Computational Text Analysis for Content Analysis

By Vaishali Kushwaha, Yana Mommadova, Margarida Rodrigues, and Yunus Emre Tapan

Digital Integration Teaching Initiative (DITI)

For ENGL 1400: Intro to Literary Studies

Mary Loeffelholz

Fall 2022

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Introduction to definitions and key terms in computational text analysis (What)
- Discussion on its applications and uses in research (Why)
- Demonstration of web-based text analysis tools (How)
  - Word Counter, Word Trees, Voyant, Lexos

Slides, handouts, and data available at:

https://bit.ly/fa22-loeffelholz-textanalysis

Northeastern University
NULab for Texts, Maps, and Networks

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

Computational Text Analysis

Text analysis is **a process to make inferences based on textual data**. 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, nGrams, and sentiment analysis.
- Is conducted using web-based tools or coding languages like Python and R

Northeastern University
NULab for Texts, Maps, and Networks

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 formal continuities or discontinuities in literary genres, or textual similarities across genres.** For example, computational tools reveal textual similarities between detective fiction and science fiction over long periods of time.

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

Key Terms

- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stop words include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **nGram:** A continuous sequence of *n* items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.
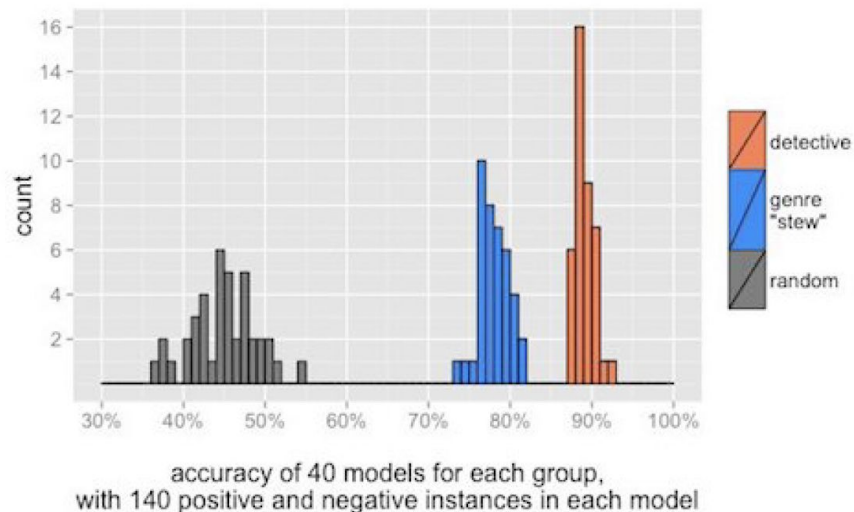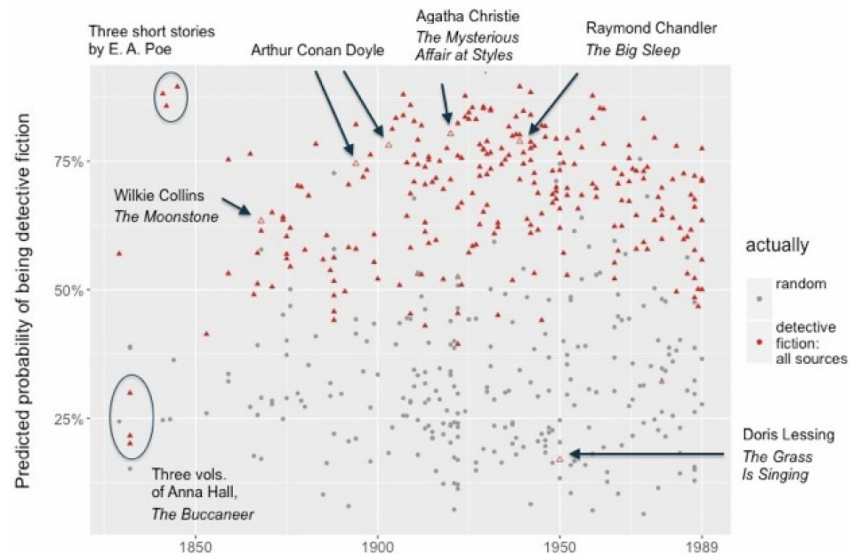
Feel free to ask questions at any point during the presentation!

Examples from Practice

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

Key Topics in Critical Literature Studies

Histogram plotting the accuracy of 40 models for three different putative "genres." For each model, 140 positive instances were selected randomly from a longer list.

Predicted probabilities of texts coming from the "detective fiction" genre. This analysis shows how likely a text is of being classified as detective fiction over the years.

Ted Underwood, "The Life Cycles of Genres," *Cultural Analytics* May 23, 2016.

