Introduction to Computational Text Analysis

Professor Griffin Zimmerman
ENGW1111, Spring 2025
Sean P. Rogers and Sara Morrell
Digital Integration Teaching Initiative (DITI)

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

For more information, please see: https://bit.ly/handout-text-resources
For all module materials, check out:
https://github.com/NULabNortheastern/digitalassignmentshowcase/tree/main/text-analysis/sp25-zimmerman-engw1111-textanalysis

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

What is Computational Text Analysis?

Computational Text Analysis

Computational text analysis refers to the **array of methods used to "read" texts with a computer.** It is similar to statistical analysis, but the data is texts (words) instead of numbers.

Text analysis:

- Involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- Includes methods such as word count frequency, keywords in context, computational modeling (with machine learning), and sentiment analysis.
- Is conducted using web-based tools or coding languages like Python and R.

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

Why Computational Text Analysis?

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

For example: "Gendered Language in Teacher Reviews" by Ben Schmidt shows stark differences in the ways that male and female professors are reviewed on "Rate My Professor."

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

Language Used in Climate News

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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

Climate News: Discussion

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

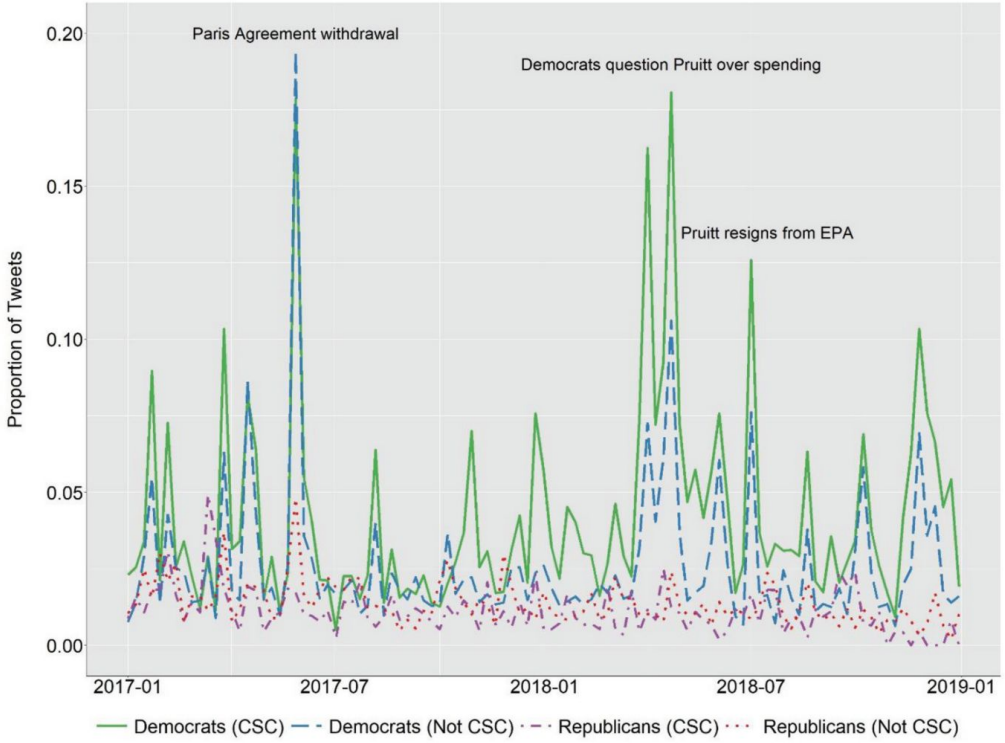
- What do you notice about the TV coverage of these terms over time? What is surprising?
- How do you think political values affects climate language?
- How might this language shape policies?

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

U.S. Environmental Politics (1/2)

