# Computational Text Analysis for Content Analysis

By Vaishali Kushwaha and Adam Tomasi
Digital Integration Teaching Initiative (DITI)

For POLS 7346 Resilient Cities
Daniel Aldrich
Spring 2021

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 **http://bit.ly/diti-spring2021-aldrich**



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

# Workshop Objectives

- Understand best practices for collecting and storing textual data when performing basic computational text analysis
- Understand how web-based computational text analysis programs work, such as in their behind-the-scenes data preparation
- Understand how to interpret the results from your text analysis explorations



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

# Workshop Outline

- Introduction
- Examples from Practice
- Text Preparation
- Word Counter
  - Demo
- Word Trees
  - Demo
- Voyant
  - Demo
- Lexos
  - Demo
- Conclusion



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

# Introduction

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 making inferences based on textual data. Computational text analysis refers to an array of methods that can be used to "read" texts with a computer. It is similar to statistical analysis, but the data are texts.

- It involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- It includes methods such as word count frequency, nGrams, and sentiment analysis.



*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** and **discover patterns** in texts.

From collections of texts, researchers can discover keywords that serve as a proxy for major trends in societies, cultures, and policies. Computational tools can reveal patterns on how public officials communicate policies, which issues are of concern, which phrases do leaders regularly associate with, and much more. Researchers may find surprising results that they would not have discovered from close reading or traditional methods alone.



*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

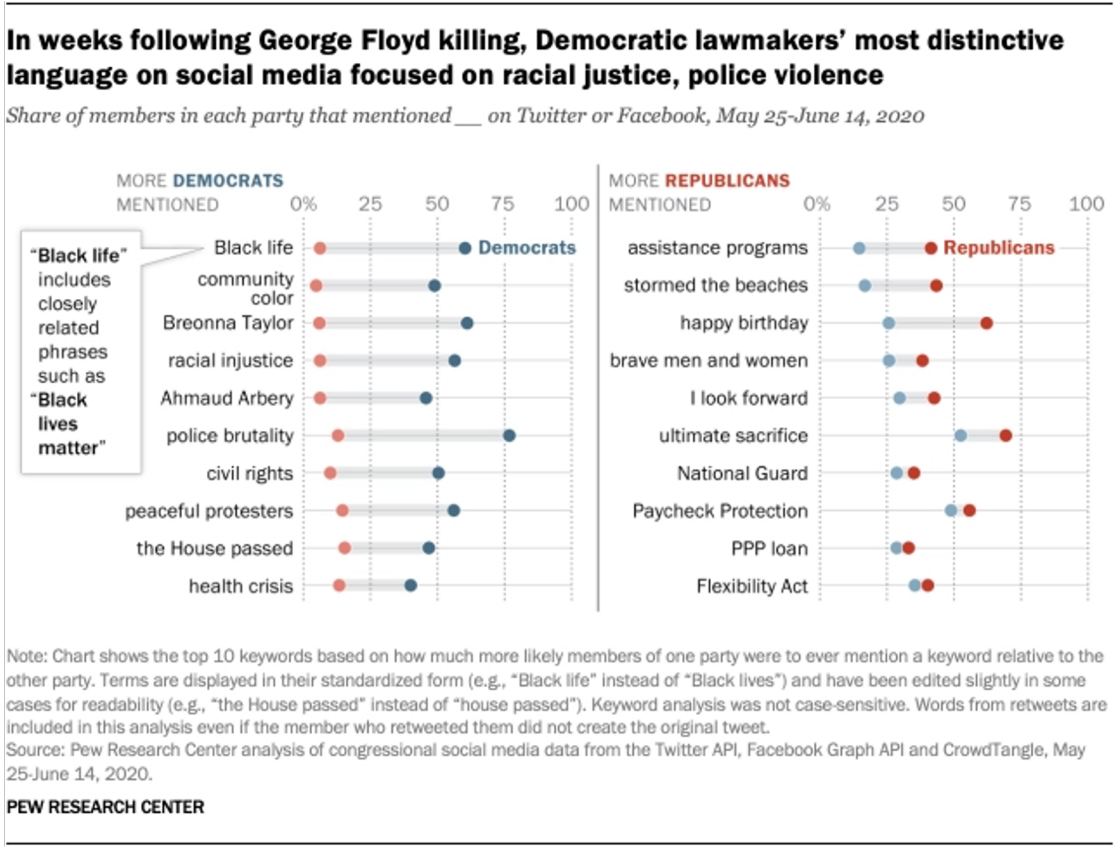
Northeastern University
*NULab for Texts, Maps, and Networks*

*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

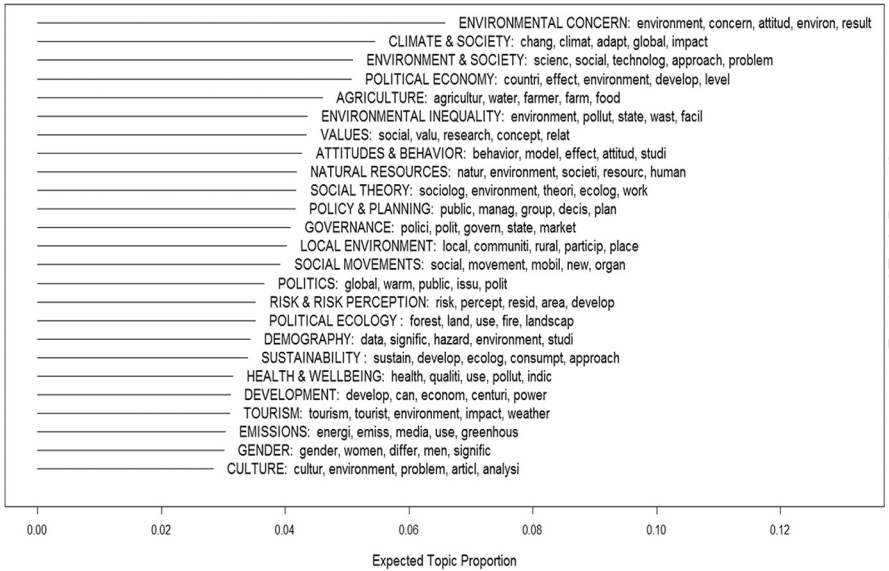
- Pew Research Center July 16, 2020 article
- Methodology



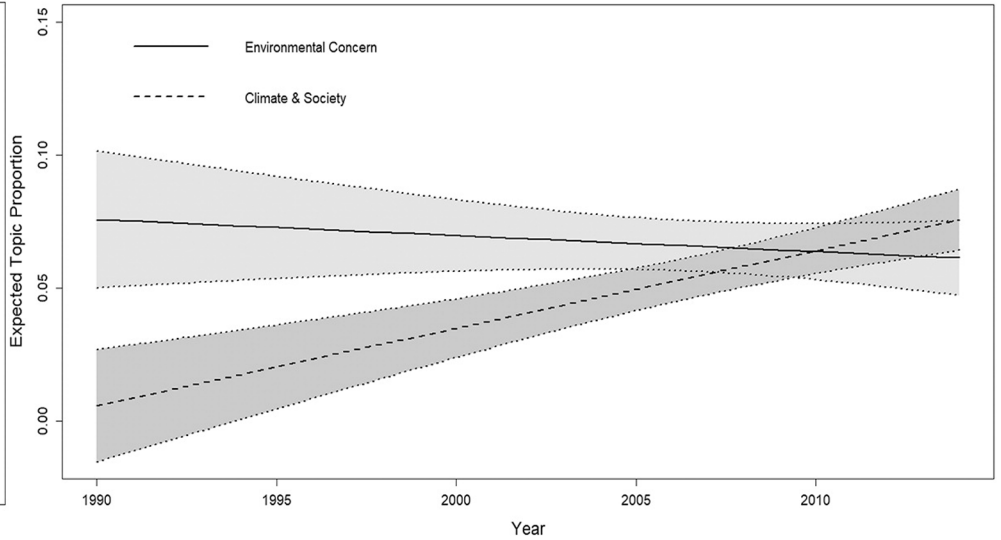



*Feel free to ask questions at any point during the presentation!*

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




25 topics ranked from most to least prevalent in the corpus of 815 environmental sociology articles, including the top five associated word stems. The *x*-axis represents the proportion of each topic within the overall corpus.