Northeastern University
NULab for Texts, Maps, and Networks

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

Posts mentioning 'Black lives matter' spiked on lawmakers' social media accounts after the death of George Floyd

- [Pew Research Center July 16, 2020 article](#)
- [Methodology](#)

In weeks following George Floyd killing, Democratic lawmakers' most distinctive language on social media focused on racial justice, police violence

Share of members in each party that mentioned ___ on Twitter or Facebook, May 25-June 14, 2020
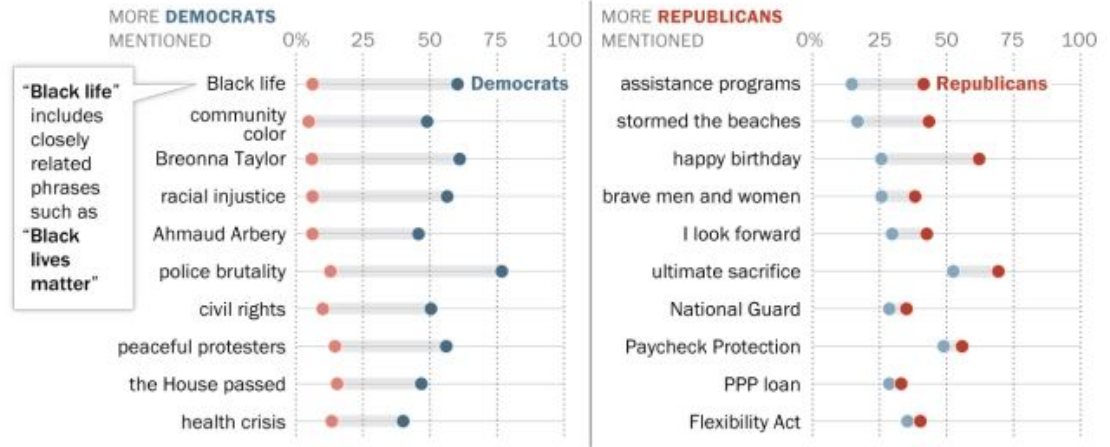
Note: Chart shows the top 10 keywords based on how much more likely members of one party were to ever mention a keyword relative to the other party. Terms are displayed in their standardized form (e.g., "Black life" instead of "Black lives") and have been edited slightly in some cases for readability (e.g., "the House passed" instead of "house passed"). Keyword analysis was not case-sensitive. Words from retweets are included in this analysis even if the member who retweeted them did not create the original tweet.

Source: Pew Research Center analysis of congressional social media data from the Twitter API, Facebook Graph API and CrowdTangle, May 25-June 14, 2020.

PEW RESEARCH CENTER

Northeastern University
NULab for Texts, Maps, and Networks

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. 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.

Northeastern University
NULab for Texts, Maps, and Networks

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

Our Corpus: source files

For our corpus, we will work with the following texts:

Frederick Douglass, *Narrative of the Life of Frederick Douglass, an American Slave*

Benjamin Franklin, *Autobiography of Benjamin Franklin*

Harriet Jacobs, *Incidents in the Life of a Slave Girl*

Jonathan Edwards, *Personal Narrative*

Herman Melville, *Benito Cereno*

Northeastern University
NULab for Texts, Maps, and Networks

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

Sample Corpus

The following .txt files are available on:

https://drive.google.com/drive/folders/11NQXGeLec9AV8DRC-DJ6d4f2ft16AVcB

Northeastern University
NULab for Texts, Maps, and Networks

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!
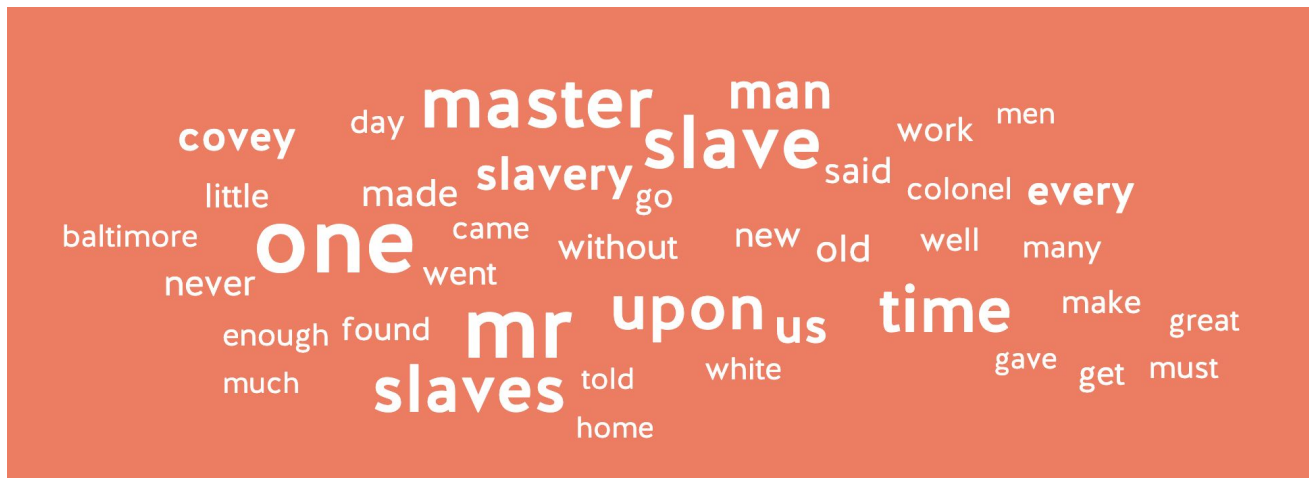
Word Counter

- 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 stopwords
- It can be run with and without stopwords

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

Word Counter Examples

This is a **word cloud**. It is helpful to get a sense of the **most used words in a document**. Words used more often are bigger, and ones used less often are smaller.

