Computational Text Analysis for Content Analysis

Taught By Sara Morrell and Johan Arango-Quiroga
Digital Integration Teaching Initiative (DITI)

POLS 7346 Resilient Cities
Daniel Aldrich
Spring 2025

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-pols7346

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?

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.

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.

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

Language Used in Climate News

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?

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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)

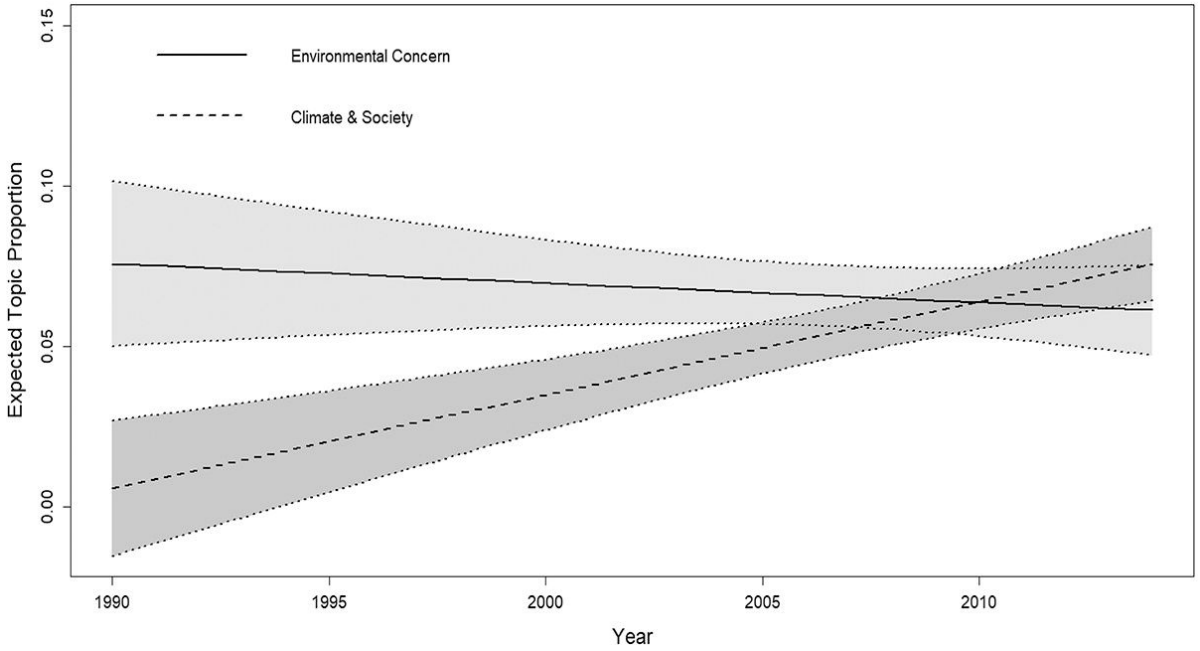
- **nGram:** A continuous sequence of *n* items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular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

Key Topics in environmental sociology



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

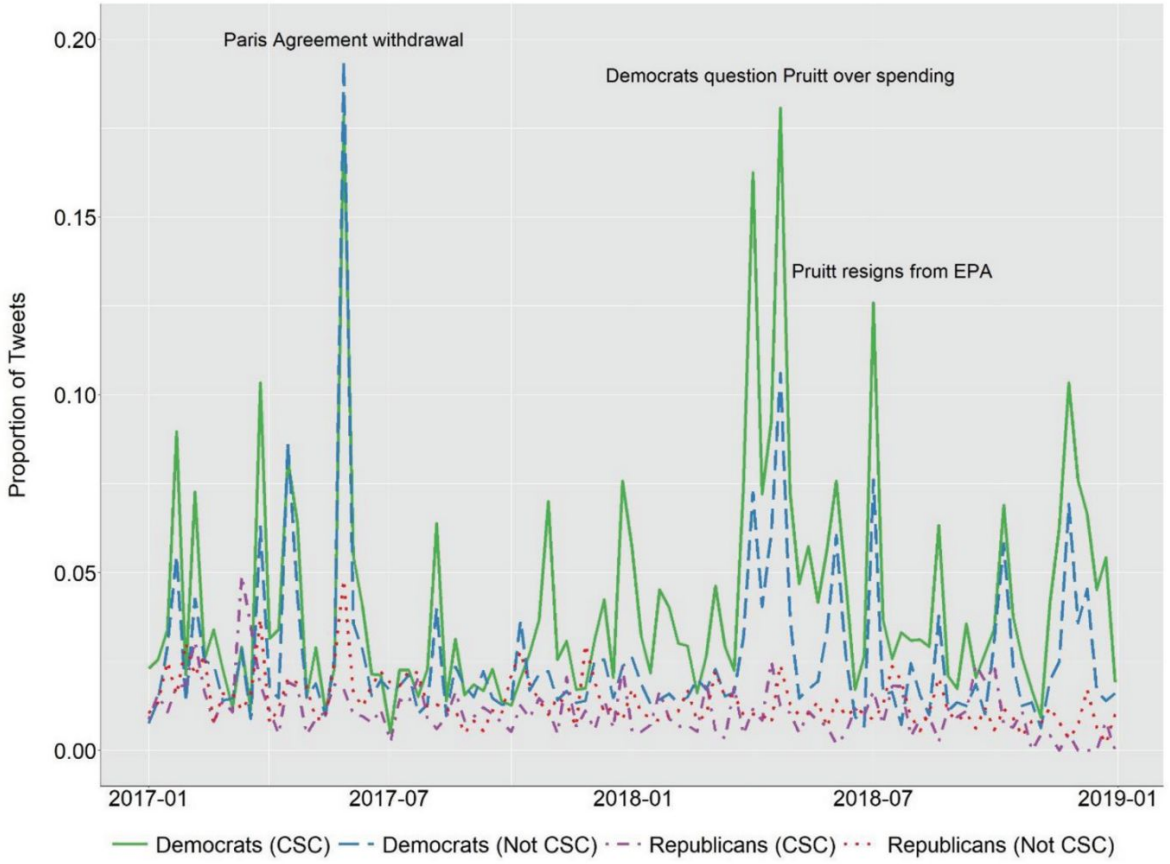
Northeastern University
NULab for Digital Humanities and Computational Social Science

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

U.S. Environmental Politics (1/2)