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

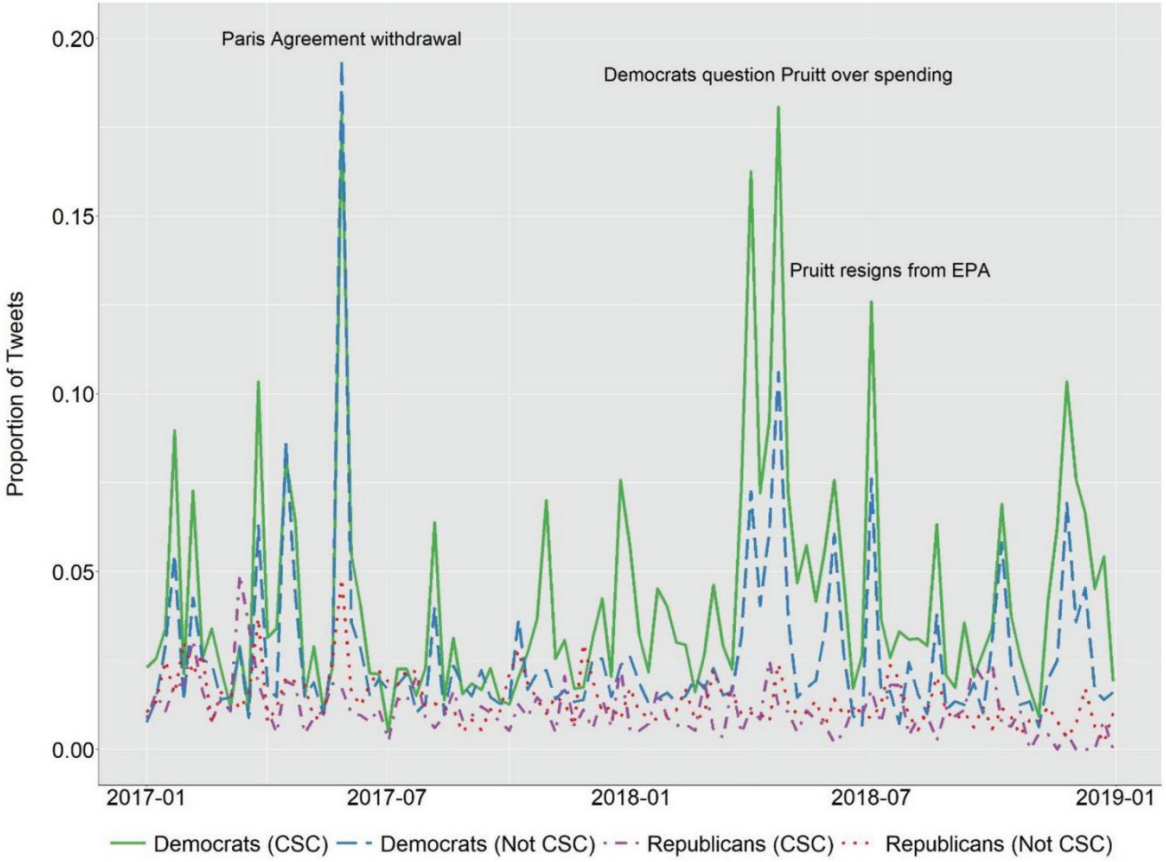
[Key events and challenges: a computational text analysis of the 115th house of representatives on Twitter](#) - Jeremiah Bohr in Environmental Politics (2021), 30 (3): 399-422



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

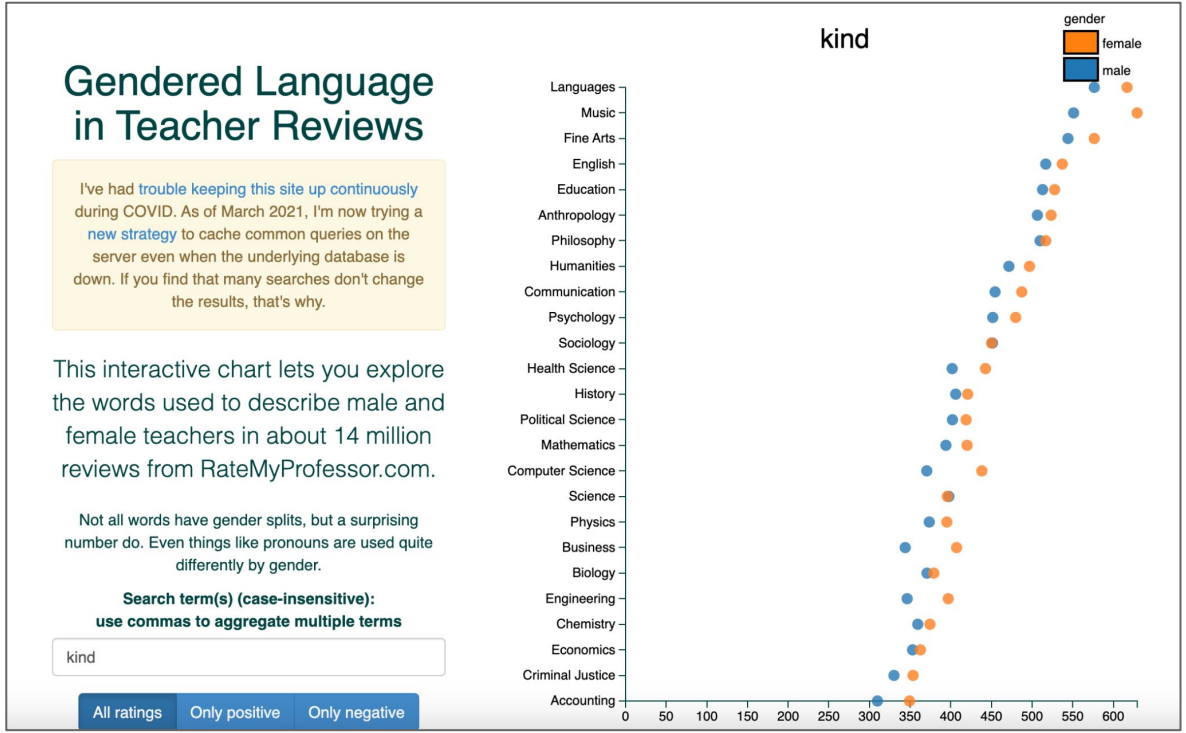
U.S. Environmental Politics (2/2)

