Computational Text Analysis for Digital Histories

Taught by Tieanna Graphenreed and Colleen Nugent

ENGL 3161: Composite Aesthetics -- Race as Technology

Eunsong Kim

Fall 2021

Northeastern University

NULab for Texts, Maps, and Networks

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

Today’s Agenda

- Define computational text analysis
- Follow along in a demonstration of web-based text analysis tools by DITI Fellows
  - Word Counter, Word Trees, Lexos, Voyant
- Experiment with text analysis tools on your own

Slides, handouts, and data available at **http://bit.ly/diti-fall2021-kim-textanalysis**

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

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

Computational Text Analysis

Computational text analysis refers to an array of methods that can be used to "read" texts with a computer. This form of analysis can range from basic word frequency counts to more advanced techniques like machine learning.

Text analysis is often used on a **corpus**, or a collection of multiple texts, and provides a glimpse into patterns across the texts. Some people also perform text analysis on larger individual documents, like novels or autobiographies.

Why Computational Text Analysis?

Computational text analysis can help us analyze very large amounts of data and discover **patterns** in texts.

Particular disciplines care deeply about the language that writers use and how this language may reach intended audiences. Text analysis provides another method for approaching these questions.

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

Our Text

Our text is a plain text (.txt file) of the Introduction and Chapter 1 of *Dark Matters: On the Surveillance of Blackness* by Simone Browne, 2015. *Dark Matters* is a monograph on contemporary surveillance technologies, using blackness and Black life as a site upon which surveillance functions (and is resisted).

In the version of the text used for the examples below, the chapter titles and frontispiece lists were removed as part of data preparation. Data prep is incredibly important for text analysis; always be thoughtful about what you specifically want to analyze.

Creating a Corpus

You will not need to create a corpus today, since we'll be working with **one text**, but the steps are actually the same!

Steps:

1. Choose the text(s) you'd like to use in your corpus.
2. Save the original texts in a folder, with consistent naming conventions, where you can easily retrieve them. **Note: It's good practice to keep unmodified copies of the original documents (PDF, .docx, etc.) in case you need to recreate your plain-text files (.txt)**
3. Open a plain text editor (Notepad for PC, TextEdit for Mac)
4. Copy-paste the contents into individual plain-text files, and save with the appropriate .txt extension (ex. browne_chapter-1.txt)

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

Tips for creating a corpus

- .txt files are ideal because they standardize and remove formatting -- HTML files are often easier to copy/paste than PDFs
- TextEdit on Macs: You must make sure it is configured to work with plain text files. To do this, open Text Edit and go to “Preferences” and make sure “plain text editor” is selected. Then, restart TextEdit.
- Only copy one text into each new plain text file. Make sure not to put any spaces in the names of the files as you save them. Use underscores or hyphens to mark spaces between words instead.
- Make sure to use detailed and consistent names for your files, and think about how you might want to take advantage of sorting by filename.

Preparing Your Text

1. Navigate to your digital version of Simone Browne's *Dark Matters* (you may access this through the Northeastern Library search function; be sure to sign in with your student login)
2. Copy and paste the text into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
   a. Mac users, you will need to make your Text Edit into a plain text editor. Open Text Edit, go to Preferences, and make sure "plain text" is selected
3. 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!

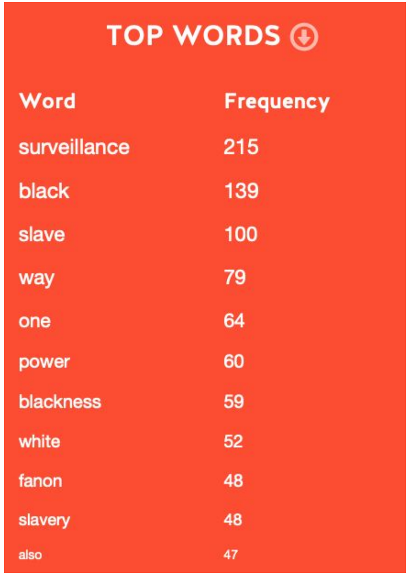
Exploratory Tools

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
- Can be run with and without **stopwords**