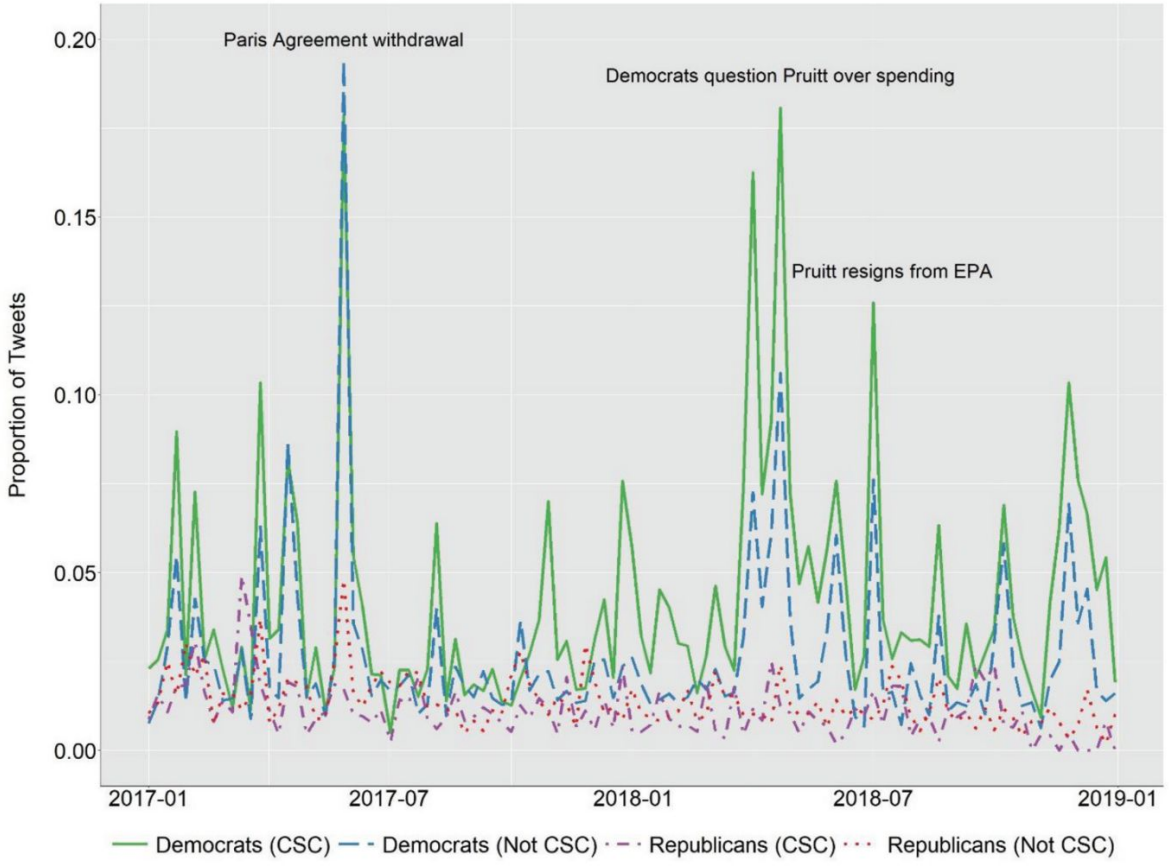
Weekly proportions of tweets discussing environmental issues sent by the 115th House of Representatives.

[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

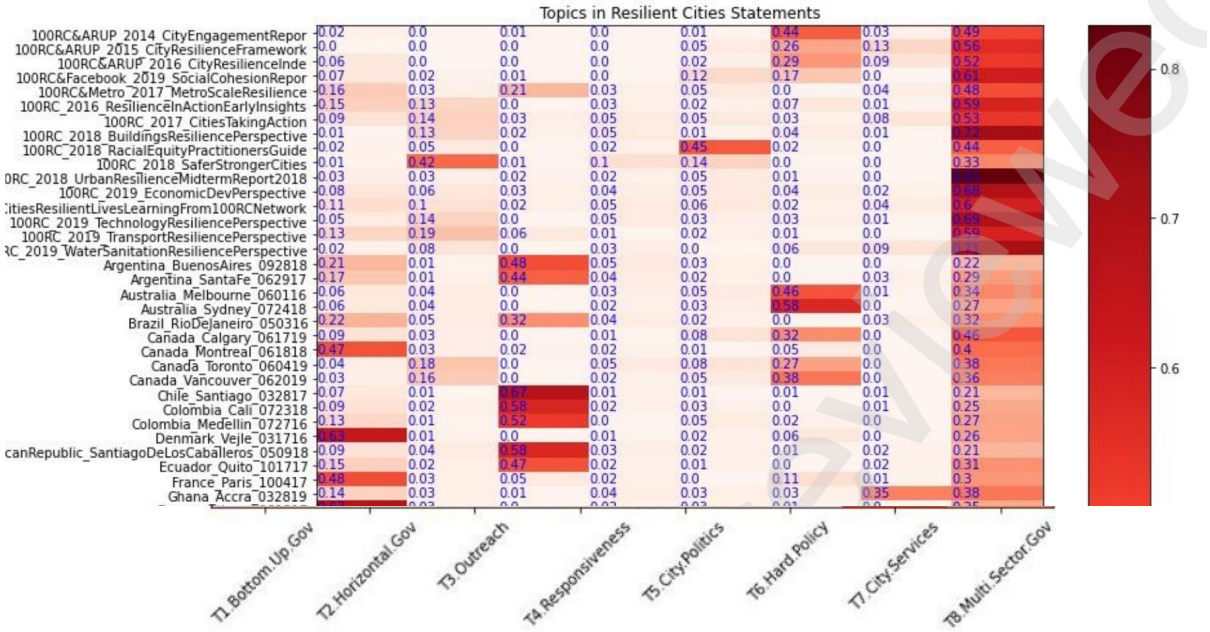


U.S. Environmental Politics (2/2)

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



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 Digital Humanities and Computational Social Science

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.

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?

For more information, see our Corpus Building Handout.

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. 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.

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

Our Text

We will use three metropolitan climate action plans published between 2023 and 2024:

- Greater Boston Priority Climate Action Plan, March 2024
- PlaNYC Getting Sustainability Done, April 2023
- Priority Climate Action Plan for the Chicago Metropolitan Statistical Area, March 2024

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-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.

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

Initial Corpus Analysis

Open any one of the texts from the sample corpus:

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

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

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

Exploratory Tool: Word Counter

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

Word Counter

- https://databasic.io/en/wordcounter/
- A user-friendly **basic word counting tool**
- Allows you to count words, bigrams, and trigrams in plain text files and to download spreadsheets with your results
- The max file upload is 10MB
- The default is to lowercase all words and apply stop words, but you can change those settings
- For more information, please see: https://bit.ly/handout-data-basics-suite

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

Word Counter Example

What seems significant in the most frequent terms from the Greater Boston Priority Climate Action Plan?

This is a **word cloud**, used to get a sense of the **most used words in a document**. Words used more often are bigger than those used less often.

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

“Tokenizing” text

Before words can be counted, they must be “tokenized” or divided into components that programs can treat as distinct segments. Different programs will have different standards for tokenization—this one uses both white spaces and punctuation marks (such as commas) to separate **words** into **tokens. What are some limitations of this approach?**

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

Data preparation

Go to the upload/paste screen for WordCounter and un-click the “ignore stop words” and “ignore case” options, then upload Boston’s climate action plan and count the words again.