Word cloud from *Narrative of the Life of Frederick Douglass,* by F. Douglass

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

Word Counter Examples

TOP WORDS

Word	Frequency
one	181
mr	171
slave	143
master	140
slaves	123
time	121
upon	118
man	93
us	85

Shows the top words in the text.

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

bigram❓	Frequency
of the	312
in the	202
i was	144
to the	142
to be	110
it was	108
he was	97
at the	86

trigram❓	Frequency
st michael s	21
as well as	20
colonel lloyd s	19
one of the	17
it was a	16
my old master	15
he was a	14
the end of	14

The top trigram is St. Michael's, a town in Talbout County, Maryland, the location of Colonel Lloyd's plantation, where Douglass was enslaved. Note that Colonel Lloyd's is also a top trigram.

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

Northeastern University
NULab for Texts, Maps, and Networks

Exploratory Tool: Word Trees

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

Word Trees

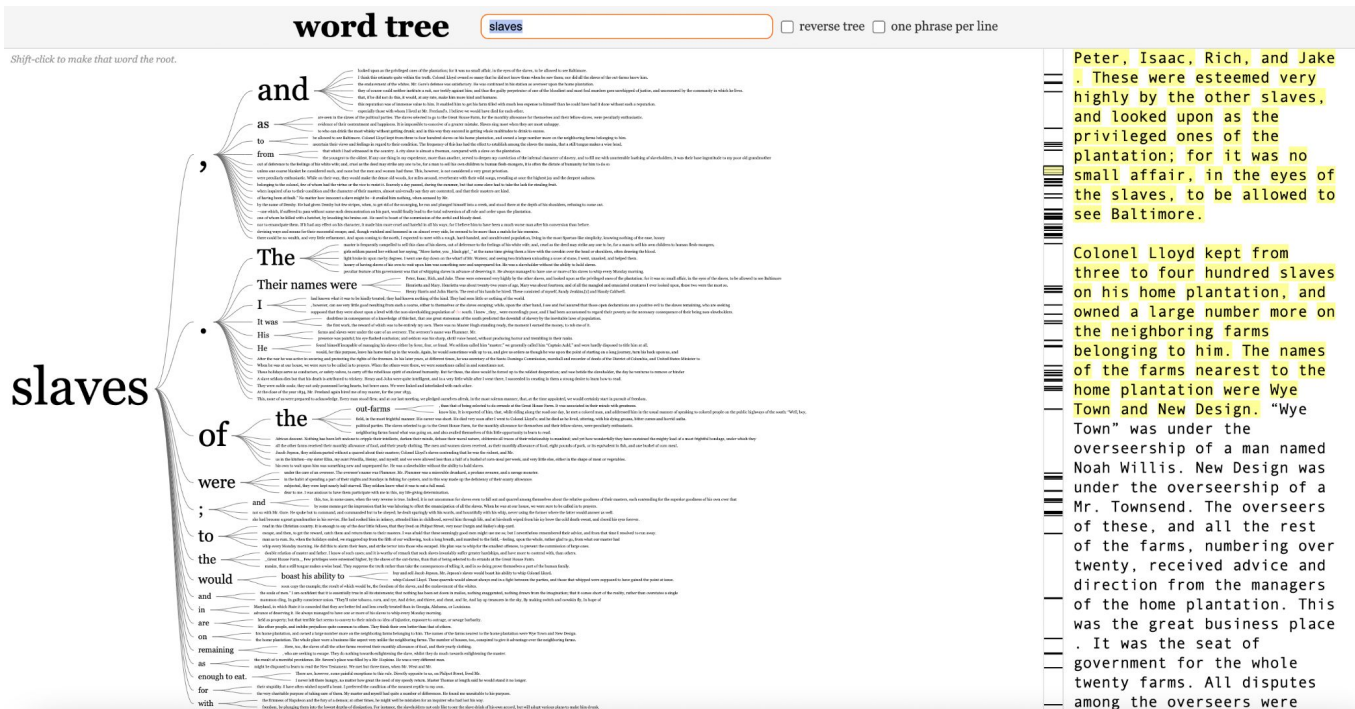- https://www.jasondavies.com/wordtree/
- A word tree **depicts multiple parallel sequences of words**
- This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest.
- There are some restrictions in size: fewer than 1 million words should work

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

Word Tree Example

Reflects descriptive language on the daily lives of enslaved people.

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 your search term. To do this click "reverse tree" next to the search bar.

Here, 'the', 'his', 'my fellow-', and 'their' are the dominant words preceding the word 'slaves'.

