

# Computational Text Analysis for Content Analysis

---

Taught by: Emily Sullivan & Hunter Moskowitz  
ENGW 1410: Intro to Writing Studies Research

Dr. Neal Lerner  
Spring 2024



# 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/sp24-lerner-engw1410-text-analysis>



# What is Computational Text Analysis?



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 are 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](#).
- 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.



# 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.



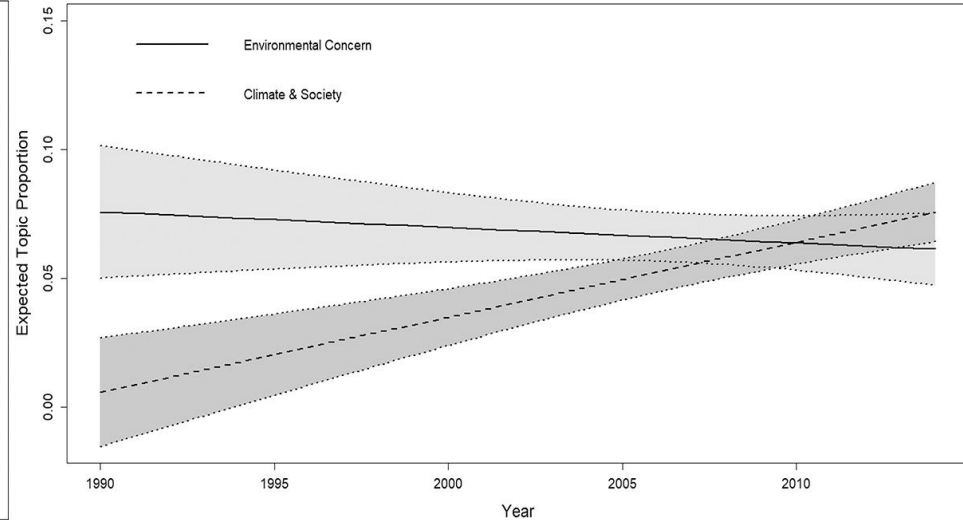
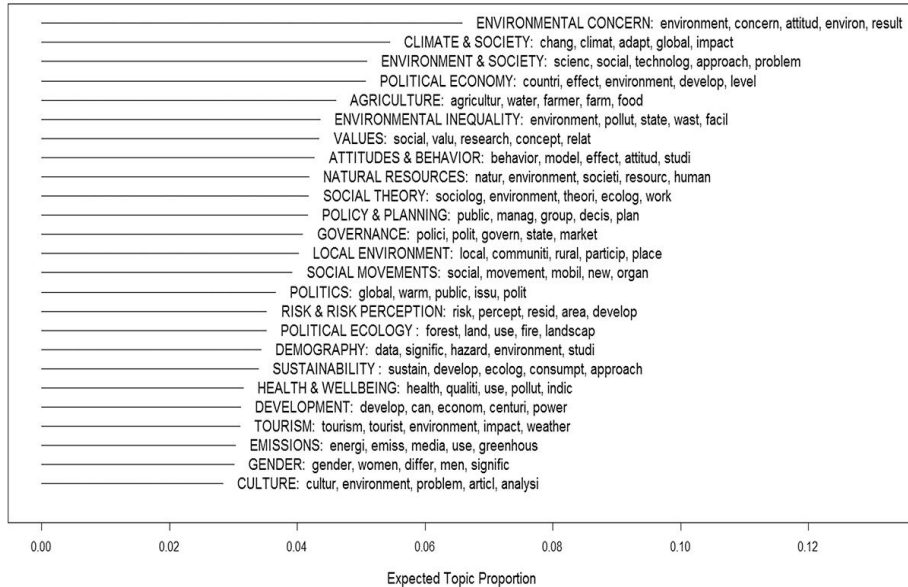
# Examples from Practice