What happened? Why do you think the default is to ignore stop words and remove differences between upper/lowercase words? Can you think of any limitations to this approach?

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

Bigrams and Trigrams

TOP WORDS		BIGRAMS		TRIGRAMS	
Word	Frequency	bigram	Frequency	trigram	Frequency
energy	398	in the	147	advisory group members	54
emissions	213	of the	117	in the region	39
municipal	183	https www	103	the greater boston	28
community	182	the region	91	https www mass	26
transportation	182	for the	82	www mass gov	26
program	172	advisory group	81	department of energy	20
programs	168	new hampshire	75	ghg emissions reductions	19
communities	162	ghg emissions	75	measure advisory group	19
buildings	155	renewable energy	62	climate action plan	18

In addition to single words, it is also useful to consider **bigrams** and **trigrams** which include additional context.

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Exploratory Tool: Word Tree

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

Word tree starting with “Boston” from Boston’s climate action plan

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Tools for corpus exploration: Voyant

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

Voyant

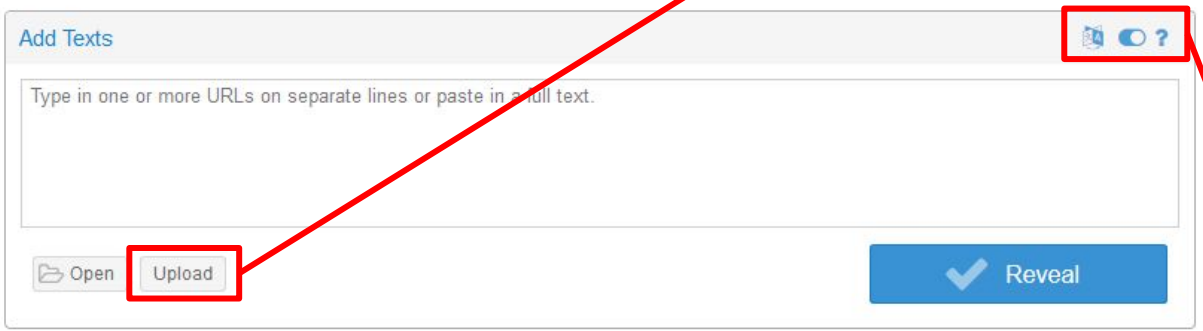
Voyant makes it possible to **perform analyses on one or multiple files in many ways**, including word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances.

https://voyant-tools.org/

For more information, see: https://bit.ly/handout-voyant-intro

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

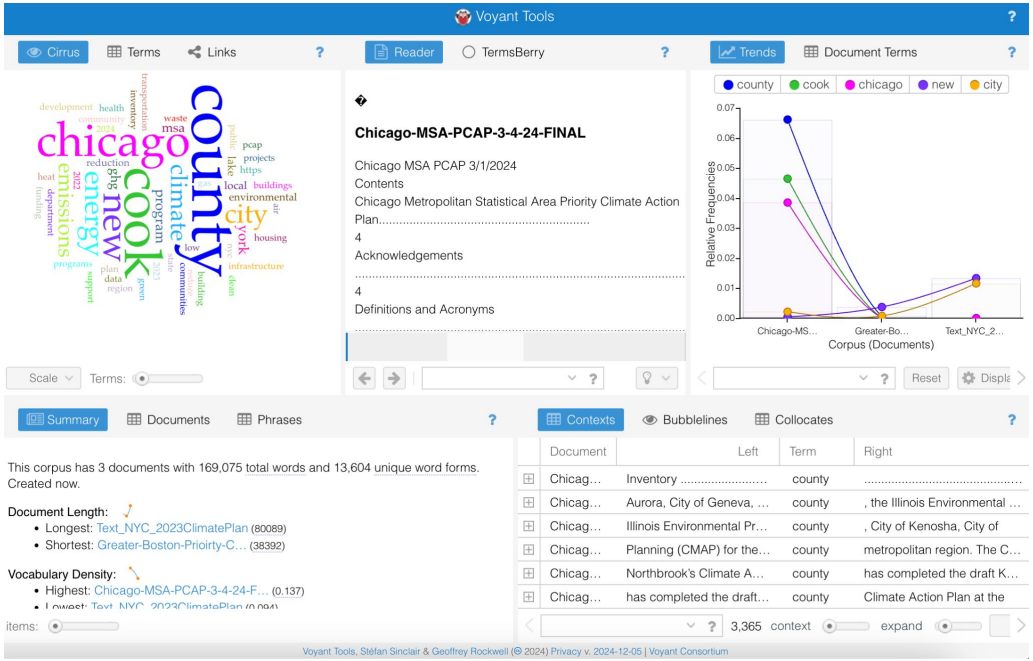
Feel free to ask questions at any point during the presentation!

Voyant: Dashboard

Results:

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

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



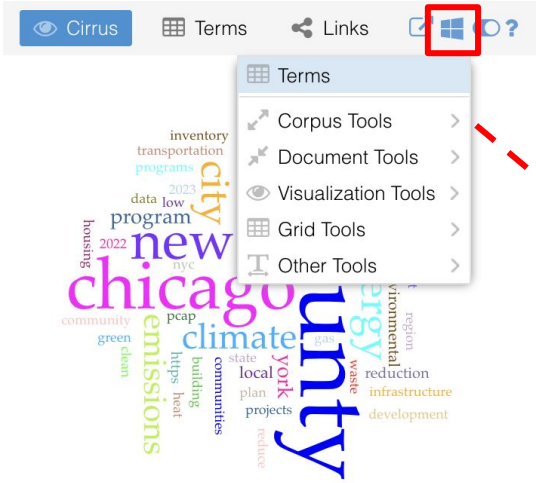
These boxes can all be changed!

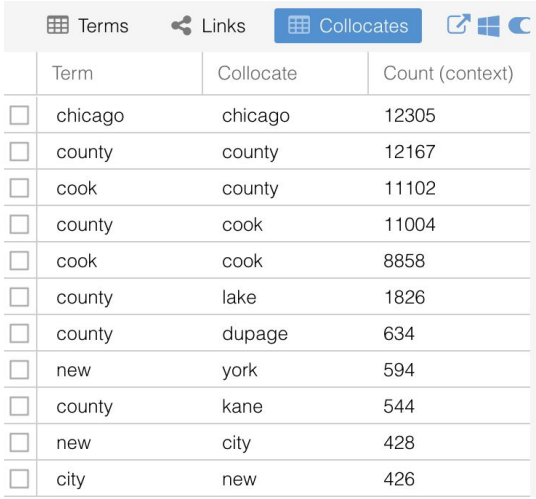
Feel free to ask questions at any point during the presentation!