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

Word Counter Examples

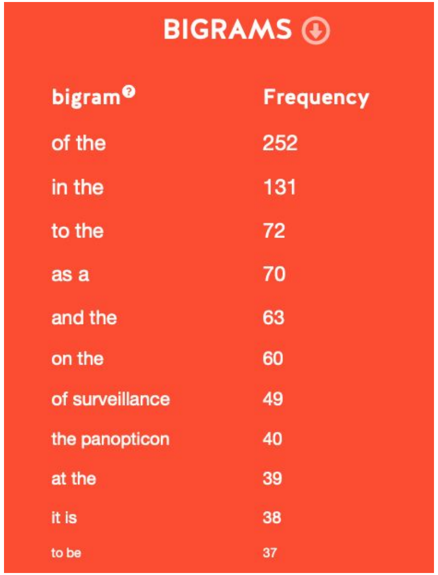


TOP WORDS

Word	Frequency
surveillance	215
black	139
slave	100
way	79
one	64
power	60
blackness	59
white	52
fanon	48
slavery	48
also	47

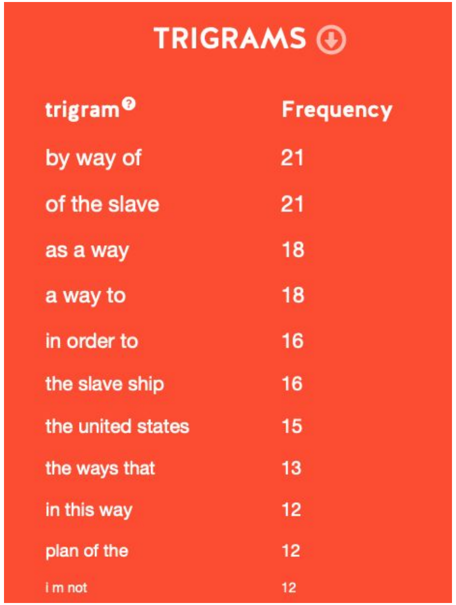
Shows the top words in the text.

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



BIGRAMS

bigram	Frequency
of the	252
in the	131
to the	72
as a	70
and the	63
on the	60
of surveillance	49
the panopticon	40
at the	39
it is	38
to be	37



TRIGRAMS

trigram	Frequency
by way of	21
of the slave	21
as a way	18
a way to	18
in order to	16
the slave ship	16
the united states	15
the ways that	13
in this way	12
plan of the	12
i m not	12

It is interesting how many of the trigrams reference movement and locations.

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

Word Trees

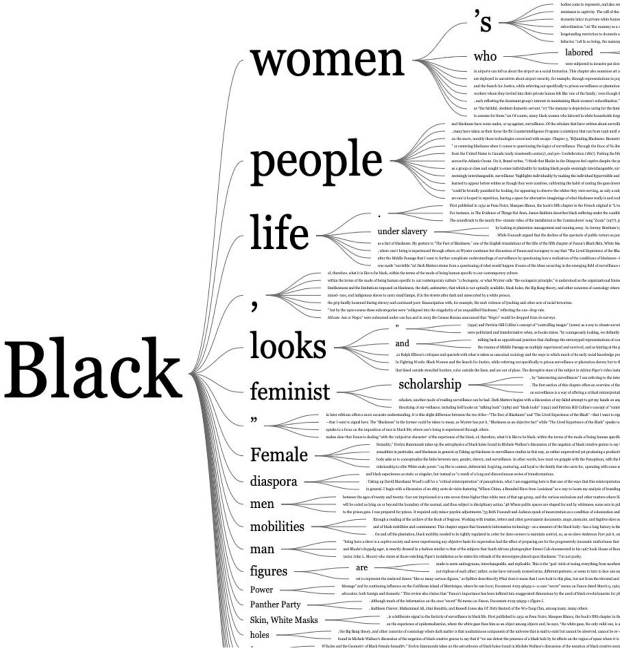
- https://www.jasondavies.com/wordtree/
- A word tree depicts multiple parallel sequences of words
- This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest.
- There are some restrictions in size: fewer than 1 million words should work