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 Trees!**

Discussion Prompts

- What limitations are you observing? What functionalities do you wish these tools might offer?
- Even with these limitations, how can you apply these simple tools in your research and exploration?

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

Voyant

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

Voyant

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

https://voyant-tools.org/

VOYANT
see through your text

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

Alternatively, insert URLs or full text into the textbox.

Add Texts   🗺 ◑ ?

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

📂 Open | Upload

✓ Reveal

Click here for help and advanced options

Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant: Single text Dashboard

Results:

From Benjamin Franklin's autobiography you can see the default results page with multiple panes:

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

These boxes can all be changed!
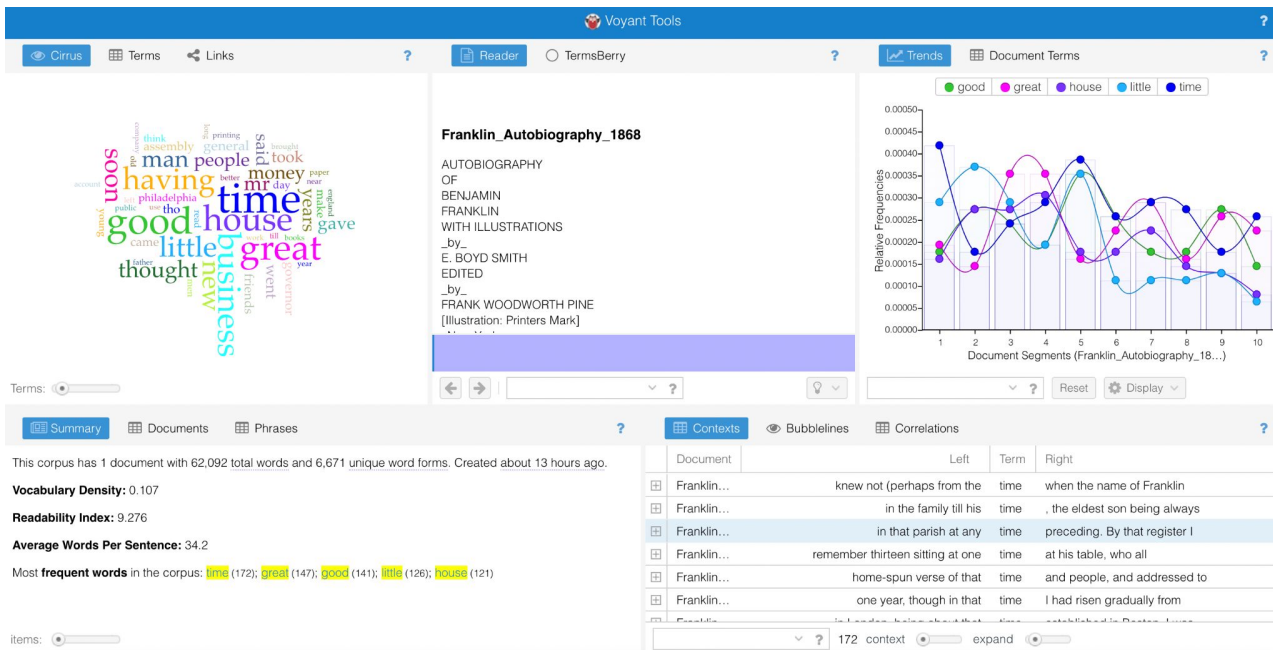
Benjamin Franklin's Autobiography

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 "time" appears in the text and the contexts in which it appears.

Document	Left	Term	Right
Franklin…	knew not (perhaps from the	time	when the name of Franklin
Franklin…	in the family till his	time	, the eldest son being always
Franklin…	in that parish at any	time	preceding. By that register I
Franklin…	remember thirteen sitting at one	time	at his table, who all
Franklin…	home-spun verse of that	time	and people, and addressed to
Franklin…	one year, though in that	time	I had risen gradually from
Franklin…	in London, being about that	time	established in Boston, I was
Franklin…	to be with him some	time	on liking. But his expectations
Franklin…	often regretted that, at a	time	when I had such a
Franklin…	and I still think that	time	spent to great advantage. There
Franklin…	brother. I stood out some	time	, but at last was persuaded
Franklin…	last year. In a little	time	I made great proficiency in
Franklin…	or wanted. And after some	time	an ingenious tradesman, Mr. Matthew
Franklin…	one another again for some	time	, I sat down to put
Franklin…	endeavor at improvement. About this	time	I met with an odd
Franklin…	should have acquired before that	time	if I had gone on
Franklin…	into verse; and, after a	time	, when I had pretty well
Franklin…	think I might possibly in	time	come to be a tolerable

Bubblelines *Correlations* **Contexts**

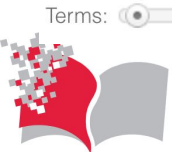*Feel free to ask questions at any point during the presentation!*