Topical prevalence of 'Environmental Concern' and 'Climate & Society' over time, 1990 – 2014 (with 95% confidence intervals).

Jeremiah Bohr & Riley E. Dunlap (2018) Key Topics in environmental sociology, 1990–2014: results from a computational text analysis, Environmental Sociology, 4:2, 181-195, DOI: 10.1080/23251042.2017.1393863

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

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

# 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?



*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

Our text is plain text (.txt file) of *Climate Ready Boston*, the final report dated December 2016. The primary objective is to explore this text using web-based computational text analysis tools.

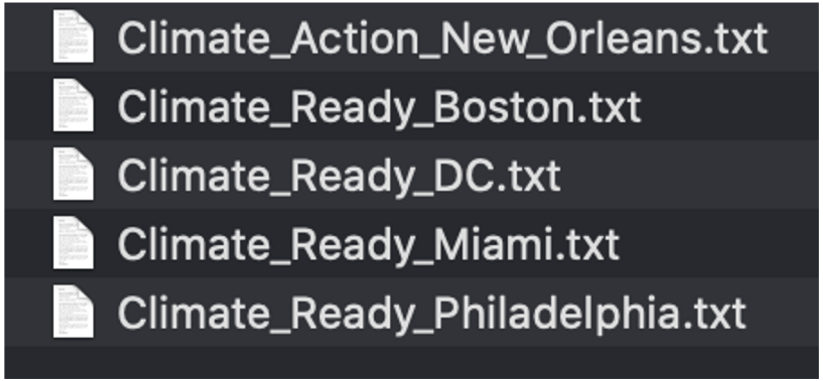
We will also use climate change plans of DC, Philadelphia, New Orleans, and Miami, to see how a corpus can be analyzed. The primary objective is to compare the climate change plans of these cities with Boston.



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

# Sample Corpus

The following .txt files are available on: **http://bit.ly/diti-spring2021-aldrich**






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

# 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

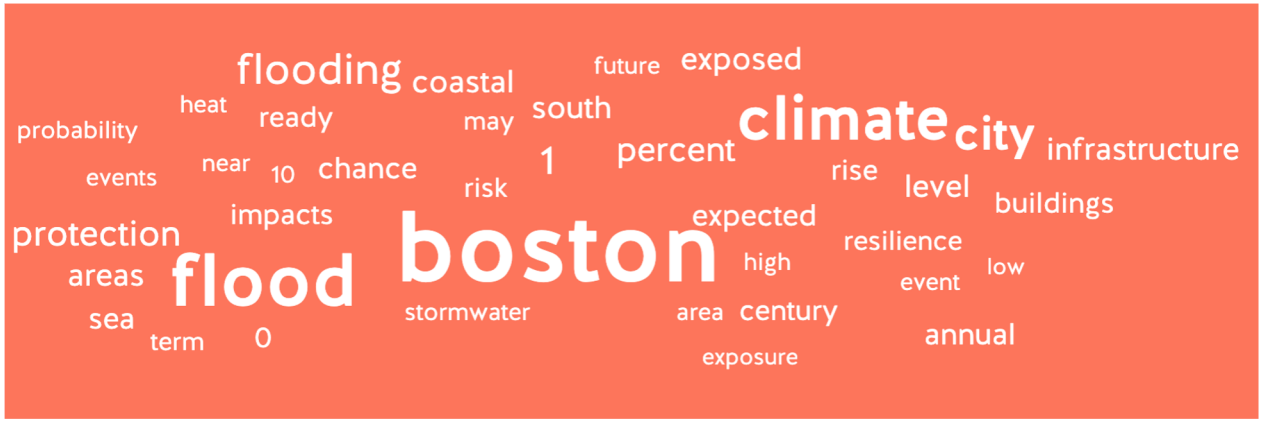
- https://databasic.io/en/wordcounter/
- A user-friendly **basic word counting tool**
- It allows you to count words, bigrams, and trigrams in plain text files and to download spreadsheets with your results
- The max file upload is 10MB
- Default is lowercase all words and apply stopwords
- It can be run with and without stopwords



*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.






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

# Word Counter Examples

**TOP WORDS**

| Word | Frequency |
|---|---|
| boston | 1680 |
| flood | 1203 |
| climate | 991 |
| city | 686 |
| flooding | 600 |
| 1 | 562 |
| protection | 475 |
| percent | 453 |
| exposed | 396 |

Shows the top words in the text.

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

**BIGRAMS**

| bigram | Frequency |
|---|---|
| in the | 685 |
| of the | 546 |
| flood protection | 400 |
| the city | 399 |
| annual chance | 327 |
| climate ready | 322 |
| sea level | 310 |
| exposed to | 305 |

**TRIGRAMS**

| trigram | Frequency |
|---|---|
| sea level rise | 294 |
| climate ready boston | 275 |
| percent annual chance | 233 |
| city of boston | 215 |
| boston climate ready | 193 |
| of boston climate | 189 |
| 1 percent annual | 177 |

It is interesting that though 'flood' is second most used term, it does not appear in top few trigrams. Also, '1' gets a meaningful context here!



