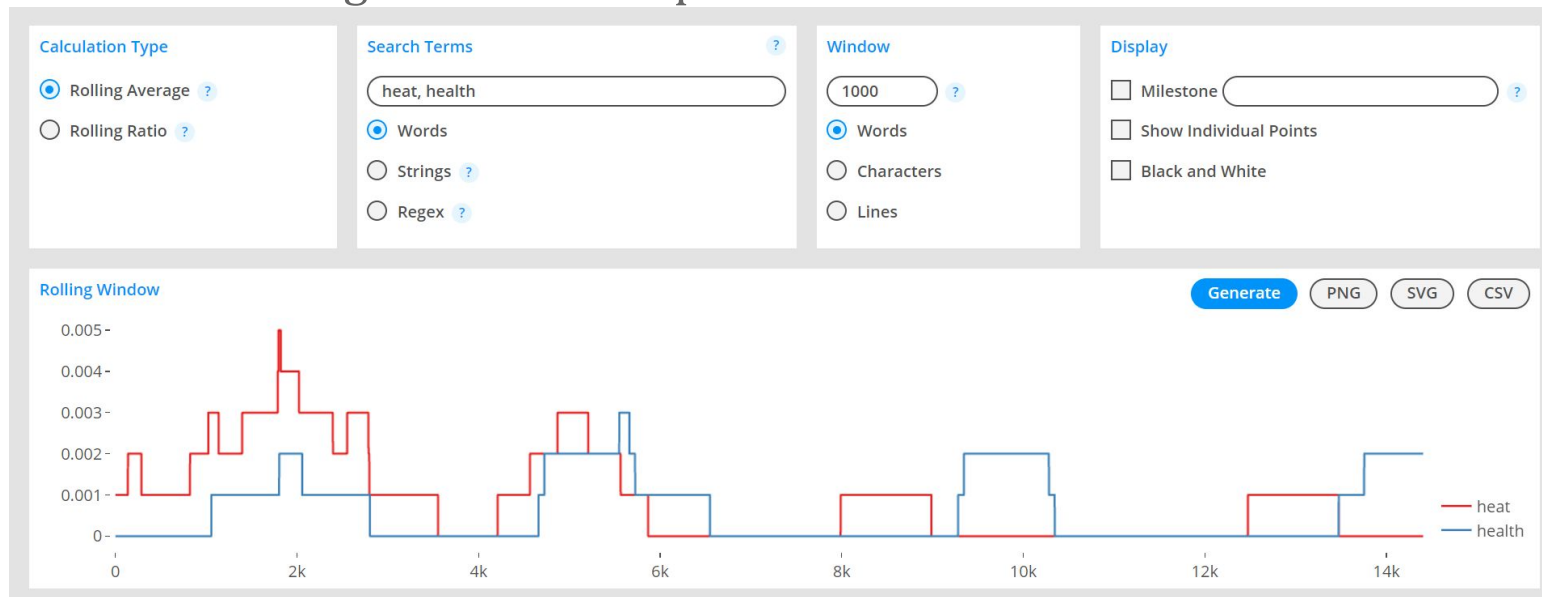
Rolling windows allow you to look at word trends across **one** document. To use a rolling window, first select a single text in the "Manage" screen, then:

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 (heat, health, flood, storm)
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"



Lexos: Rolling Window Results

Using the 2019 Boston Climate Plan, and searching for the words 'heat' and 'health' with a window of 1000 (since this is a large document), we can get an idea of how these terms work together in the report.