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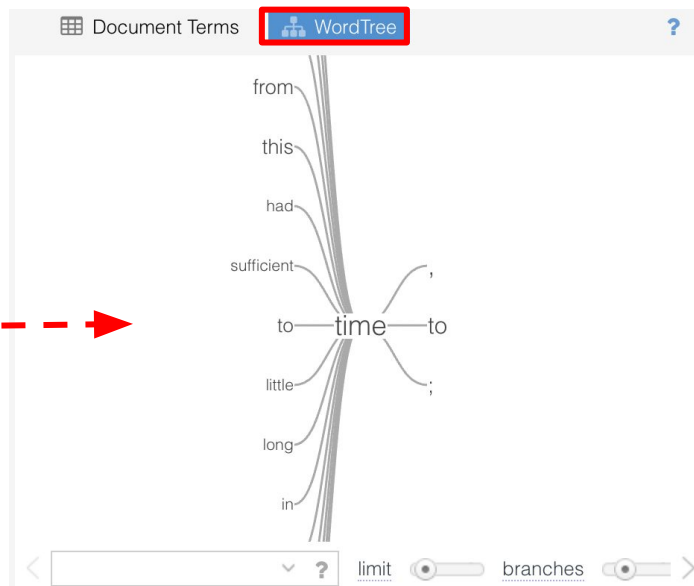
For our new pane option, we have chosen the WordTree visualization from the 'visualization tools' dropdown sub-menu. You can select the number of "branches" by dragging the scroll button at the bottom.

Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant: Corpus Dashboard

Results page of the corpus.
- A word cloud: combining all texts
- Reader section: scroll down all texts
- **Trends: relative frequency of terms across text —good for comparison**
- **Document Summary—good for comparison**
- Word Contexts: seperate for all texts

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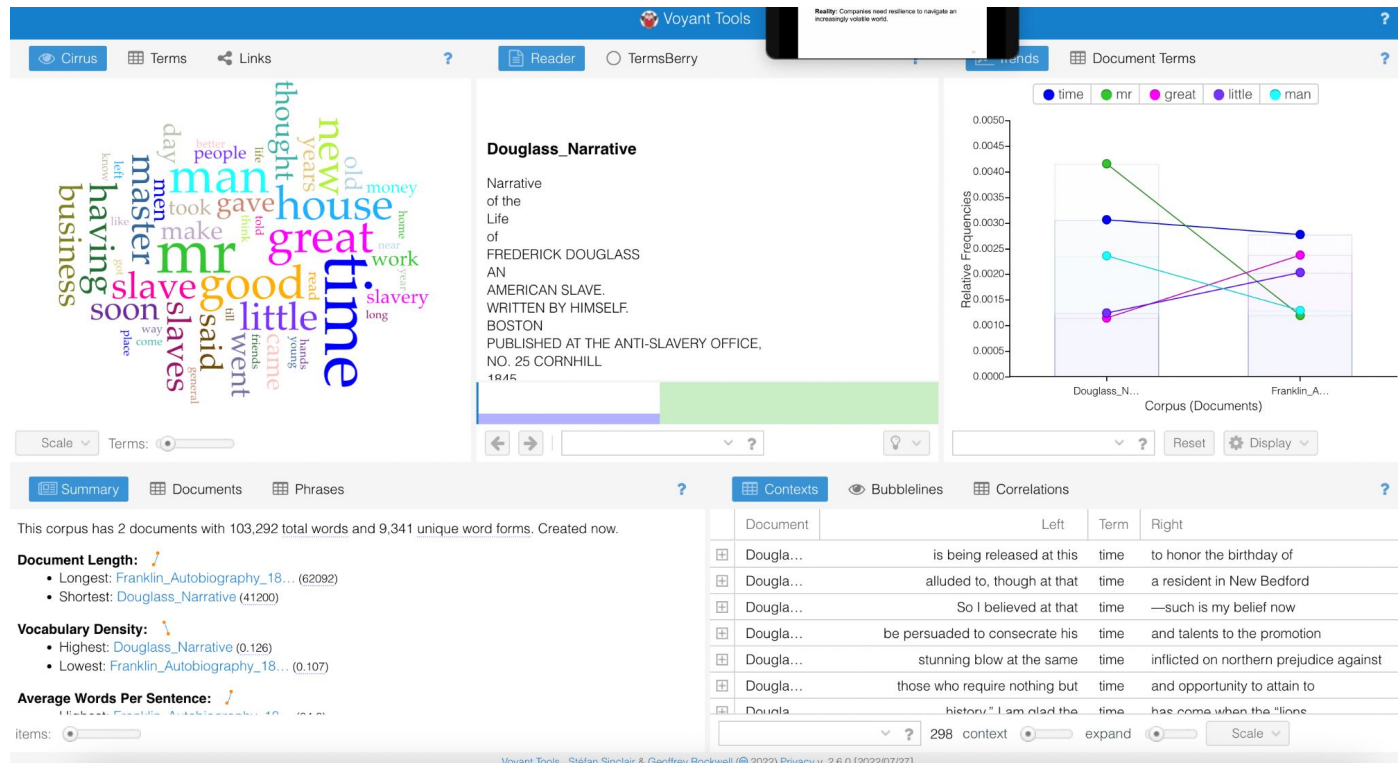
Your Turn!