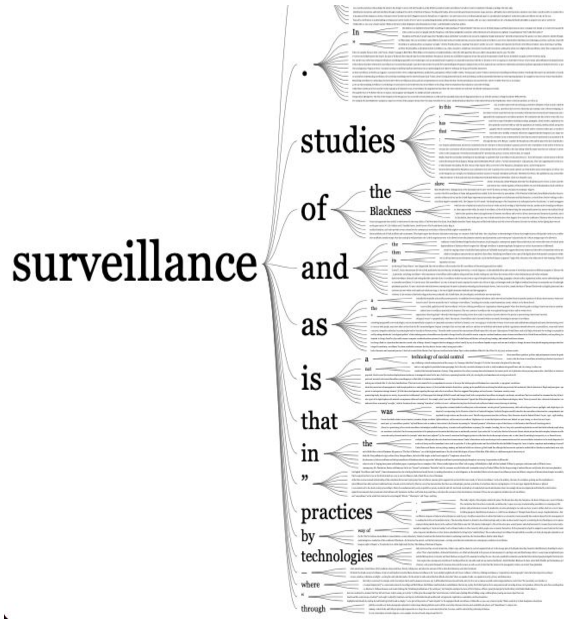
Word Tree Examples

Reflects the focus of the book as on the experiences of Black people, particularly of **Black women**.

Even though this book is largely about the experiences of Black people writ large, discussions about Black men are less frequent. We can attribute this likely to **Browne's feminist approach**.

The punctuation following **surveillance** suggests it is often the end of the sentence.