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

# 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

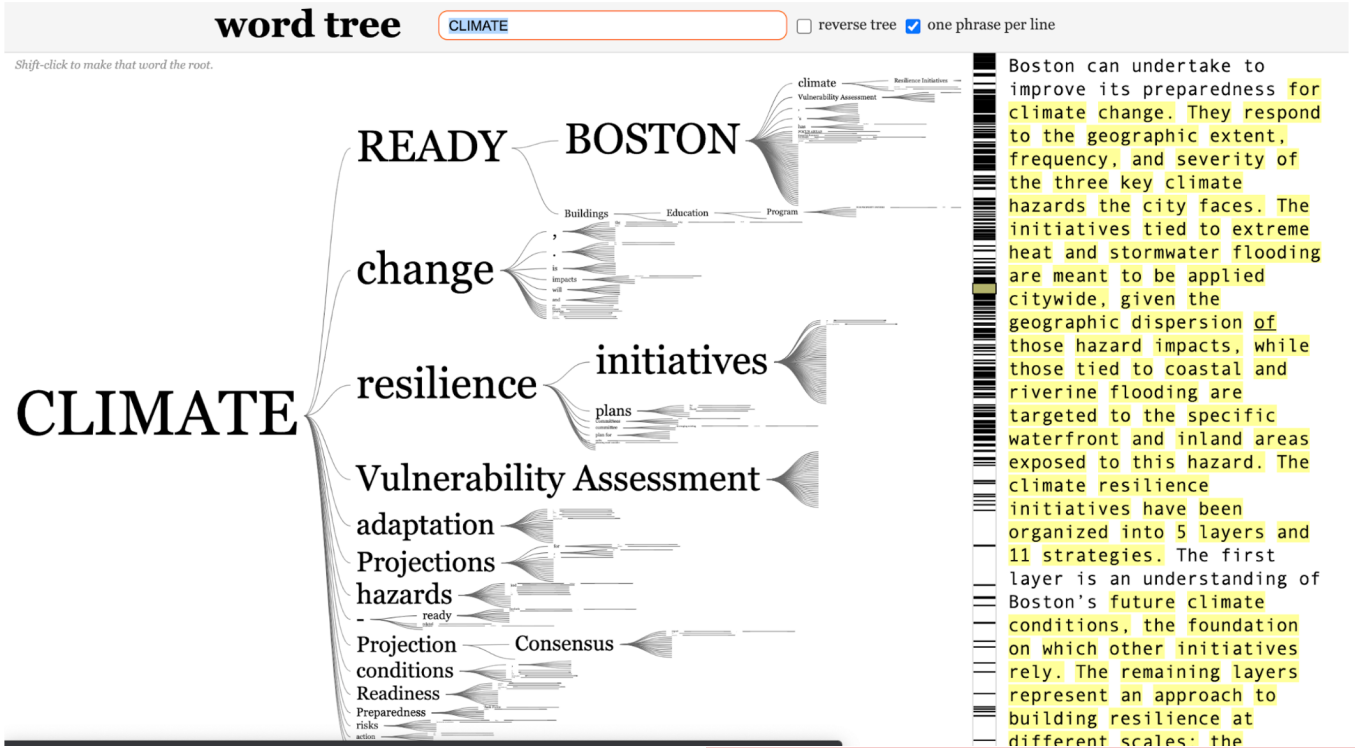
- https://www.jasondavies.com/wordtree/
- A word tree **depicts multiple parallel sequences of words**
- This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest.
- There are some restrictions in size: fewer than 1 million words should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



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

# Word Tree Example

Reflects the focus of the report on climate ready boston, climate change, resilience initiatives, vulnerability assessments, etc.

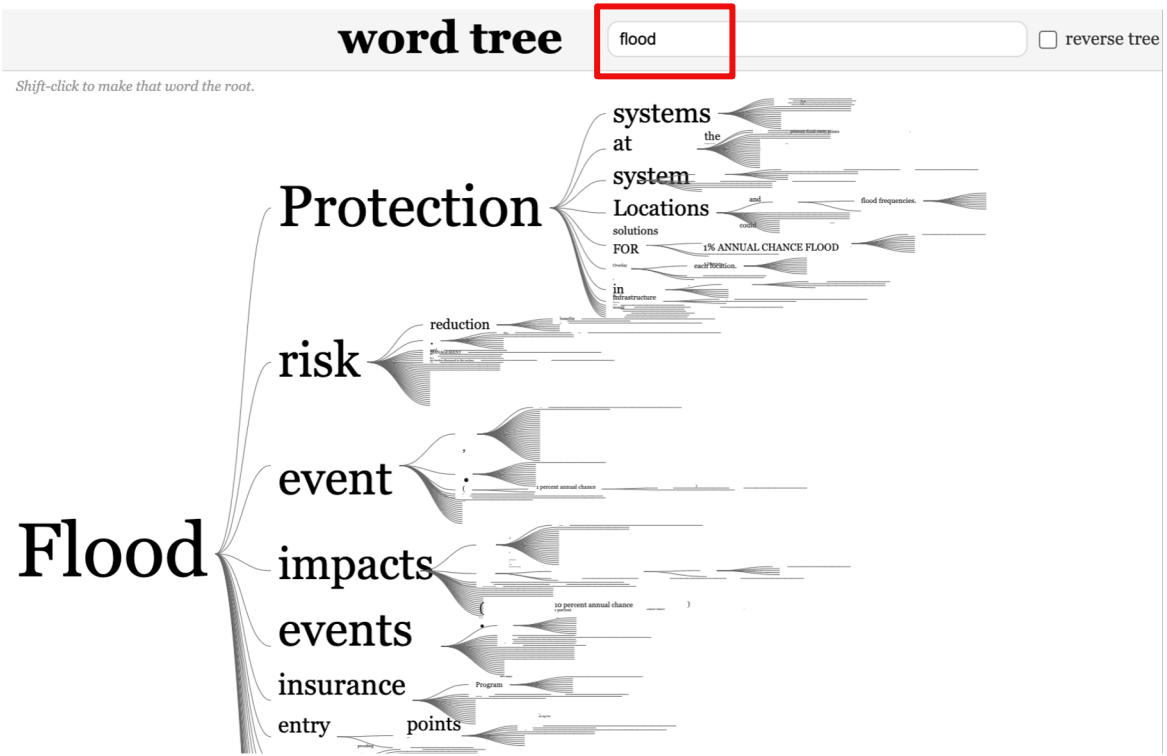





*Feel free to ask questions at any point during the presentation!*

# Word Tree Examples

'Flood' is the second most used word in the text. It is followed by various patterns: protection systems/system/locations, risk, event/s, impacts and insurance.



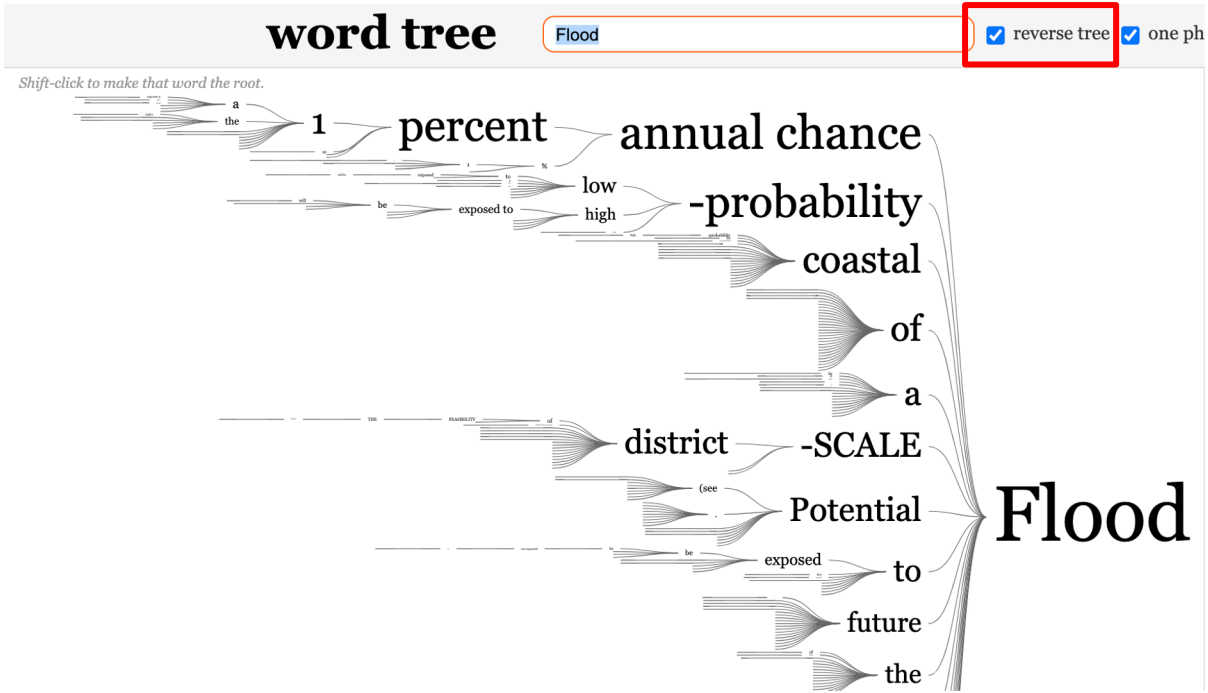



*Feel free to ask questions at any point during the presentation!*

# Word Tree: Reverse Trees

It is worth reversing the tree to see the words that often precede it. To do this click “reverse tree” next to the search bar.