Northeastern University
NULab for Digital Humanities and Computational Social Science

Voyant: Changing Displayed Results

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





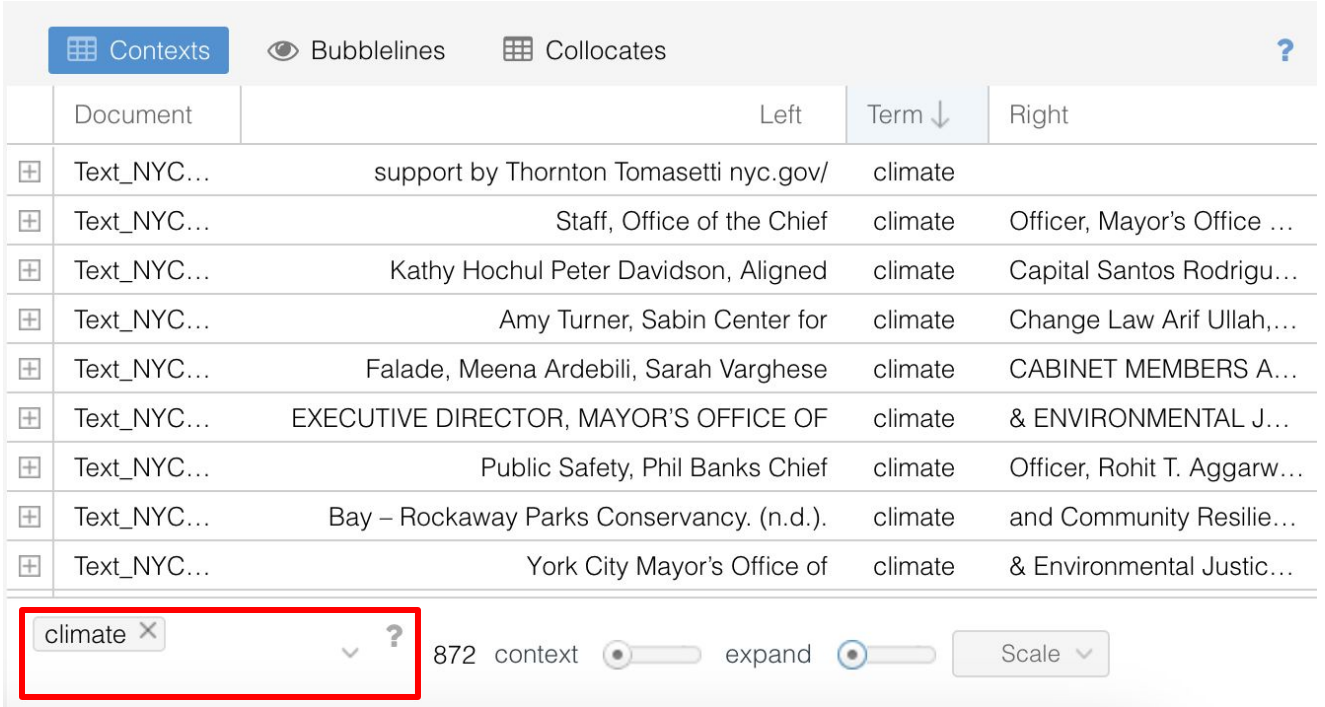
Term	Collocate	Count (context)
chicago	chicago	12305
county	county	12167
cook	county	11102
county	cook	11004
cook	cook	8858
county	lake	1826
county	dupage	634
new	york	594
county	kane	544
new	city	428
city	new	426

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: Tools for further exploration

- Voyant's Getting Started guide
- Voyant's List of Tools, showing all the features possible with Voyant including descriptions of each
- Some useful tools to explore:
  - MicroSearch
  - Topics
  - Correlations
  - Collocates Graph

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 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?

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

Tools for corpus exploration: 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

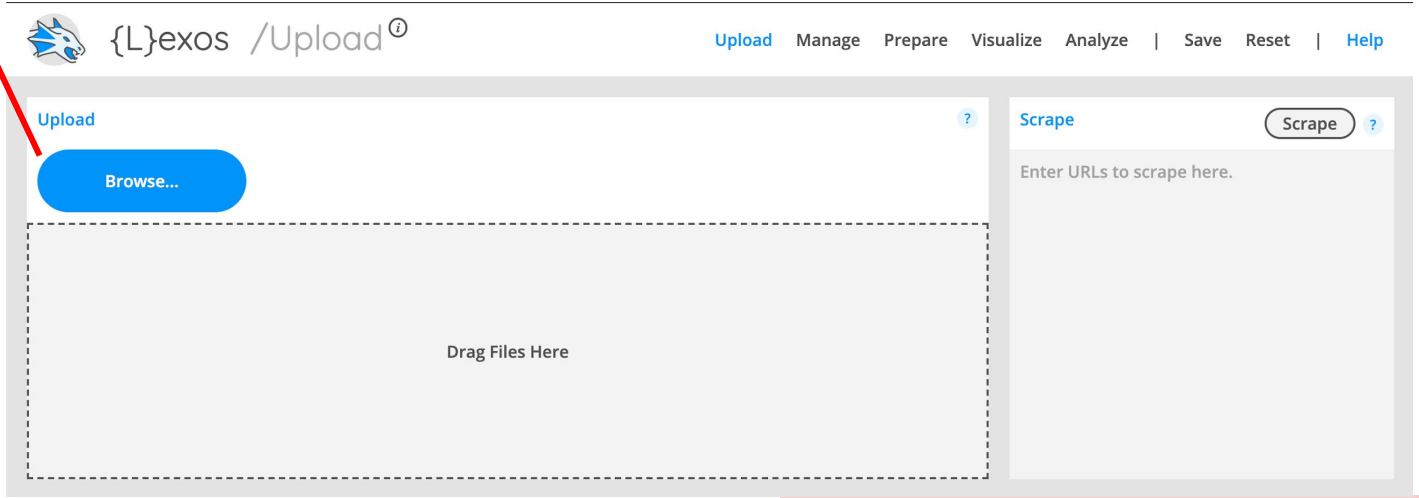
For more information, please see: https://bit.ly/handout-Lexos-intro

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.