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



Northeastern University
NULab for Digital Humanities and Computational Social Science

Gendered Language



Go to bit.ly/schmidt-gender and try a few queries. For example:

- Smart
- Ditzy
- Unprofessional
- Nice

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

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

Key Terms

- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stop words include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **nGram:** A continuous sequence of *n* items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'

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 on building a corpus, see **this handout**.

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

Our Corpus

For our corpus, we will work with a set of State of the Union addresses from 1990 to 2019.

https://drive.google.com/drive/folders/1-1at6fwDylv4GKc7s4N6nNuosN_9u0Kl?usp=sharing

The easiest way to work with these files is to choose "Download all" and open them with a plain-text editor (TextEdit on Mac, Notepad on Windows). Mac users should be able to click on the zip file to expand it; Windows users will need to right-click and choose "Extract all."

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 Tools: Word Counter and Word Trees

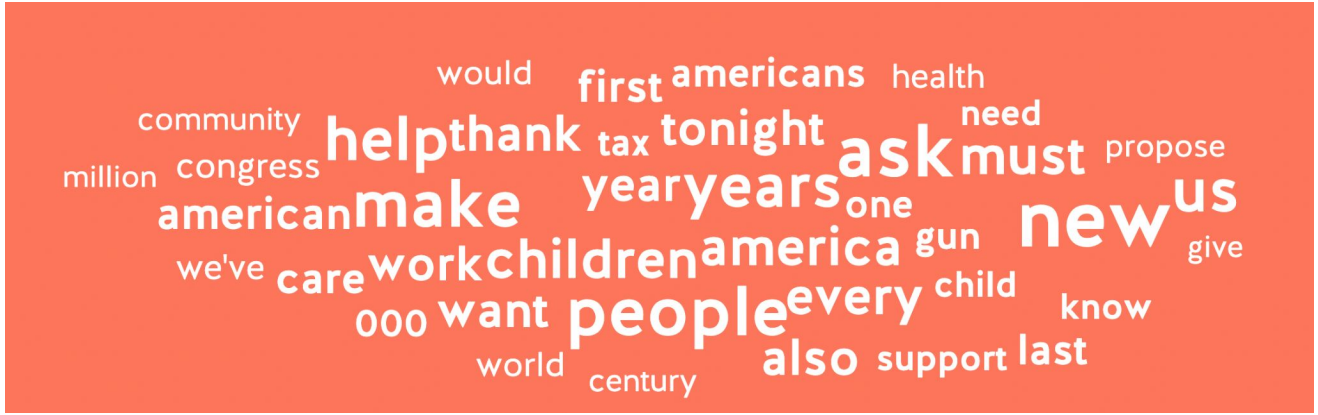
Word Counter

- 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 stopwords, 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

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.



What seems significant in the most frequent terms from Clinton's 2000 State of the Union Address?

Northeastern University
NULab for Digital Humanities and Computational Social Science

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

“Tokenizing” text

Why do you think that “000” is one of the most common words in Clinton’s 2000 SotU address? Open the .txt file and search for “000” to check your guess.

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 back to the upload/paste screen for WordCounter and un-click the “ignore stopwords” and “ignore case” options, then count the words again.

What happened? Why do you think the default is to ignore stopwords 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

In addition to single words, it is also useful to consider **bigrams** and **trigrams**. Why do you think the phrase "I ask you" appears so often in the 2000 State of the Union Address? What about "we should"?

TOP WORDS		BIGRAMS		TRIGRAMS	
Word	Frequency	bigram	Frequency	trigram	Frequency
new	47	in the	40	i ask you	23
ask	43	i ask	32	ask you to	23
people	40	ask you	30	i want to	15
make	38	you to	30	want to thank	10
years	35	of the	27	tonight i propose	8
us	35	of our	26	thank you for	7
help	35	we should	24		