Here annual chance, probability, coastal etc. are some of the terms preceding the word ‘flood’.






*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!

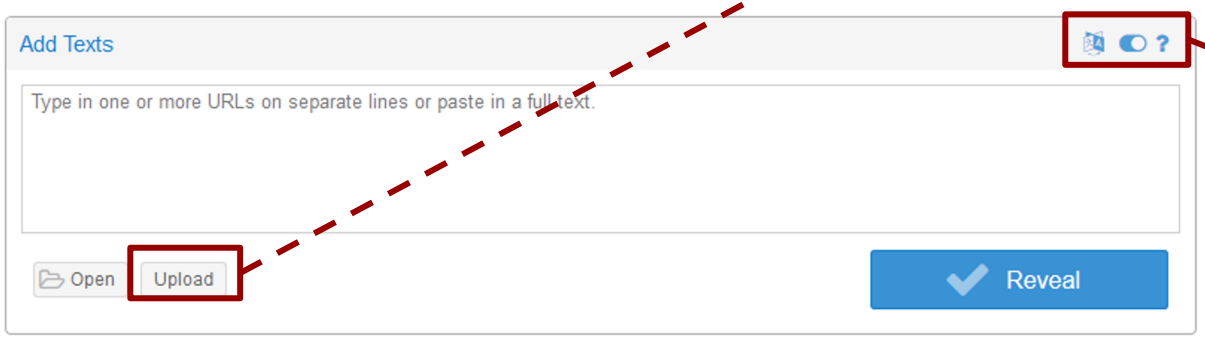
**https://voyant-tools.org/**



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

# 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.

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!*

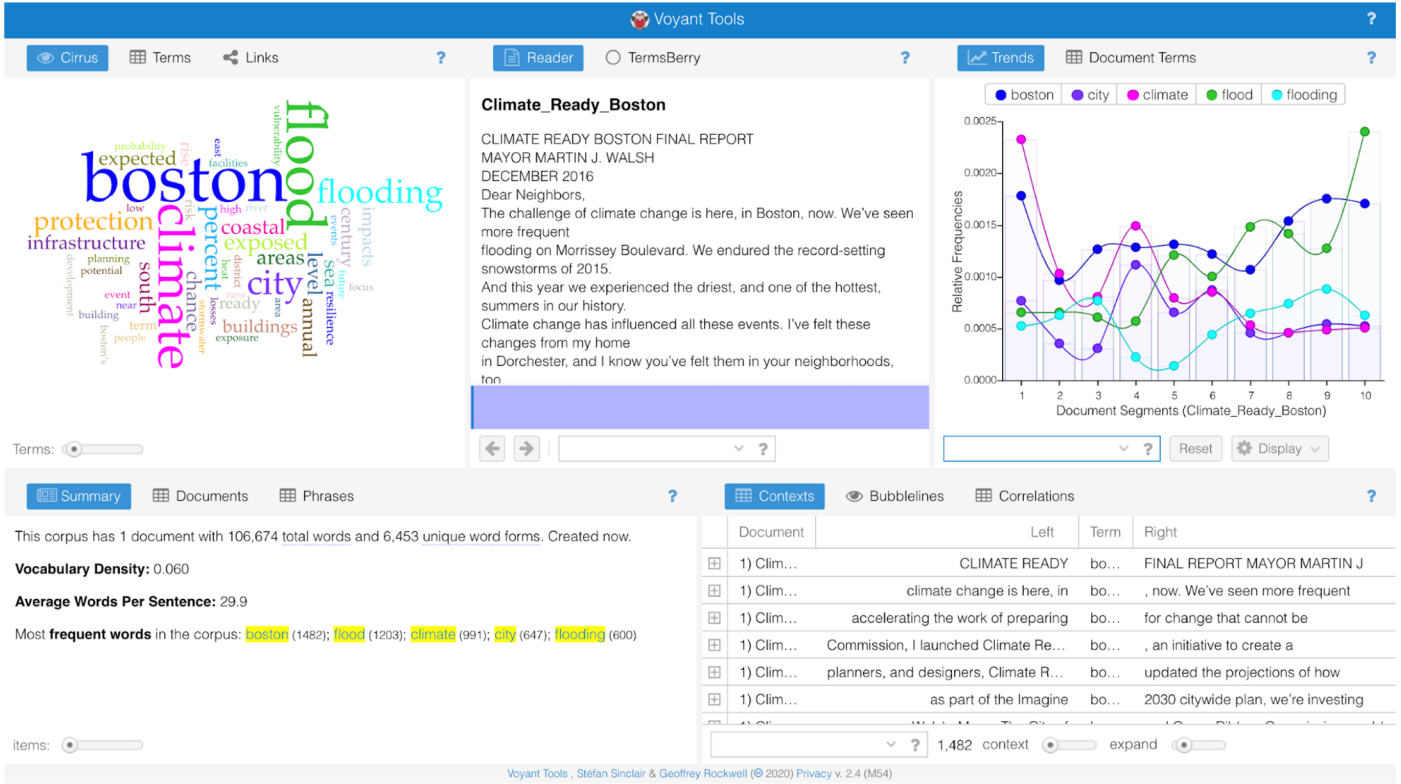
# Voyant: Understanding the Dashboard

Results:

From Climate Ready Boston 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!