Northeastern University  
*NULab for Texts, Maps, and Networks*

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

# Key Topics in Environmental Sociology



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!*



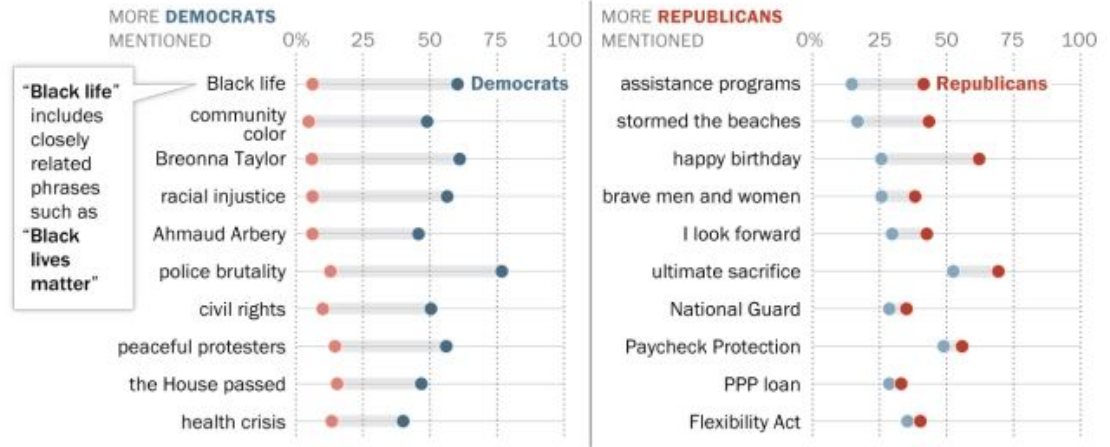
Content warning: police violence, racism

## 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



# 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 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?



# 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

Filippo Tommaso Marinetti, “[The Joy of Mechanical Force](#)”

McKenzie Wark, “[A Hacker Manifesto](#)”



# Sample Corpus

The following .txt files are available on:

<https://bit.ly/sp24-lerner-engw1410-text-analysis>

- For each file, click “Raw” in the top right corner.
- Right-click (PC) or Ctrl-click (Macs) on the text and choose “Save As.”
- Save as a .txt file on your computer.



# 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
- Default is lowercase all words and apply stopwords
- It can be run with and without stopwords





# 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 Counter Examples

## TOP WORDS ⬇

Word	Frequency
class	98
information	82
property	59
production	40
form	39
politics	39
hacker	37
new	32
hack	31
free	31

Shows the top words in the text.

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

## BIGRAMS ⬇

bigram <sup>2</sup>	Frequency
of the	81
the hacker	33
to the	31
as a	31
is the	30
of information	25
in the	24
hacker class	22
the hack	20
of a	19

## TRIGRAMS ⬇

trigram <sup>2</sup>	Frequency
the hacker class	22
the vectoralist class	13
the production of	11
of the hack	10
the possibility of	9
form of property	9
as a class	8
the means of	8
the form of	8
the politics of	8

The top two trigrams 'the hacker class' and 'the vectoralist class' both contain the words 'the' and 'class'. 'The' is a stopwords, and 'class' is the dominant word in this text!