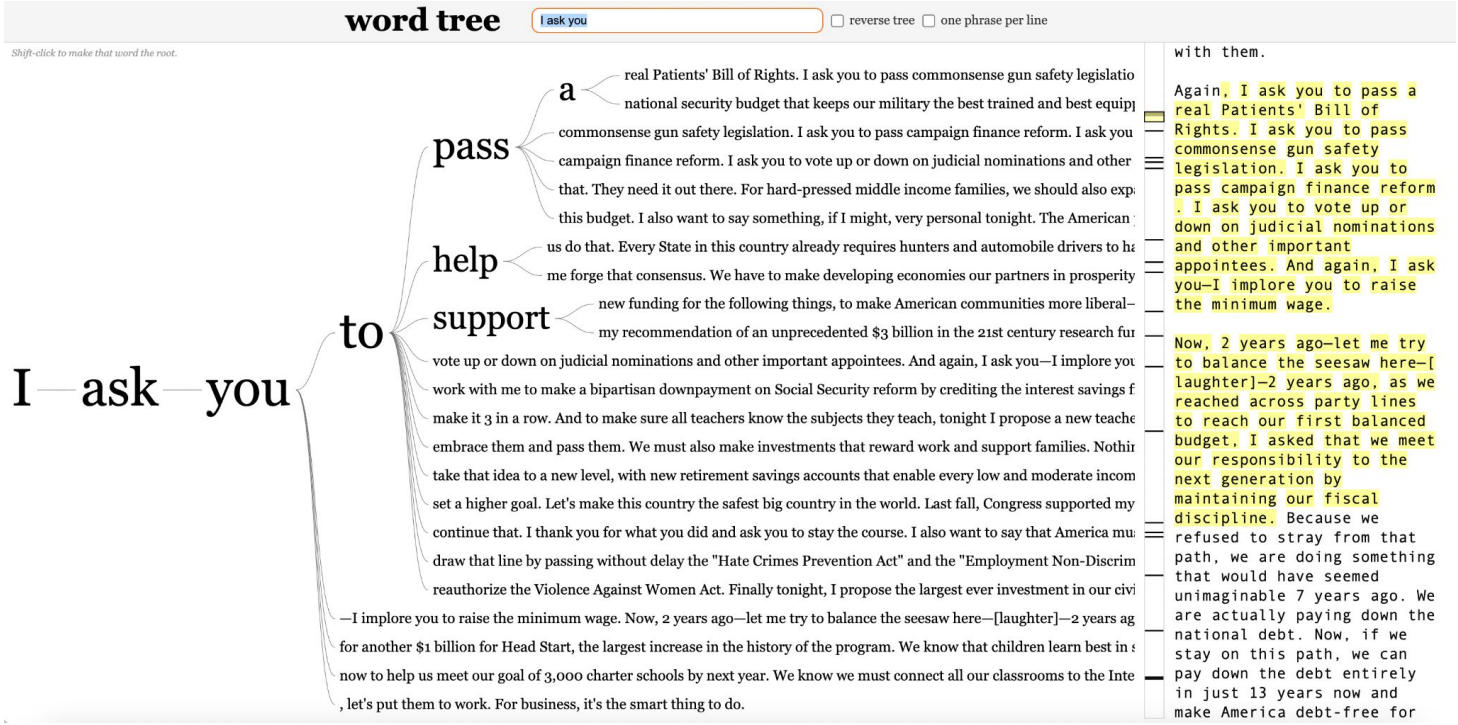
Feel free to ask questions at any point during the presentation!

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



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

Tools for corpus exploration: Voyant

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

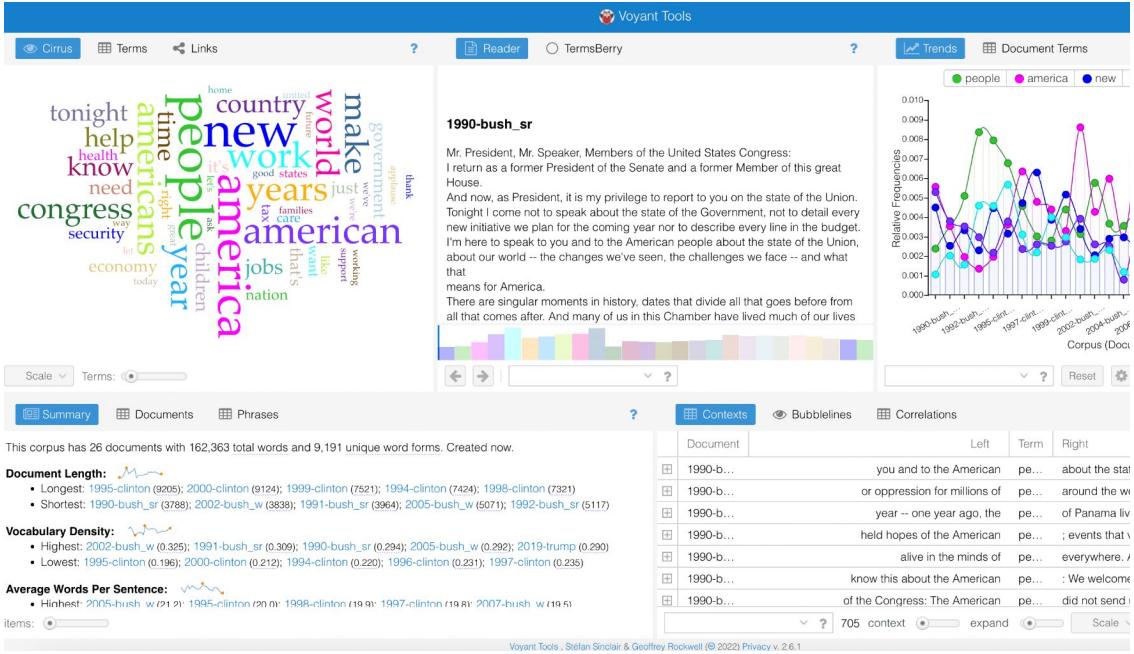
Feel free to ask questions at any point during the presentation!