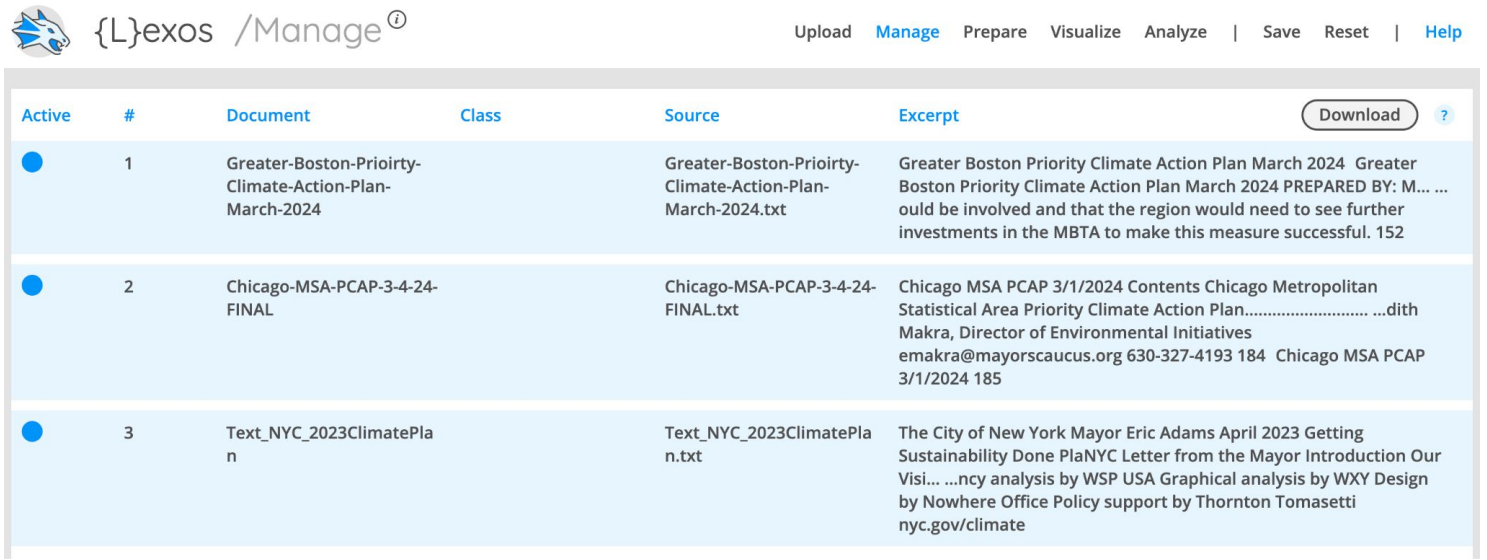


Northeastern University
NULab for Digital Humanities and Computational Social Science

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)



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)

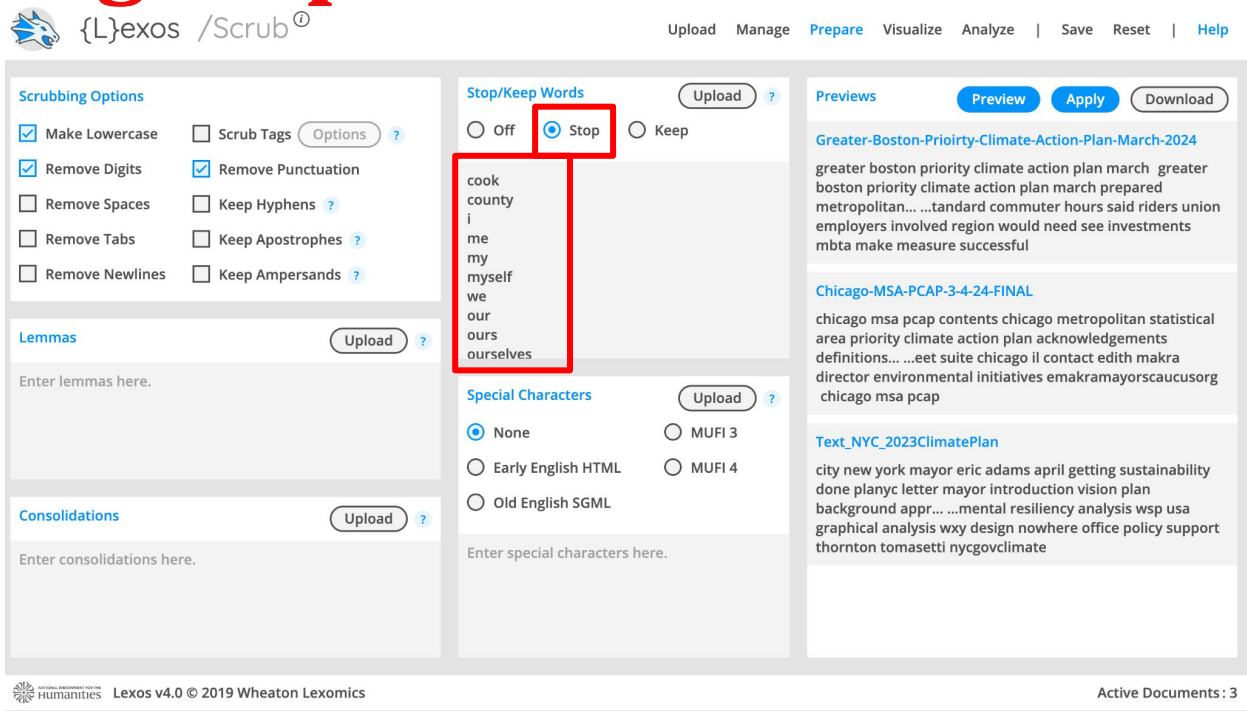
You can also stem words and remove certain words. Here are some possibilities:

- **Stop/Keep Words**: remove a list of words. Usually these would be **stop words**. With WordCounter, you had to use the stop words 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 Stop Words

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



You can also add stop words particular to your corpus.

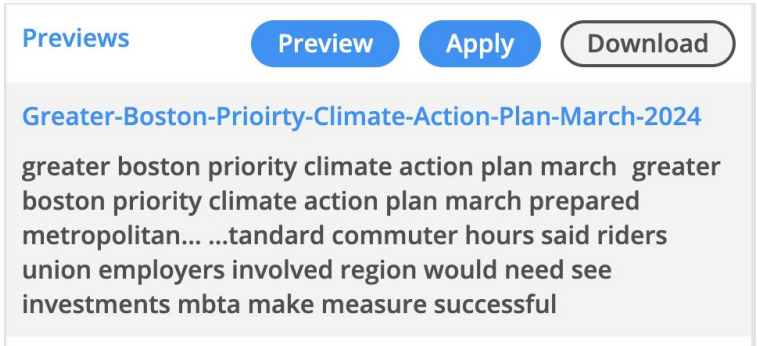
Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

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 Example

Top Words

Generate | Download | ?