# Exploratory Tool: Word Trees



Northeastern University  
*NULab for Texts, Maps, and Networks*

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

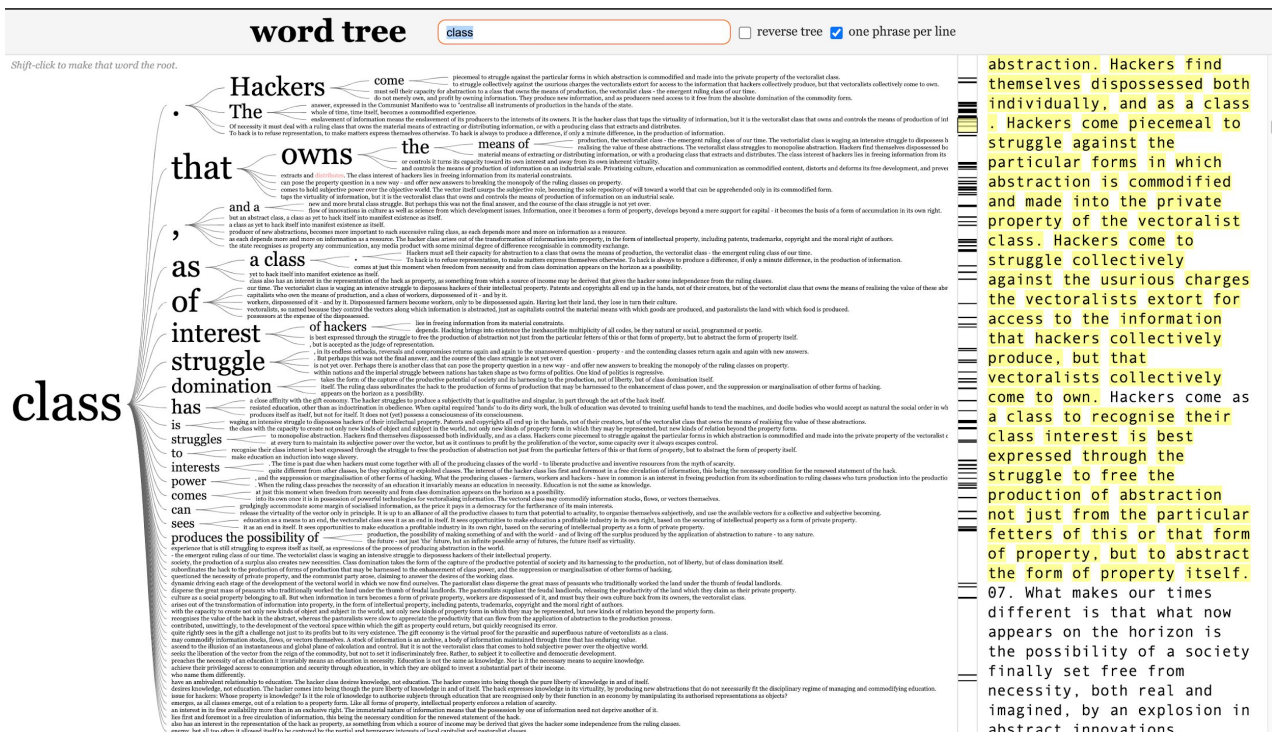
# Word Trees

- <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



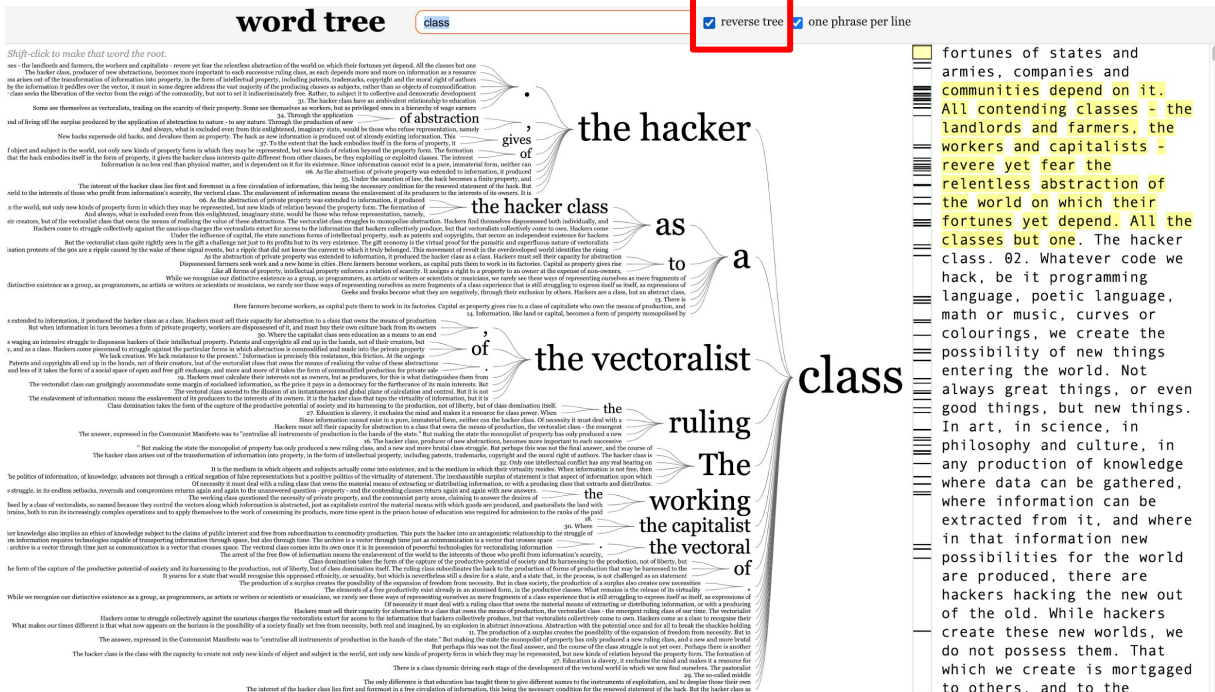
Notice the word  
'class' seems to be  
often at the end of a  
sentence, followed by  
period.