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

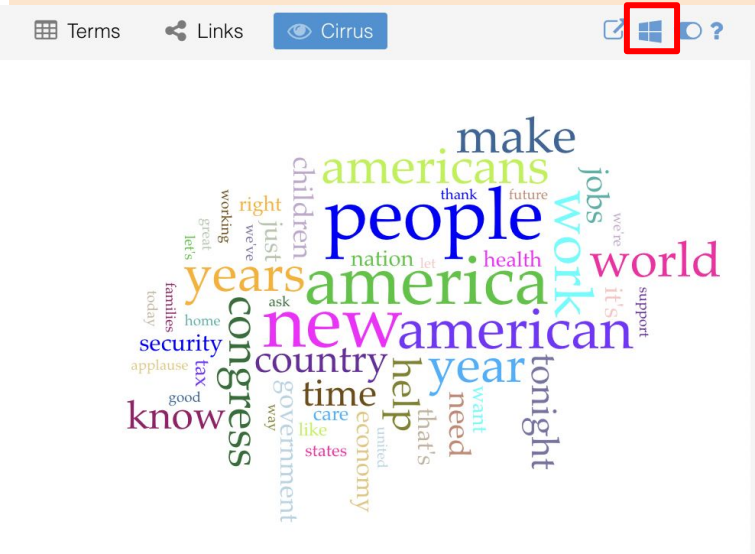


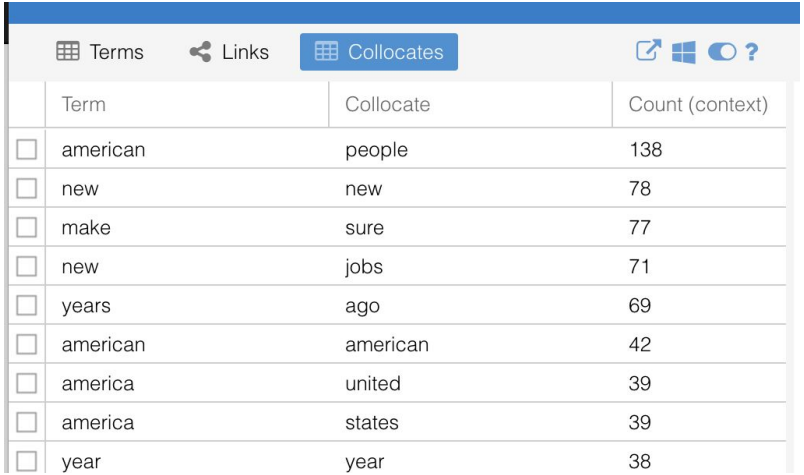
These boxes can all be changed!

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

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)
american	people	138
new	new	78
make	sure	77
new	jobs	71
years	ago	69
american	american	42
america	united	39
america	states	39
year	year	38

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!

Tools for corpus exploration: Lexos

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

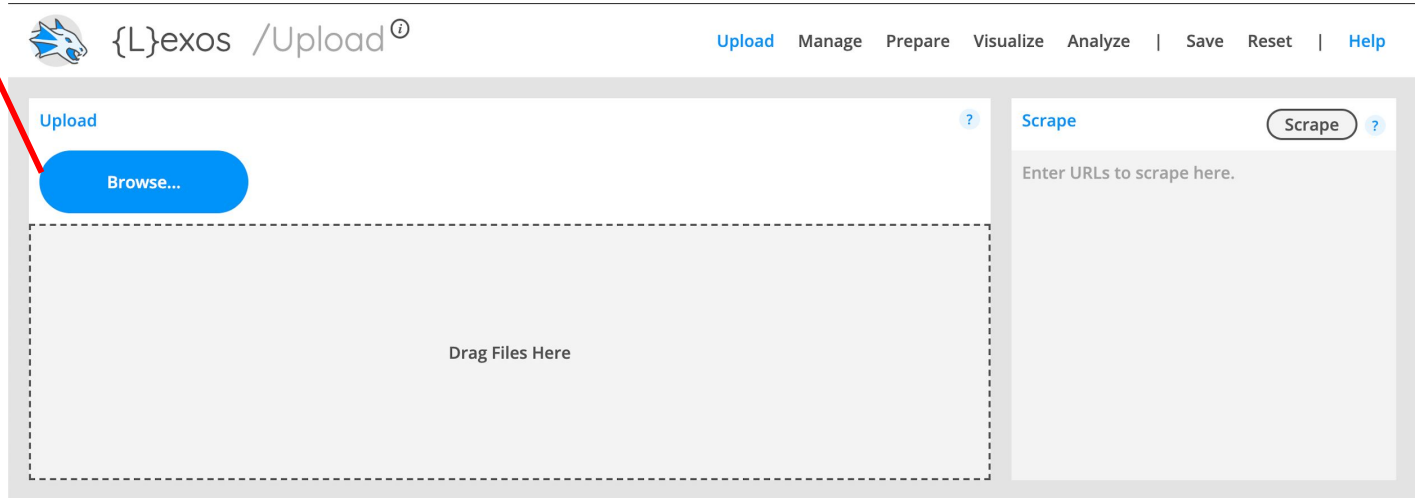
For more information, please see: https://bit.ly/handout-Lexos-intro