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



Lexos: Dendrogram

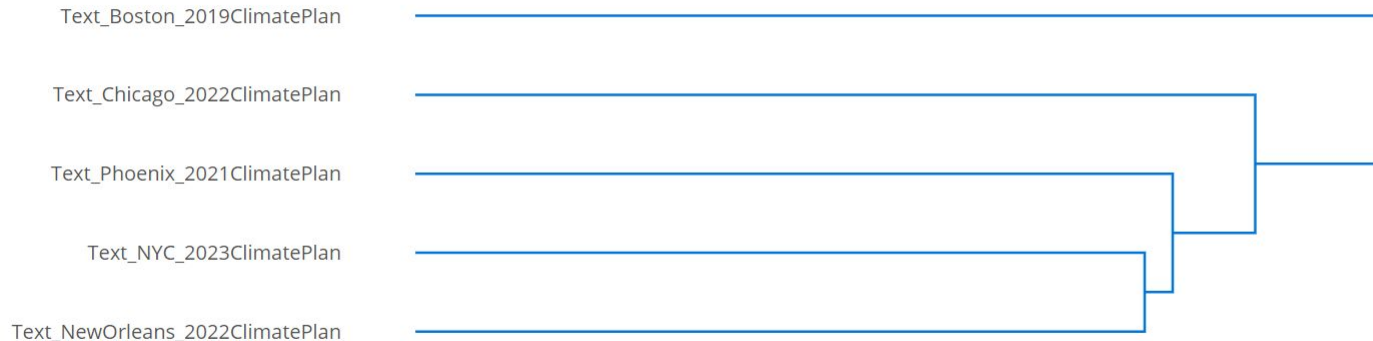
The dendrogram demonstrates similarity between the different documents.

Dendrogram

Generate

PNG

SVG



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 Lexos's features!**

Discussion Prompts

- Between Voyant and Lexos, which tool do you prefer and why?
- How would you want to use these tools for your research? Which features do you think will be useful in your analysis?
- How do you think this computational text analysis can complement your other research methods?



A Brief Introduction to Web Scraping

Slide content courtesy of Alyssa Smith



Northeastern University
NULab for Texts, Maps, and Networks

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

Why Access Internet Data?

- Internet data can give us a way to (very imperfectly) quantify people's social lives online.
 - What are people talking about?
 - Who do people interact with?
 - How do communities form?
- It is especially useful at large scales.
 - Getting this kind of information on how people associate without social media data would be very difficult, if not impossible!
- Internet data is very rich in terms of context, content, and usability.
- Internet data captures certain times, cultures, and social contexts.

This is useful when researching recent and current issues.



How can you access internet data?

Unless you want to hand copy the contents of each web page, one at a time, you will need to use a program for automatically extracting data from the web. In some cases, websites provide their own tools, called **APIs**, that are designed to let you retrieve data that you specify. In other cases, you might use general software for **scraping** the contents of websites.

It helps to understand the general principles of how APIs and web scraping work, but typically each site will have its own specifications that you will need to learn to access their data.






Access Web Data Through APIs

- An API is a way for computer programs to talk to each other.
- APIs are code wrappers, a clean way to code communication with websites that eliminates the need for more complicated scraping.
- If you are trying to get a lot of information repeatedly from somebody else's computer program, an API is the way to do it!
- This might look like:
 - An analysis of all reddit posts mentioning “electric vehicles”.
 - A program that emails you every time your elected officials in Congress post something with a negative sentiment.



API Example - NY Times

- The New York Times features a Developer tool, available here: <https://developer.nytimes.com/>
- From here, users can sign up and access a variety of NYT content through their APIs.

		
Archive API Get all NYT article metadata for a given month.	Article Search API Search for New York Times articles.	Books API Get NYT Best Seller lookup book reviews

Article Search

Use the Article Search API to look up articles by keyword. You can refine your search using filters and facets.

```
/articlesearch.json?q={query}&fq={filter}
```

[NYT Article Search API](#)



Web Scraping

- Sometimes websites don't have an API; you'll have to scrape the website.
- Scraping pulls the whole webpage—you then parse it and extract the data you want.
- This works better on structured websites that don't block bots (if you are scraping a website, you are a bot).
- Please obtain consent before scraping content from a site (or, at least, try to!)



Ethical Considerations of Scraping

- **Contextual Privacy**

- When we think about privacy online we want to think of it as contextual. What someone might be comfortable saying in one context might not be something they would say to a researcher or want to be quoted in a publication.

- **Keeping People Safe**

- It is risky to publicize the username, profile picture, or exact text of a social media post or profile.
- To show example posts etc, you can make up your own or heavily redact them.



Learn More About Web Scraping

- <https://bit.ly/ScrapingSlides>
- [Data Ethics Handout](#)
- Northeastern Library [Guide on Text and Data Mining Library Databases](#)
- Databases for learning and applying text analysis:
 - [Constellate](#)
 - [ProQuest TDM](#)



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-pols7346>

Schedule an appointment with us! <https://bit.ly/diti-meeting>

We'd love your feedback! Please fill out a short survey here:

<https://bit.ly/diti-feedback>