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



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 the hacker, the vectoralist, ruling, working etc. are the dominant words preceding the word 'class.



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

# 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 “class” appears in the text and the contexts in which it appears.

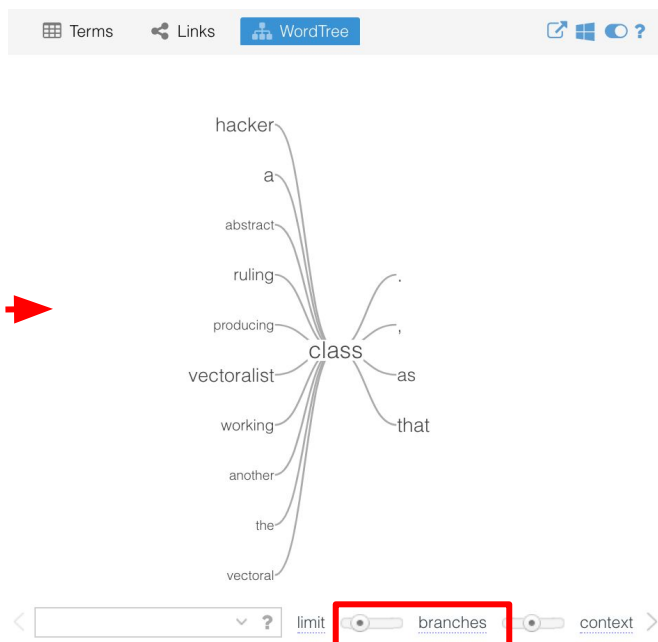
Contexts				
Bubblelines				
Correlations				
Document	Left	Term	Right	
1) Hacker_Man...	classes but one. The hacker	class	. 02. Whatever code we hack	
1) Hacker_Man...	as mere fragments of a	class	experience that is still struggling	
1) Hacker_Man...	by others. Hackers are a	class	, but an abstract class, a	
1) Hacker_Man...	a class, but an abstract	class	, a class as yet to	
1) Hacker_Man...	but an abstract class, a	class	as yet to hack itself	
1) Hacker_Man...	form, neither can the hacker	class	. Of necessity it must deal	
1) Hacker_Man...	must deal with a ruling	class	that owns the material means	
1) Hacker_Man...	information, or with a producing	class	that extracts and distributes. The	
1) Hacker_Man...	that extracts and distributes. The	class	interest of hackers lies in	

98 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.