Choose a sample text or texts and begin practicing web-browser text analysis. **Explore Voyant's features!**

Discussion Prompts

- What do you find challenging or exciting about this tool?
- What interesting or surprising results came up?
- How might you interpret those results based on what you know about your field?

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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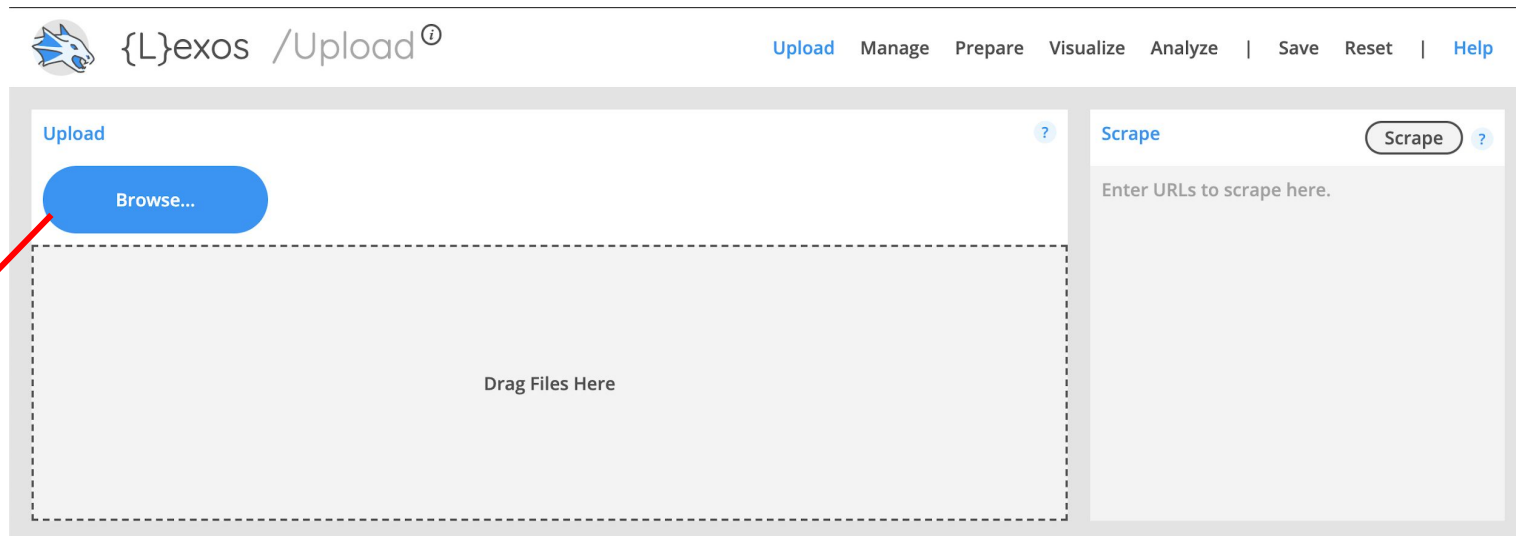
Lexos: Upload

Click Browse and select your entire text (or drag file into the "Drag Files Here" area)

You will not get a super visible notification when the upload is done—click "Manage" to double check that the text file is there.

{L}exos /Upload ⓘ

Upload   Manage   Prepare   Visualize   Analyze | Save   Reset | Help

Upload   ?

Browse...

Drag Files Here

Scrape   Scrape   ?

Enter URLs to scrape here.

Northeastern University
NULab for Texts, Maps, and Networks

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 *i*

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

Active	#	Document	Class	Source	Excerpt	
●	1	Jacobs_Narrative		Jacobs_Narrative.txt	incidents life slave girl written linda brent northerners know nothing slavery think perpetual bondage conception depth... ...even years concealment subsequent escape north resident boston living witness truth interesting narrative george w lowther	Download ?
●	2	Douglass_Narrative		Douglass_Narrative.txt	narrative life frederick douglass american slave written boston published antislavery office cornhill entered accord... ...ice success humble effortsand solemnly pledging self anew sacred causei subscribe frederick douglass lynn mass april end	