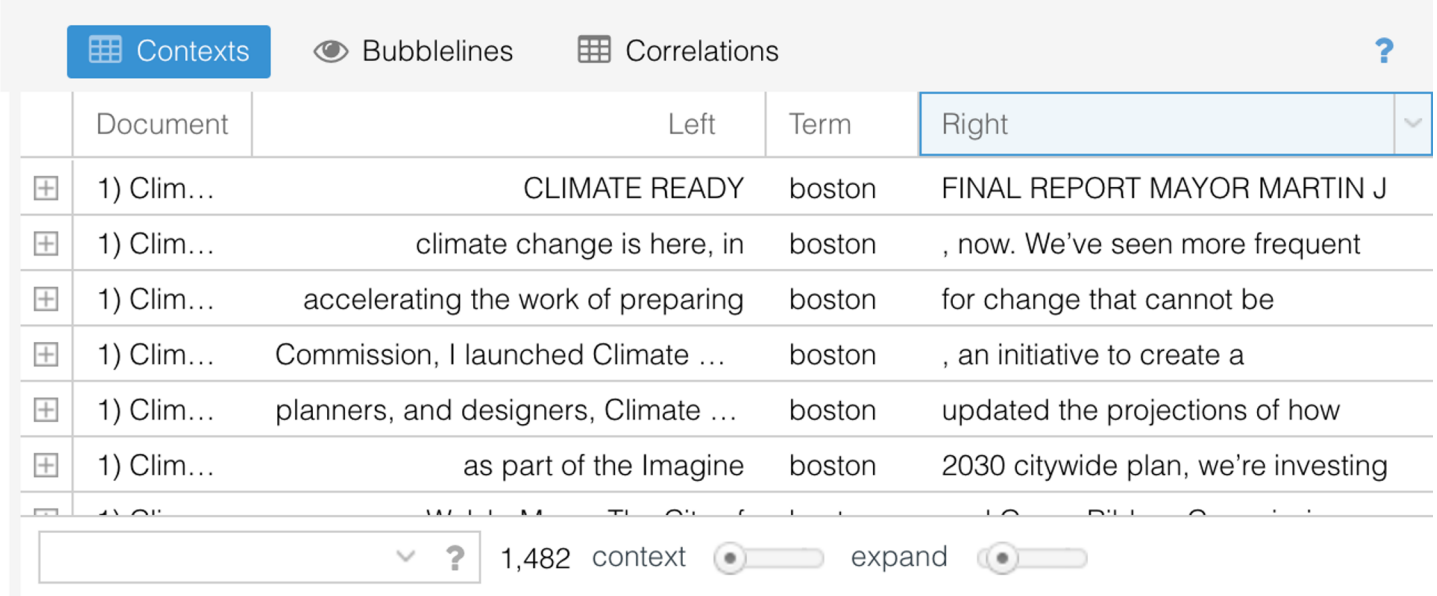





*Feel free to ask questions at any point during the presentation!*

# Voyant: Contexts (concordances)

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

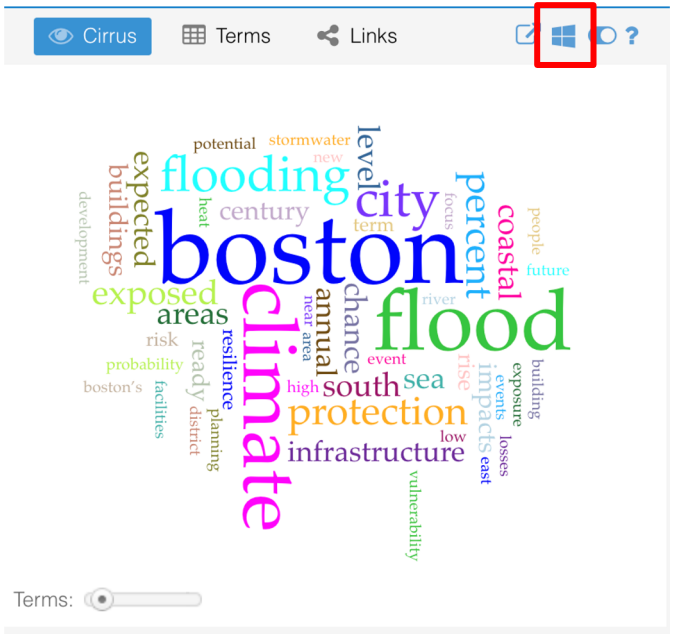


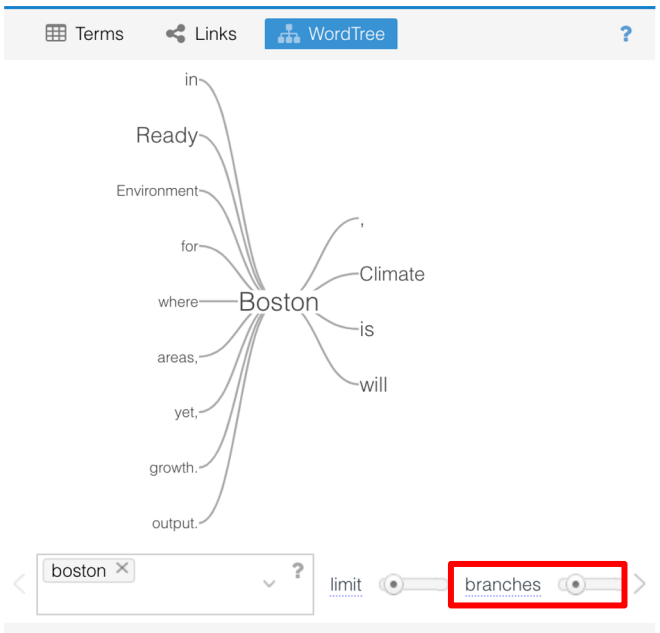



*Feel free to ask questions at any point during the presentation!*

# 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.

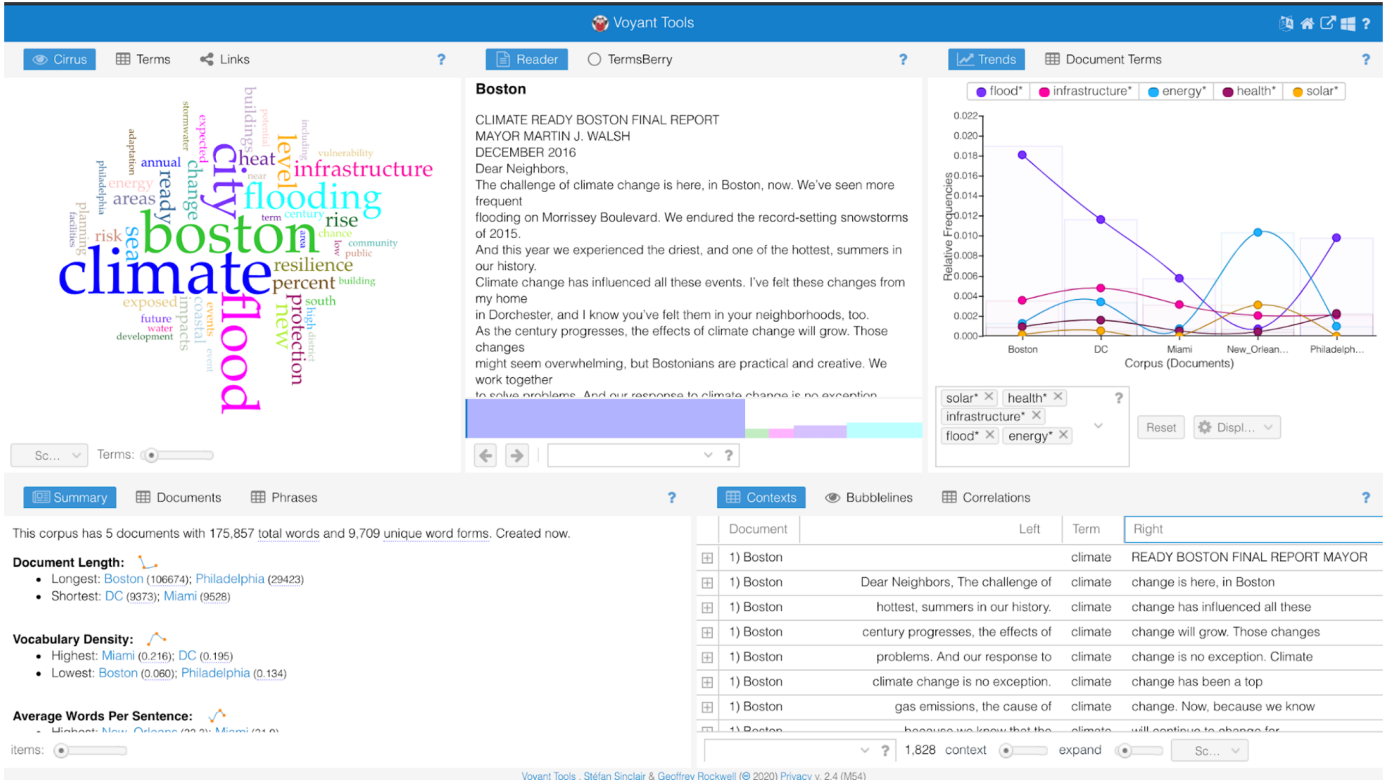


*Feel free to ask questions at any point during the presentation!*

# Voyant: Corpus Dashboard

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

- 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





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


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

# Lexos

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

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

# Lexos

Lexos provides a step-by-step guide for text uploading, preparation, and analysis.