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)

{L}exos /Manage

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

Active	#	Document	Class	Source	Excerpt	Download ?
●	3	2018-trump		2018-trump.txt	The President. Mr. Speaker, Mr. Vice President, Members of Congress, the First Lady of the United States, and my fellow America... ...er. And our Nation will forever be safe and strong and proud and mighty and free. Thank you. And God bless America. Goodnight.	
●	4	2019-trump		2019-trump.txt	Madam Speaker, Mr. Vice President, Members of Congress, the First Lady of the United States — (applause) — and my fellow Americ... ...ong all the nations of the world. Thank you. God bless you. And God bless America. Thank you very much. Thank you. (Applause.)	
●	5	2013-obama		2013-obama.txt	Please, everybody, have a seat. Mr. Speaker, Mr. Vice President, Members of Congress, fellow Americans: Fifty-one years ago, Jo... ... thors of the next great chapter of our American story. Thank you. God bless you, and God bless these United States of America.	
●	6	2014-obama		2014-obama.txt	The President. Mr. Speaker, Mr. Vice President, Members of Congress, my fellow Americans: Today in America, a teacher spent ext... ...es cast toward tomorrow, I know it is within our reach. Believe it. God bless you, and God bless the United States of America.	

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)

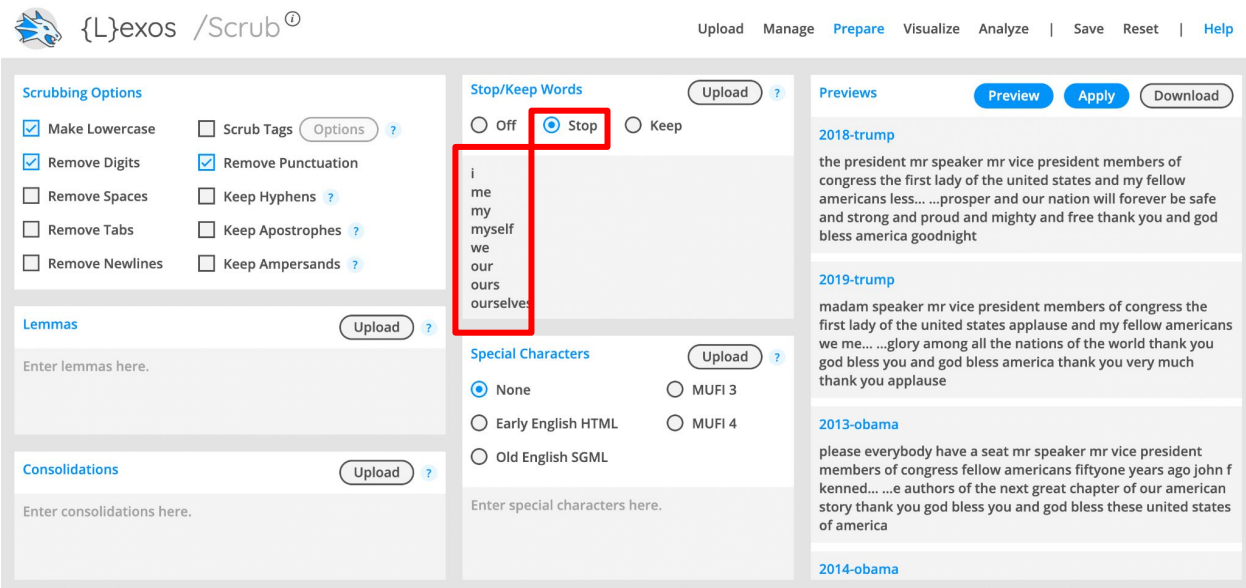
You can also stem words and remove certain words. Here are some possibilities:

- **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. Copy and paste the stopwords (hit "raw", then select all and copy) 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

2013-obama

Please, everybody, have a seat. Mr. Speaker, Mr. Vice President, Members of Congress, fellow Americans: Fifty-one years ago, Jo... ...thors of the next great chapter of our American story. Thank you. God bless you, and God bless these United States of America.