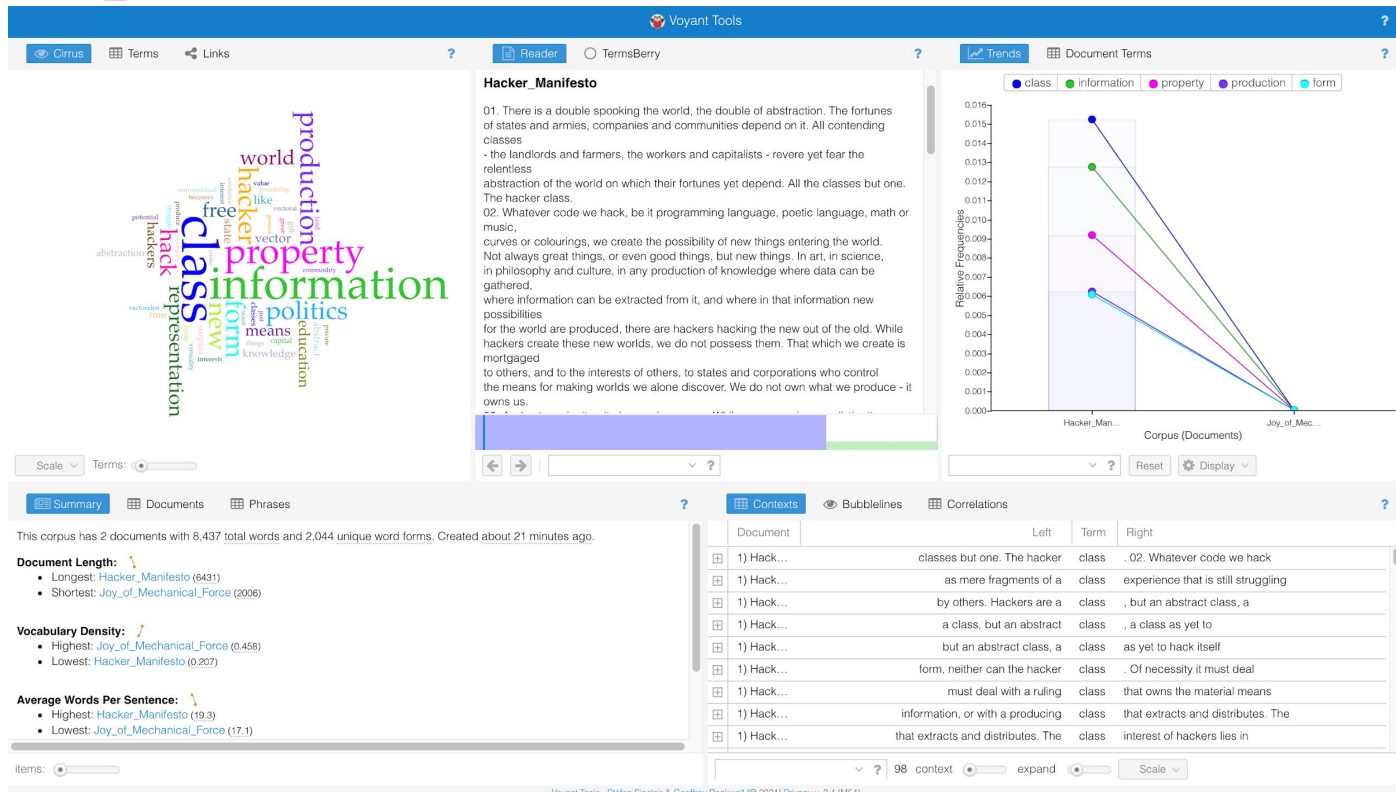
Northeastern University  
NULab for Texts, Maps, and Networks

*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: separate for all texts



# Your Turn!

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

## Discussion Prompts

- What 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?



# Thank you!

If you have any questions, contact us at [nulab.info@gmail.com](mailto:nulab.info@gmail.com)

Developed by Vaishali Kushwaha, Garrett Morrow, Emily Sullivan, & Hunter Moskowitz

- Slides, handouts, and data available at:  
<https://bit.ly/sp24-lerner-engw1410-text-analysis>
- If you have any questions, contact us at [nulab.info@gmail.com](mailto:nulab.info@gmail.com)
- Have questions? Schedule an appointment with us [here](#).