- **Upload**: upload your .txt file
- **Manage**: select the files you want to prepare and analyze
- **Prepare**: prepare your text for analysis
- **Visualize**: create visualizations of patterns across your corpus or in single texts
- **Analyze**: analyze your text

**http://lexos.wheatoncollege.edu/upload**

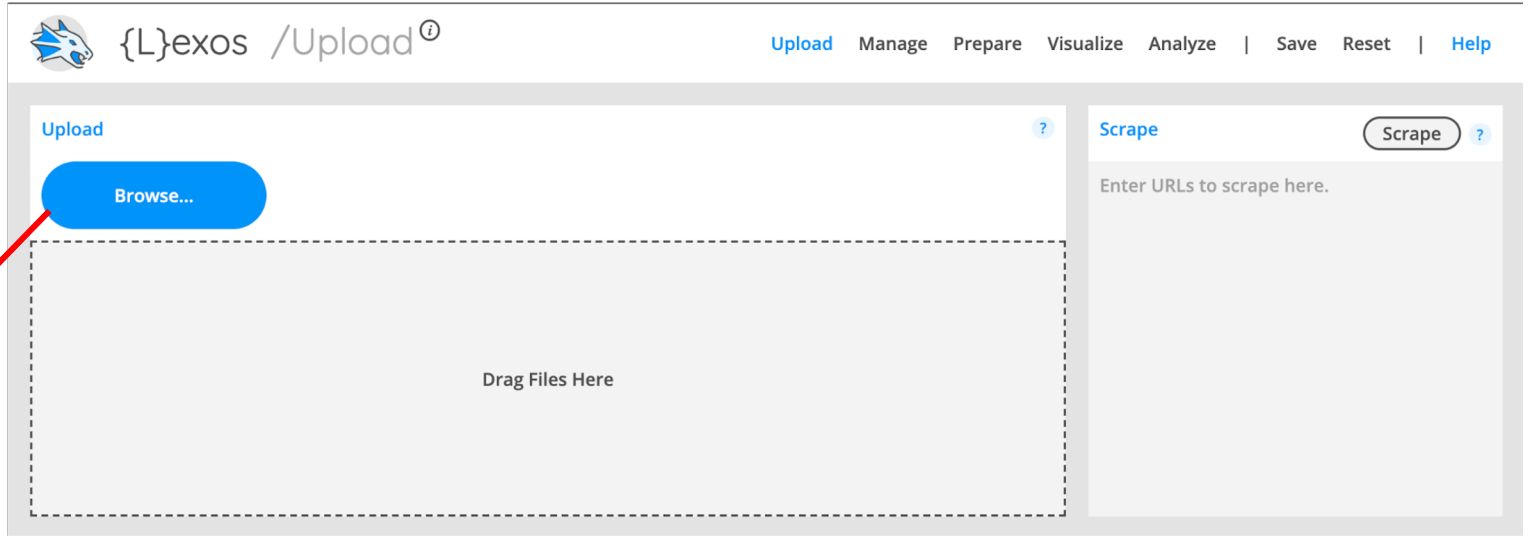


*Feel free to ask questions at any point during the presentation!*

# 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.

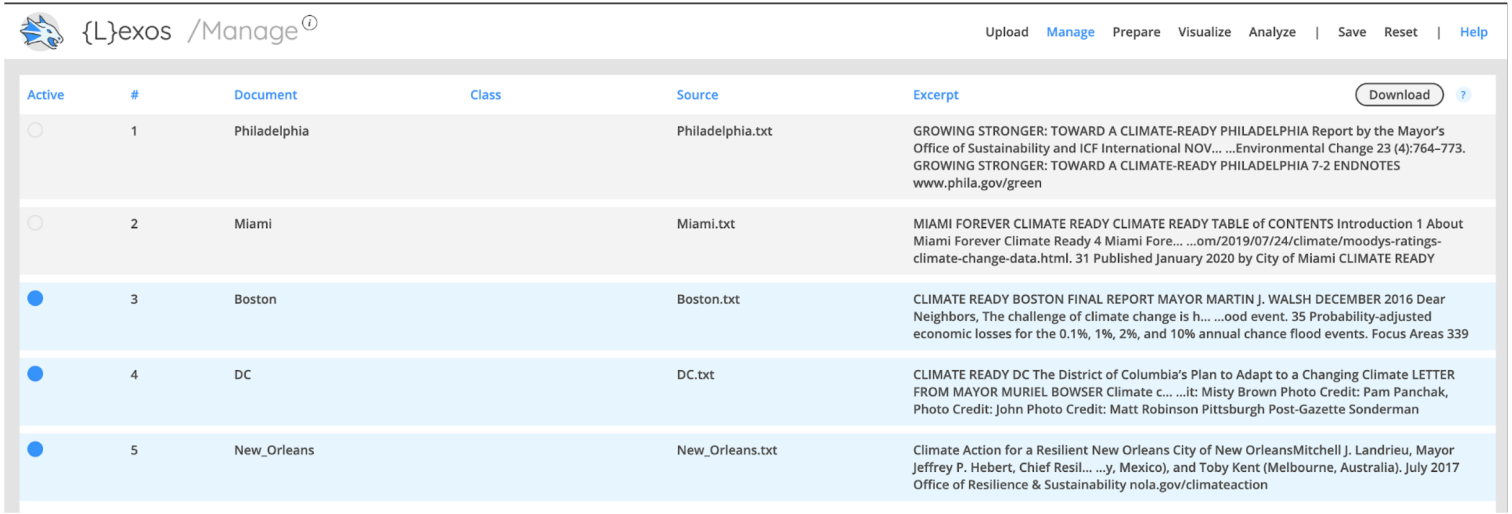





*Feel free to ask questions at any point during the presentation!*

# 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)

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:

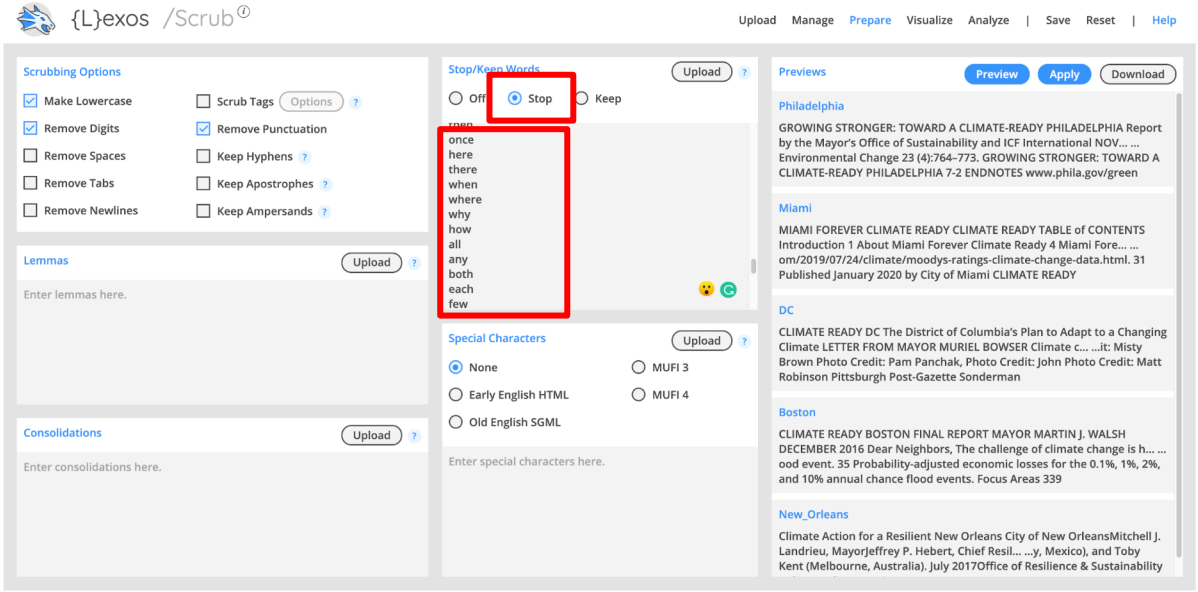
- **Make Lowercase**: make all your letters lowercase. Even though you know “A” and “a” are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation**: remove punctuation, which may influence your results.
- **Stop/Keep Words**: 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!*