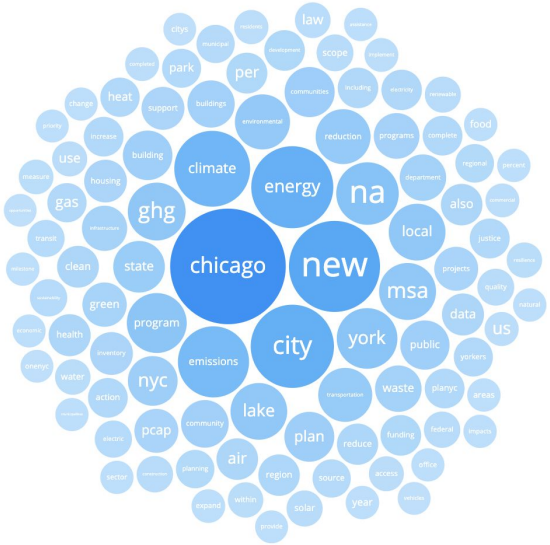
Document "Greater-Boston-Prioirty-Climate-Action-Plan-March-2024" Compared To The Corpus		Document "Chicago-MSA-PCAP-3-4-24-FINAL" Compared To The Corpus		Document "Text_NYC_2023ClimatePlan" Compared To The Corpus	
chicago	-21.7793	chicago	43.5797	chicago	-30.8436
city	-13.9094	lake	21.0948	msa	-15.1287
municipal	13.5272	msa	17.0505	lake	-14.9818
massachusetts	12.8756	new	-16.9081	new	14.0815
na	-11.712	pcap	14.1964	city	13.0095

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

Lexos: Visualize

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

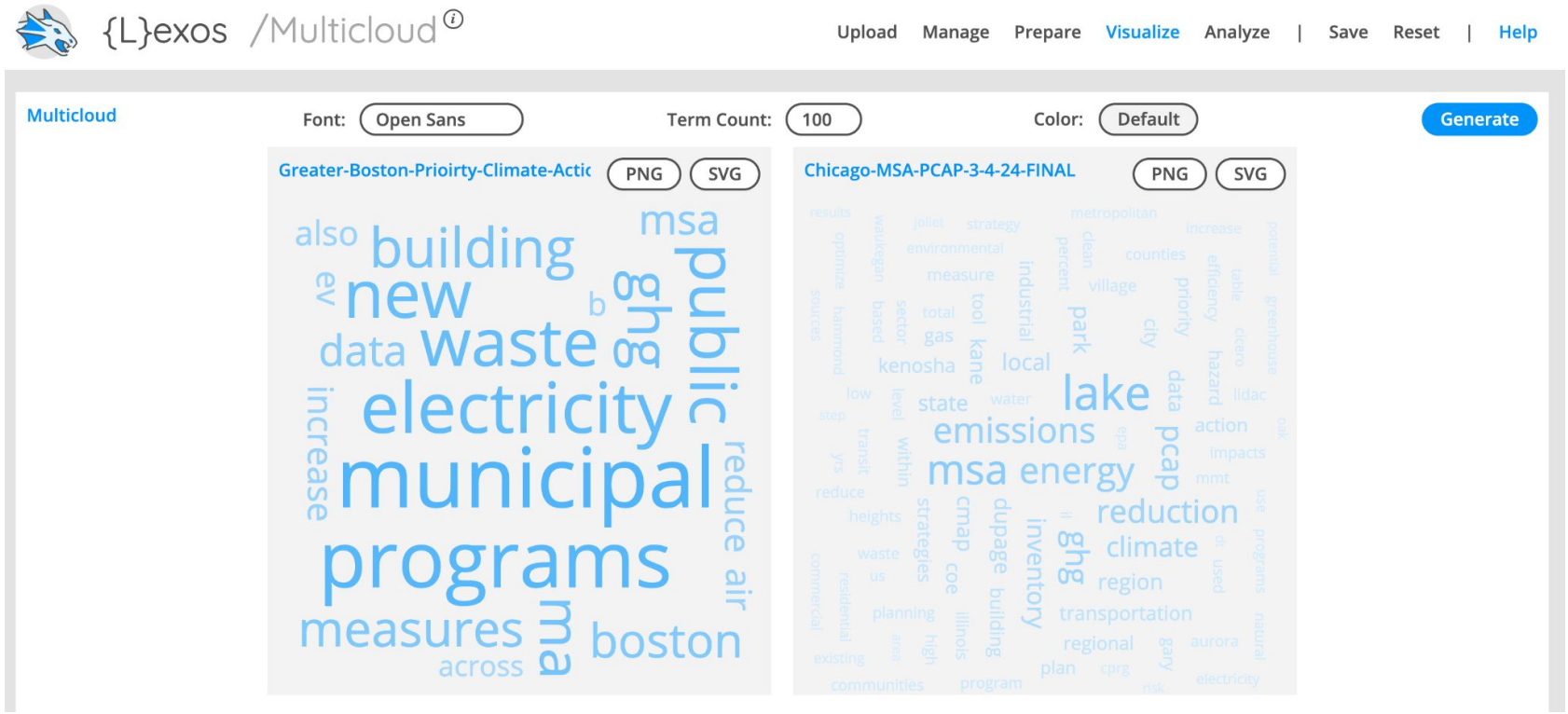
Bubbleviz: visualize word counts through bubbles across the entire text/corpus.



Northeastern University
NULab for Digital Humanities and Computational Social Science

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

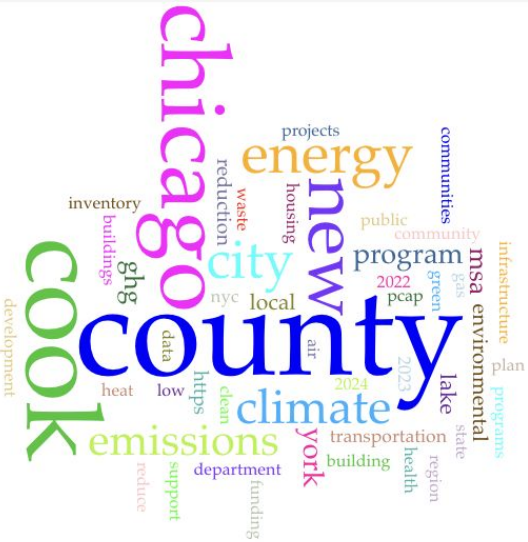
Lexos: Visualize > Multicloud



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

Voyant vs. Lexos: Word clouds

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



Lexos Word cloud

What could be causing this distinction?