2014-obama

The President. Mr. Speaker, Mr. Vice President, Members of Congress, my fellow Americans: Today in America, a teacher spent ext... ...es cast toward tomorrow, I know it is within our reach. Believe it. God bless you, and God bless the United States of America.

AFTER PREP

2013-obama

please everybody have a seat mr speaker mr vice president members of congress fellow americans fiftyone years ago john f kenned... ...e authors of the next great chapter of our american story thank you god bless you and god bless these united states of america

2014-obama

the president mr speaker mr vice president members of congress my fellow americans today in america a teacher spent extra time... ...ur eyes cast toward tomorrow i know it is within our reach believe it god bless you and god bless the united states of america

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.

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

Lexos: Analyze > Top Words

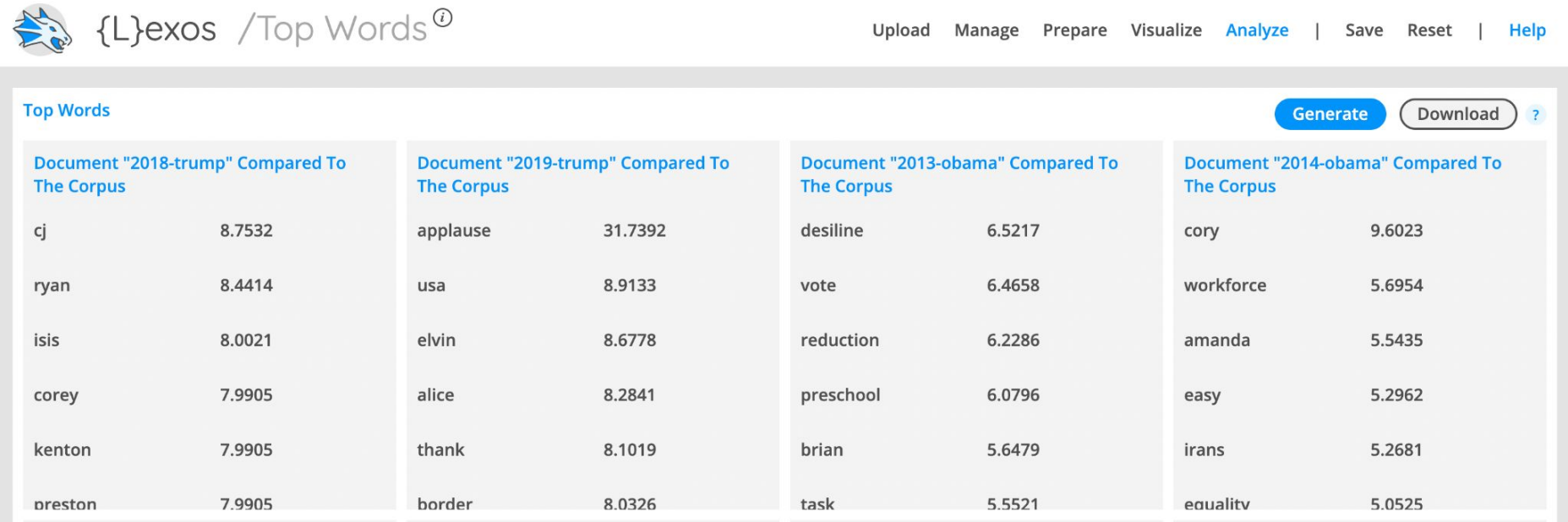
The top words tool lets you compare word usage between individual documents and your corpus as a whole. If you want to make more specific comparisons, you can also assign “classes” to subsets of tools with the “Manage” screen.

- 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



Northeastern University
NULab for Digital Humanities and Computational Social Science