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!
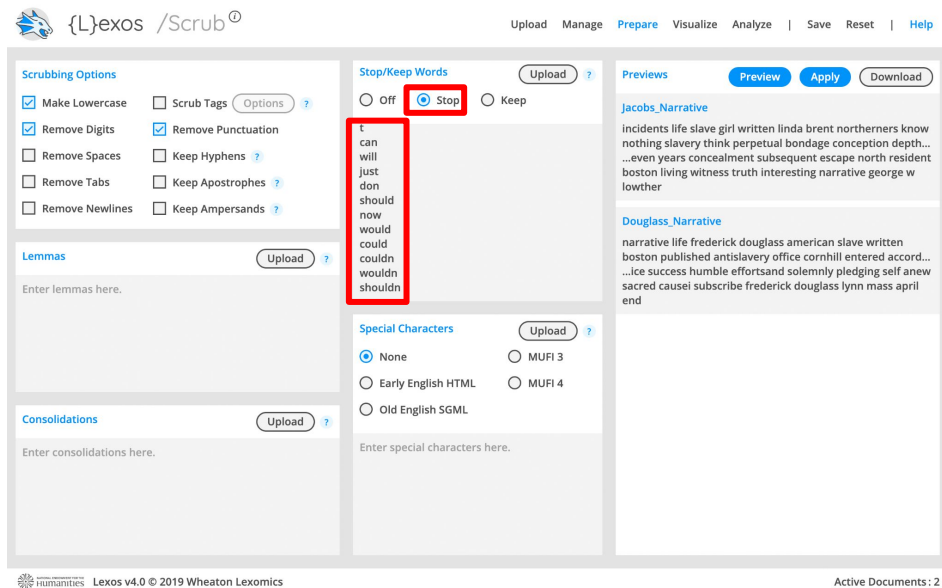
Lexos: Prepare (scrub)

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

- **Make Lowercase**: make all your letters lowercase. Even though you know "A" and "a" are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation**: remove punctuation, which may influence your results.
- **Stop/Keep Words**: remove a list of words. Usually these would be **stopwords**, or the most common words in a language (English: the, a she, her, it, him, they, etc).
- **Lemmas**: standardize to the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to "talk"

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

Lexos: Removing Stopwords

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

Lexos: Applying your Preparations

BEFORE PREP

Previews   Preview   Apply   Download

Jacobs_Narrative

Incidents in the Life of a Slave Girl. Written by Herself. Linda Brent "Northerners know nothing at all about Slavery.... ...orth. I am now a resident of Boston, and am a living witness to the truth of this interesting narrative. George W. Lowther.

Douglass_Narrative

Narrative of the Life of FREDERICK DOUGLASS AN AMERICAN SLAVE. WRITTEN BY HIMSELF. BOSTON PUBLISHED AT THE ANTI-SLAVERY... ...pledging my self anew to the sacred cause,—I subscribe myself, FREDERICK DOUGLASS. LYNN, _Mass., April_ 28, 1845. THE END

AFTER PREP

Previews   Preview   Apply   Download
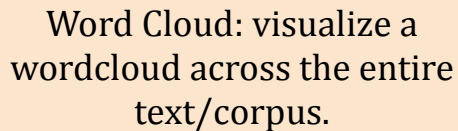
Jacobs_Narrative

incidents life slave girl written linda brent northerners know nothing slavery think perpetual bondage conception depth... ...even years concealment subsequent escape north resident boston living witness truth interesting narrative george w lowther

Douglass_Narrative

narrative life frederick douglass american slave written boston published antislavery office cornhill entered accord... ...ice success humble effortsand solemnly pledging self anew sacred causei subscribe frederick douglass lynn mass april end

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: Visualize

Word Cloud

Font: Open Sans   Term Count: 100   Color: Default   Generate   PNG   SVG

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

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

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

Northeastern University
NULab for Texts, Maps, and Networks

Lexos: Visualize > Multicloud

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** <u>document</u>. To use a rolling window:

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"

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

Lexos: Rolling Window Results

Using *Narrative of the Life of Frederick Douglass, an American Slave* and searching for the words 'slave', 'master' and 'slavery' with a window of 1000 words, we can get an idea of how these terms work together in the text.