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Lexos: Rolling Window

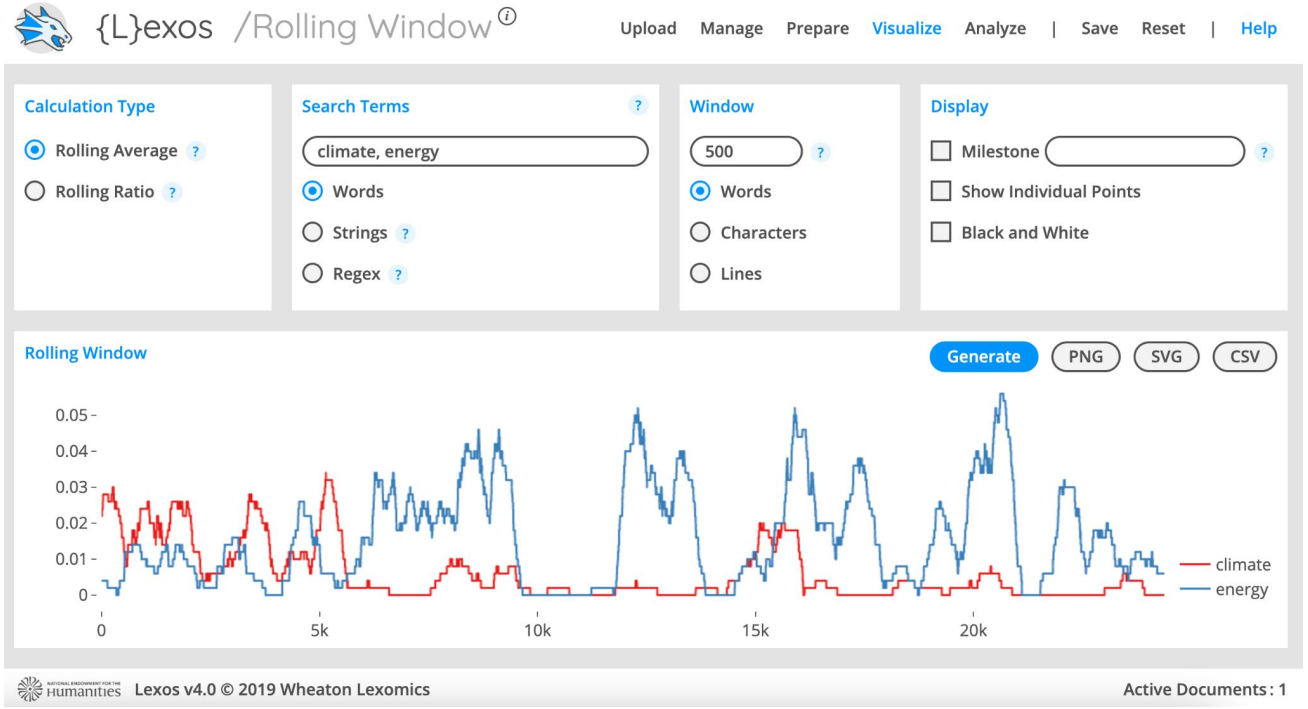
Rolling windows allow you to look at word trends across **one** document. To use a rolling window, first select a single text in the "Manage" screen, then:

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 (climate, action)
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 Boston's climate action plan, and searching for the words 'climate' and 'energy' with a window of 500, we can get an idea of how these terms work together in the document.



Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Lexos: Analyze > Dendrogram

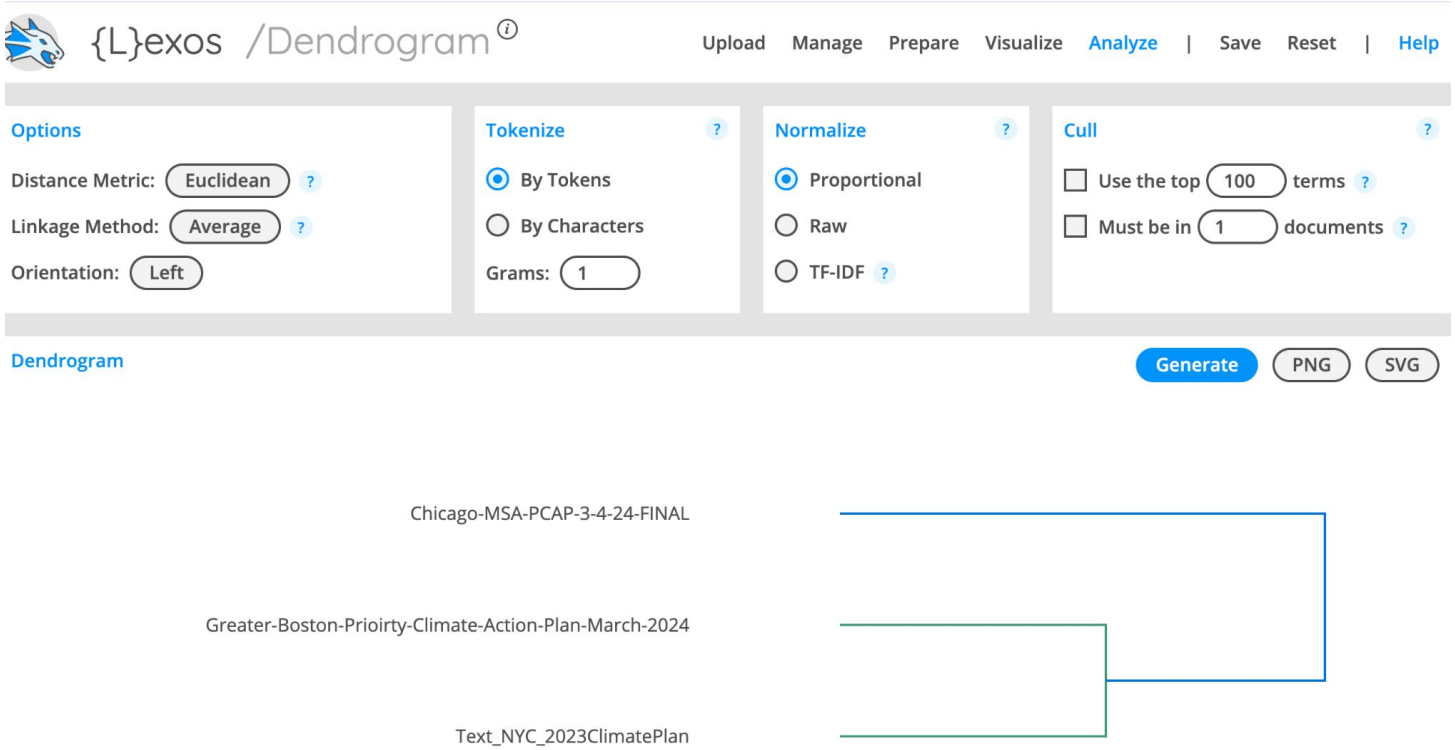
The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms are able to show the hierarchy between objects. Dendrograms show:

- 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.



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—and you can even use those downloaded text files with other tools!

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 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?

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

A Brief Introduction to Web Scraping

Slide content courtesy of Alyssa Smith

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.

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

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.

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

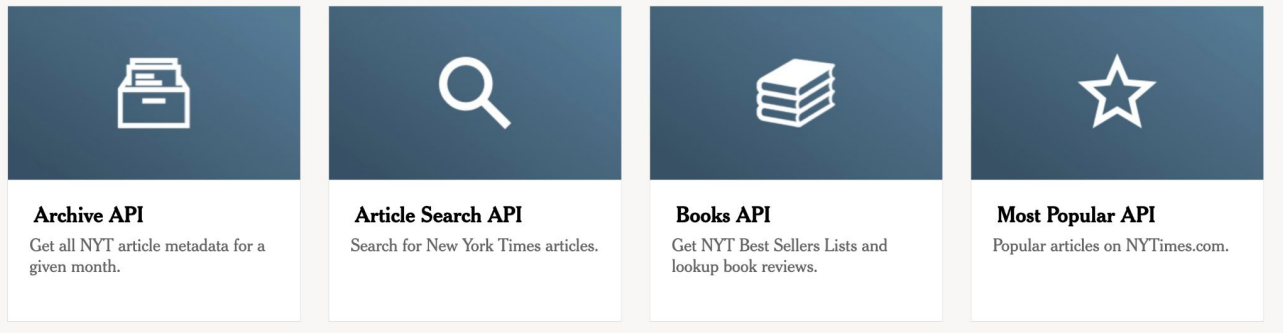
Access Web Data Through APIs

- 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.