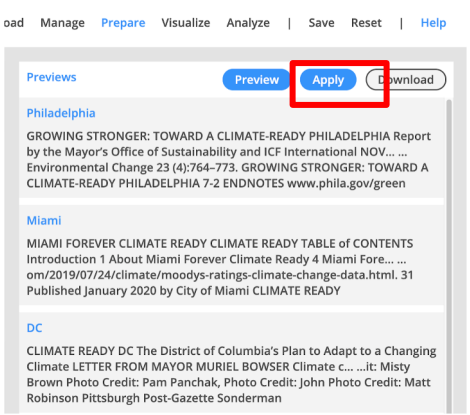
# Lexos: Removing Stopwords

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"

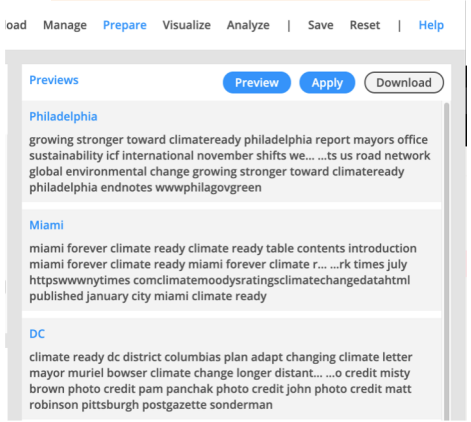


# Lexos: Applying your Preparations

BEFORE PREP

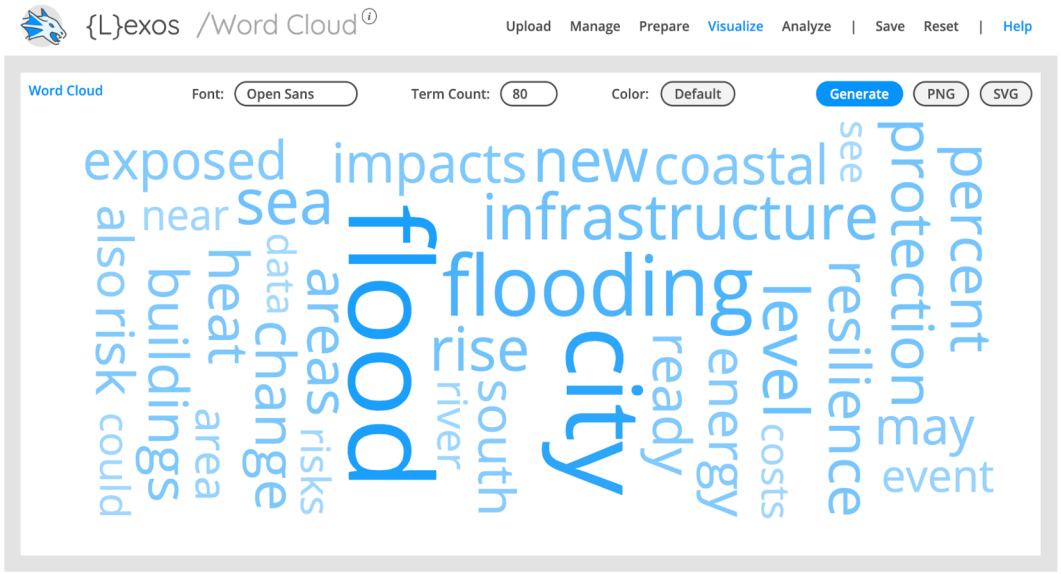



AFTER PREP



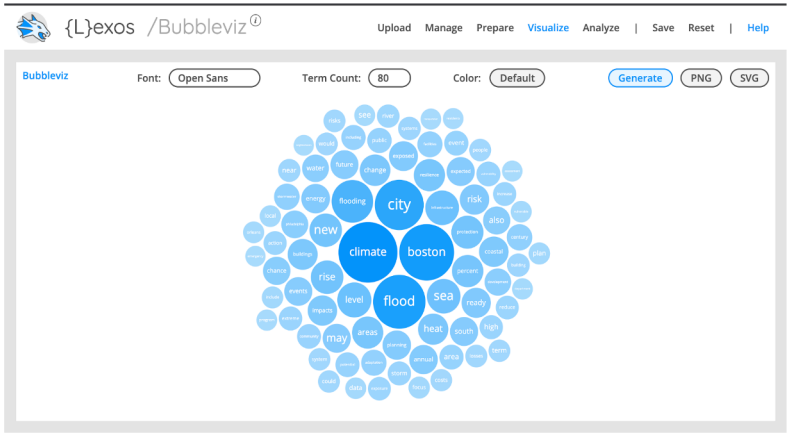

Once you have made decisions about your preparations, click "**Apply**" and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your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

# Lexos: Visualize




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

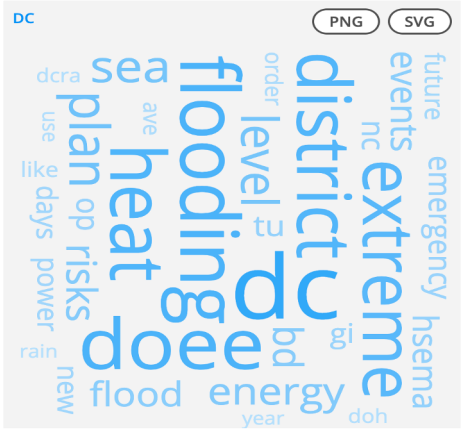
Bubbleviz: visualize word counts through bubbles across the entire text/corpus.




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

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

# Lexos: Visualize > Multicloud





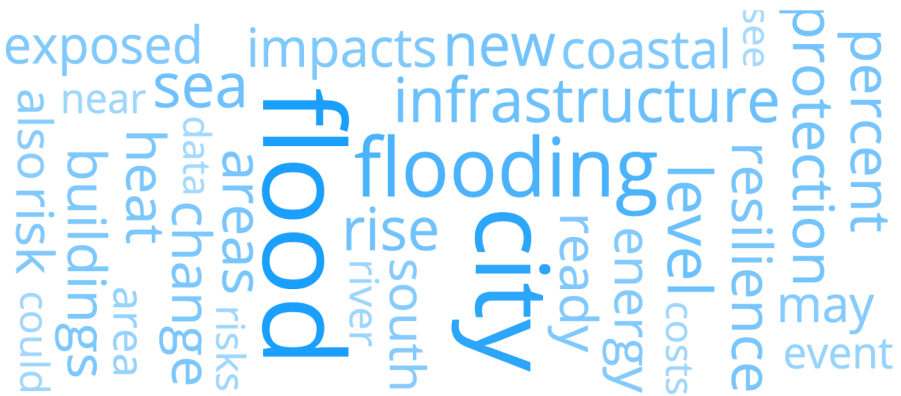
Northeastern University
*NULab for Texts, Maps, and Networks*

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

# Voyant vs. Lexos: Wordclouds

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


Lexos Wordcloud




What could be causing this distinction? This helps demonstrate the importance of understanding what a tool is doing to the texts in the background.