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

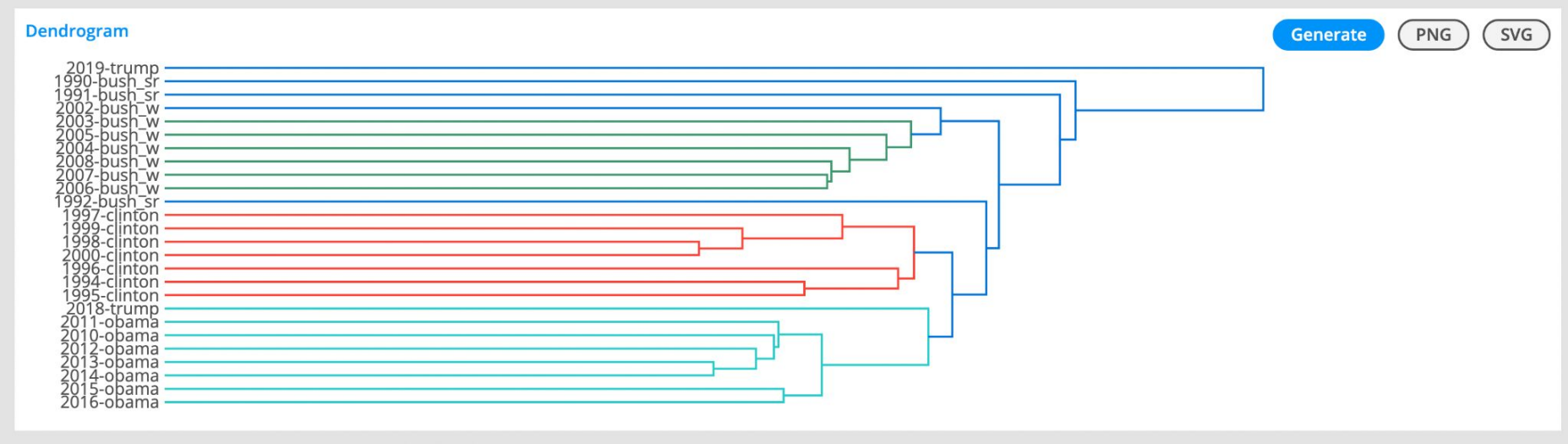
Lexos: Analyze > Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms “cluster” texts to draw out similarities:

- 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: Analyze > Dendrogram Example



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).

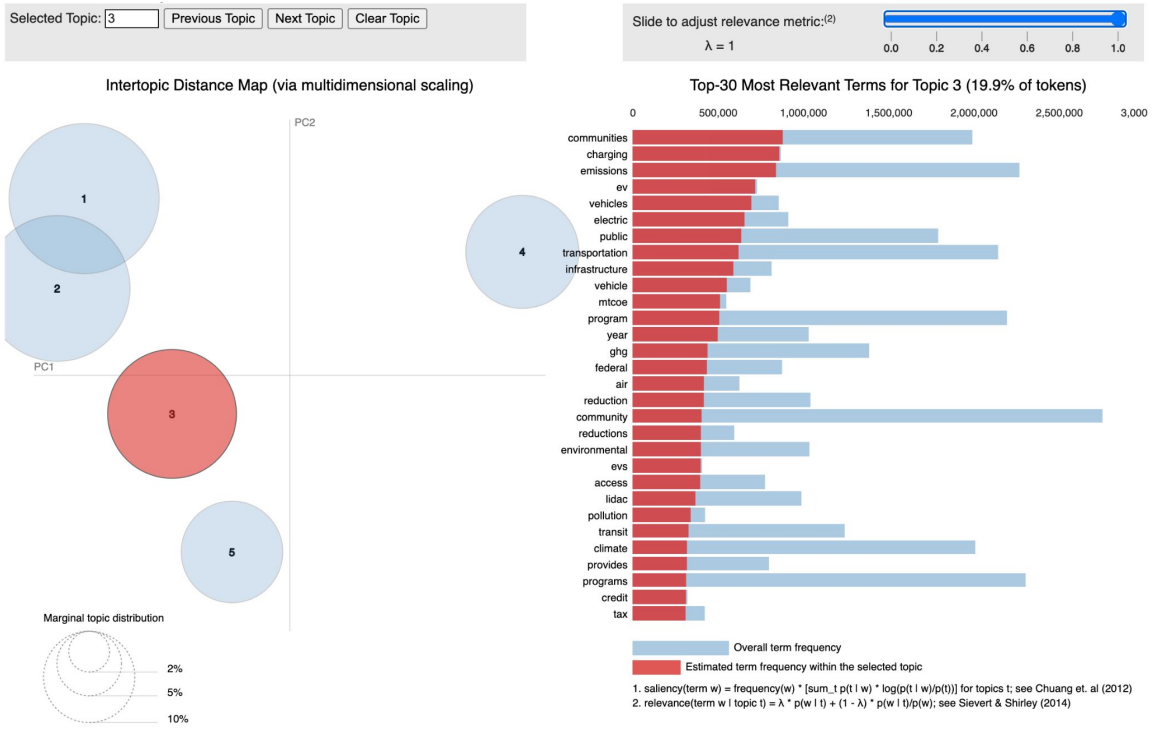
Finally, if you want to start over, you can "Reset" your Lexos dashboard.

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

Further Exploration

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.



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

Northeastern University
NULab for Digital Humanities and Computational Social Science

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.

Top 50 positive words

Show 10 entries

	word	sentiment
1	happy	8.3
2	laugh	8.22
3	excellent	8.18
4	laughs	8.18
5	rich	7.98
6	free	7.96
7	beauty	7.76
8	wonderful	7.76
9	excited	7.62
10	hope	7.38

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!

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, Jeff Sternberg, Claire Lavarreda, Sara Morrell, Sean Rogers

- 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