Feel free to ask questions at any point during the presentation!
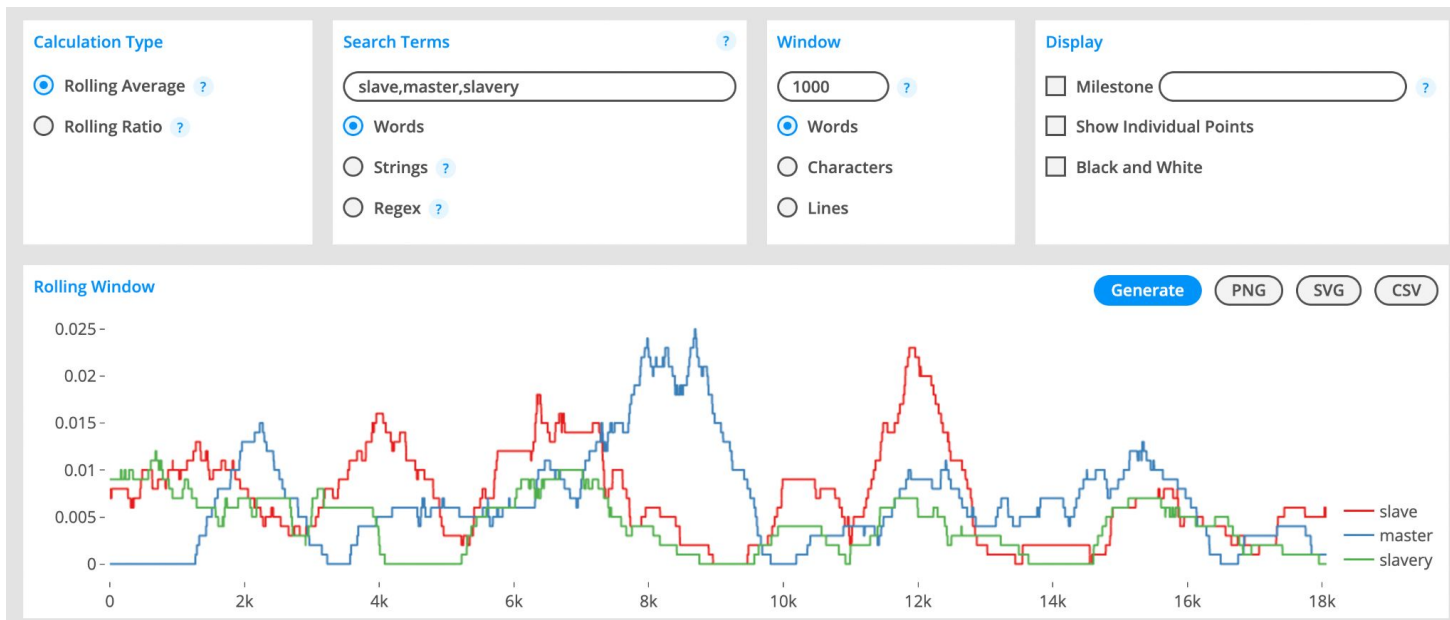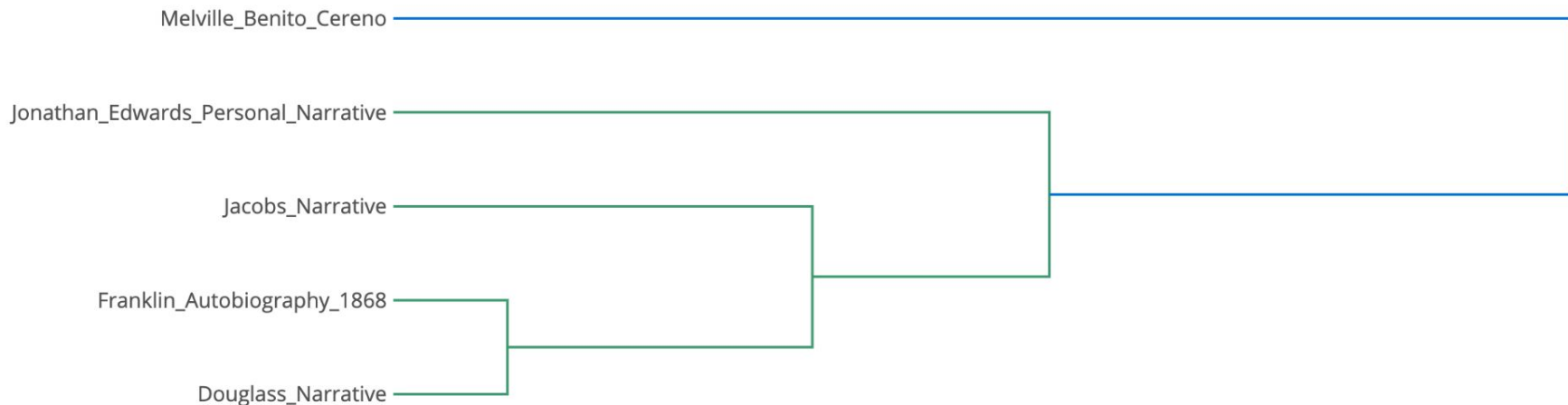
Lexos: Analyze > Dendrogram

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: Analyze > Dendrogram

This dendrogram shows that the only fictional text in the corpus, *Benito Cereno*, is identified as distinct from the rest. For the nonfiction texts, there are more similarities between the longer ones, while the briefer "Personal Narrative" of Edwards is distinctive. The two most similar texts are those by Franklin and Douglass. Note that the settings you choose will impact your results, so you should try different options!

Melville_Benito_Cereno

Jonathan_Edwards_Personal_Narrative

Jacobs_Narrative

Franklin_Autobiography_1868

Douglass_Narrative

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.

Northeastern University
NULab for Texts, Maps, and Networks

Your Turn!

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

Discussion Prompts

- What difference did you notice between Voyant and Lexos?
- Which tool do you prefer and why?
- How would you want to use these tools in this class and future?

Northeastern University
NULab for Texts, Maps, and Networks

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

Thank you!

If you have any questions, contact us at nulab.info@gmail.com

- Slides, handouts, and data available at:

https://bit.ly/fa22-loeffelholz-textanalysis

- You also have access to DITI Canvas Module on Computational Text Analysis.
- We'd love your feedback! Please fill out a short survey here: https://bit.ly/diti-feedback

Schedule an appointment with us! **https://calendly.com/diti-nu**

Northeastern University
NULab for Texts, Maps, and Networks

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