*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:

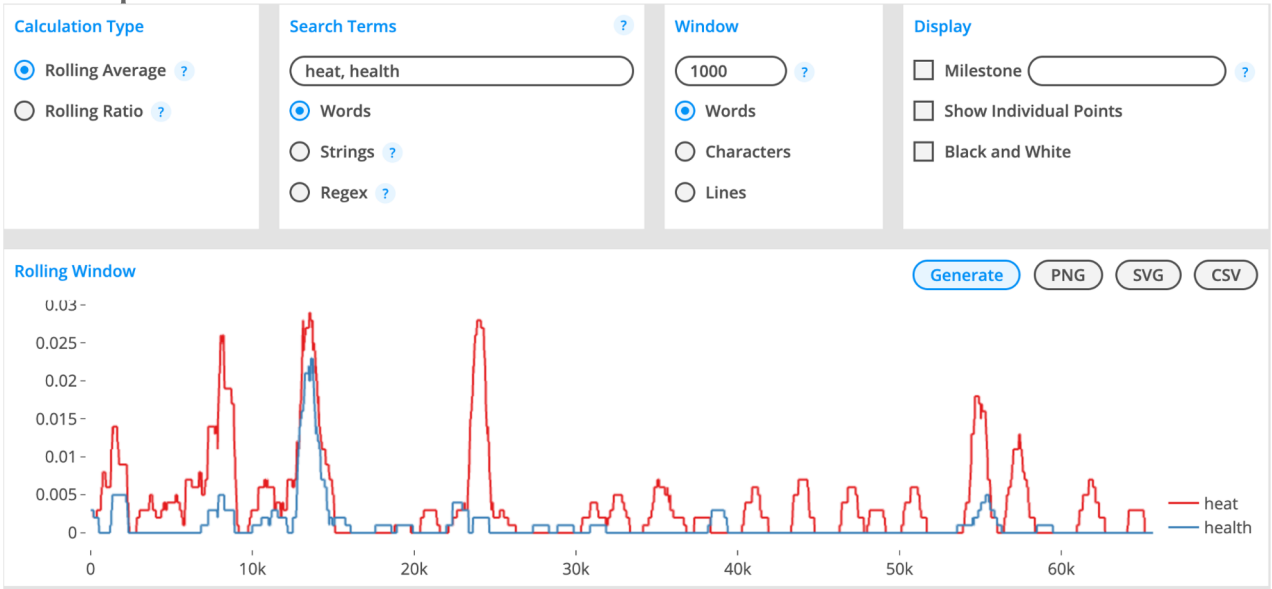
1. Go to “Visualize-> Rolling Window” and type in a search term you want to visualize. You can also search multiple terms by clicking “String” and separating words with a comma (heat, health, flood, storm)
2. Choose a Window size (the number of words each “window” contains). For shorter documents, it’s 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 *Climate Ready Boston*, and searching for the words 'heat' and 'health' with a window of 1000 (large document), we can get an idea of how these terms work together in the report.






*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:

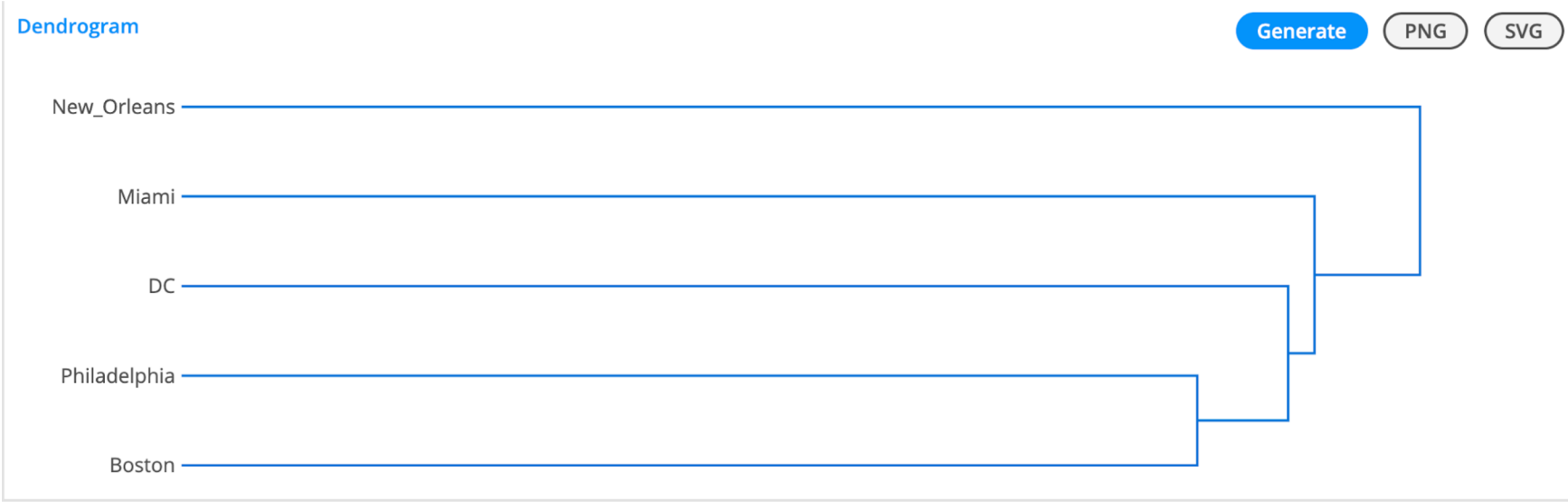
- Similarities between texts
  - The greater the distance between texts, the less similar they are
  - The smaller the distance between texts, the more similar they are



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

# Lexos: Dendrogram

The dendrogram demonstrates similarity between the different documents.






*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.



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

# Conclusion



*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 different Lexos and Voyant features!**

Discussion Prompts

- What do you find challenging or exciting about these tools?
- 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!*

# BARI Data

Boston Area Research Initiative

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TDKDJJ

| | Commerce & Economic Activity | Housing & Development | Urban Planning & Infrastructure | Crime & Disorder | Public Health |
|---|---|---|---|---|---|
| *Online Data* | | | | | |
| Craigslist | X | X | | | |
| Yelp | X | | | | |
| Airbnb | X | X | | | |
| Places of Interest | X | X | X | | |
| *Admin. Data* | | | | | |
| Property Assessments | | X | X | | |
| Building Permits | | X | X | | |
| 911 Dispatches | | | | X | |
| BOS:311 Reports | | | | X | X |
| Code and Property Violations | | | | X | X |
| Food Inspections | X | | | | X |
| CityScore | | | X | X | X |



*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

**Developed by Vaishali Kushwaha**
**Delivered by Vaishali Kushwaha and Adam Tomasi**
DITI Research Fellows
Digital Integration Teaching Initiative

Slides, handouts, and data available at **http://bit.ly/diti-spring2021-aldrich**

You also have access to DITI Canvas Module on Computational Text Analysis.
Schedule an appointment with us! **http://bit.ly/diti-office-hours**



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