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

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.



NYT Article Search API

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

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!)

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

Data privacy

- It's important to pay attention to data privacy when using digital resources
- At its simplest, **data privacy** is a person's ability to control what of their personal information is shared and with whom.
- To help you make informed decisions about interacting with digital tools in ways that honor your boundaries with your data and/or personal information, The DITI has prepared a handout on **Data Privacy**

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

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.

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

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

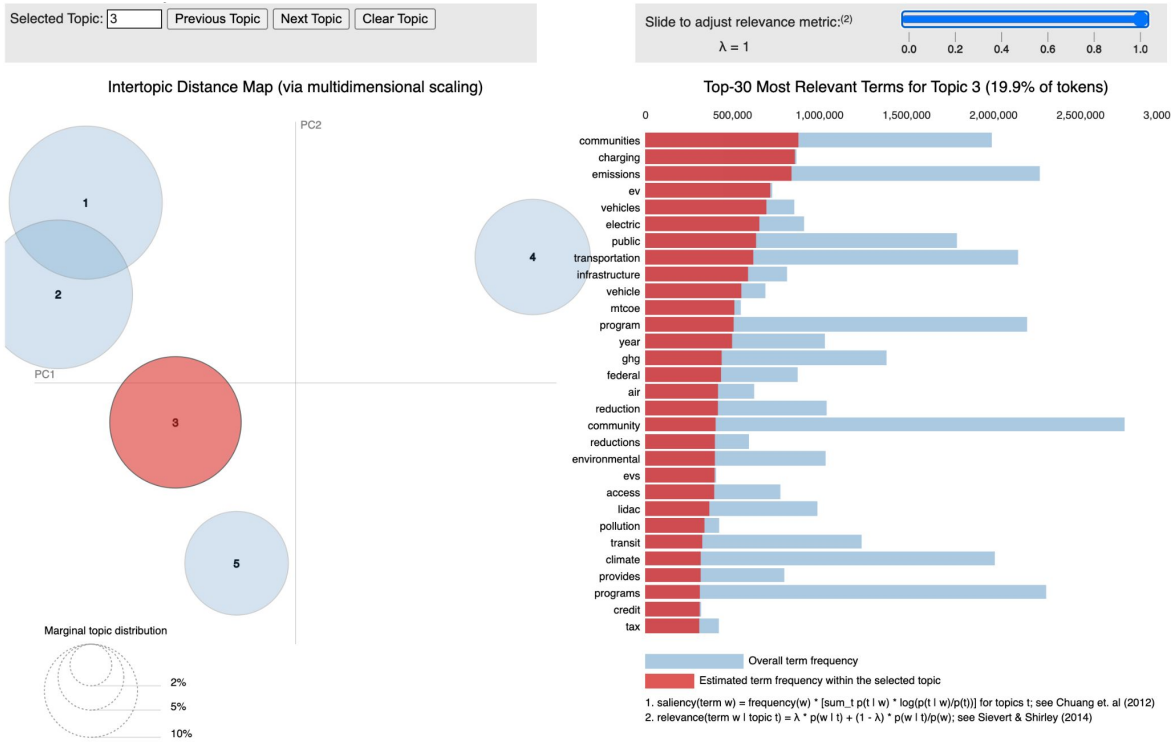
Feel free to ask questions at any point during the presentation!

Further Exploration

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

Further exploration: Topic Modeling

Topic modeling is a machine learning method that uses word co-occurrence within documents to identify "topics," or clusters of related terms. This is a topic model based on the Greater Boston Priority Climate Action Plan. In the visualization, topic 3 is selected.

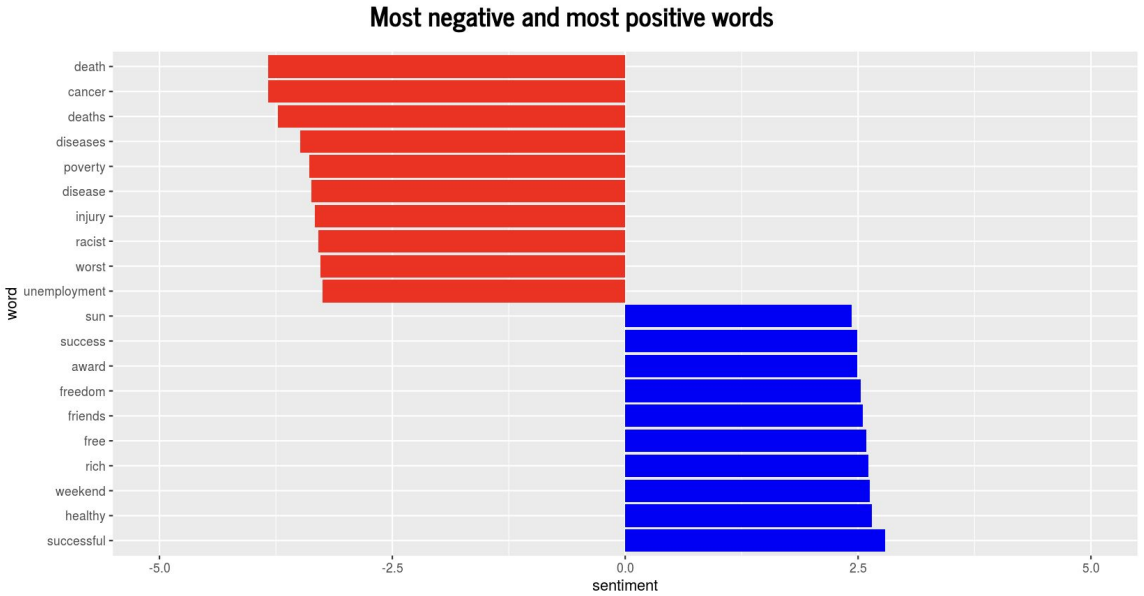


Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Further exploration: sentiment analysis

Sentiment analysis uses dictionaries, and sometimes machine learning, to assign sentiment scores (e.g., positive and negative) to documents. You can try this out with the "Drag and Drop Sentiment Analysis" tool.



Greater Boston Priority Climate Action Plan

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

For further exploration

DITI handouts on building a corpus and more links and resources for text analysis

NULab list of resources for text analysis

Programming Historian tutorials

"Data-Sitters' Club" tutorials

Library subject guides on text mining and analysis: guide on getting started, guide on vendor policies

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

Thank you!

—**Developed by** Cara Marta Messina, Juniper Johnson, Sara Morrell, Ayah Aboelela, and Jeff Sternberg

- For more information on 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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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