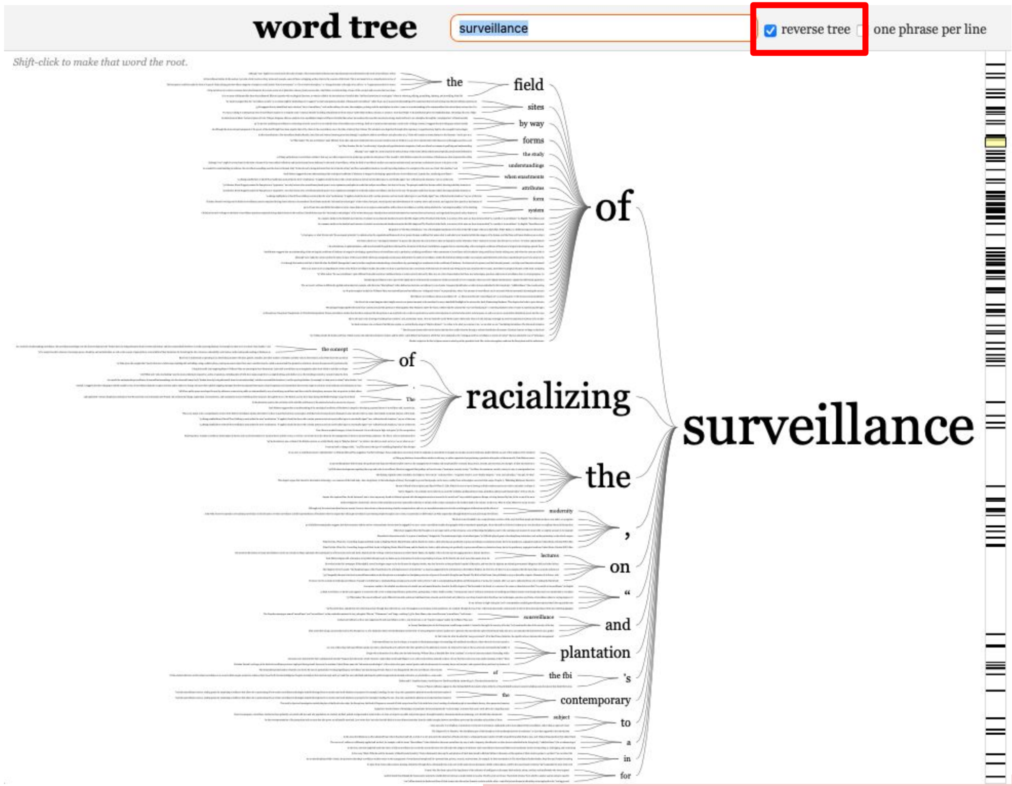


Northeastern University
NULab for Texts, Maps, and Networks

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

Word Tree: Reverse Trees

When words are commonly followed by punctuation, 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.



word tree

surveillance

☑ reverse tree □ one phrase per line

Shift-click to make that word the root.

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

Lexos

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

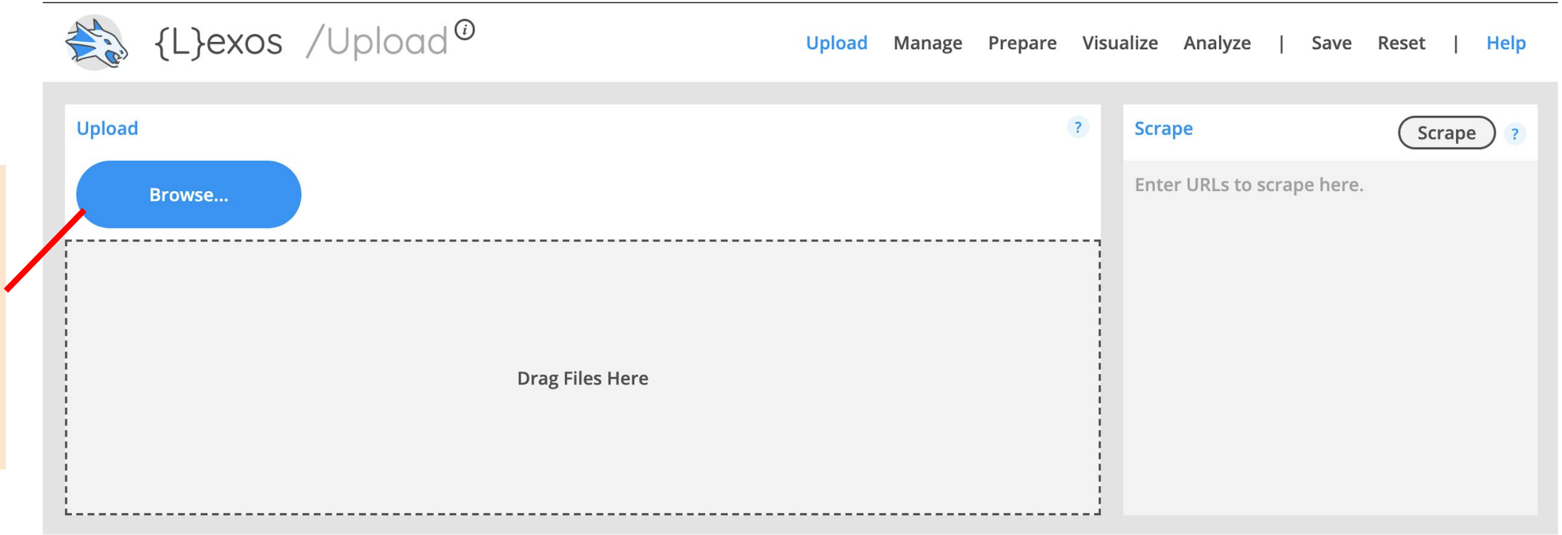
Lexos: http://lexos.wheatoncollege.edu/upload

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

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

Lexos: Upload

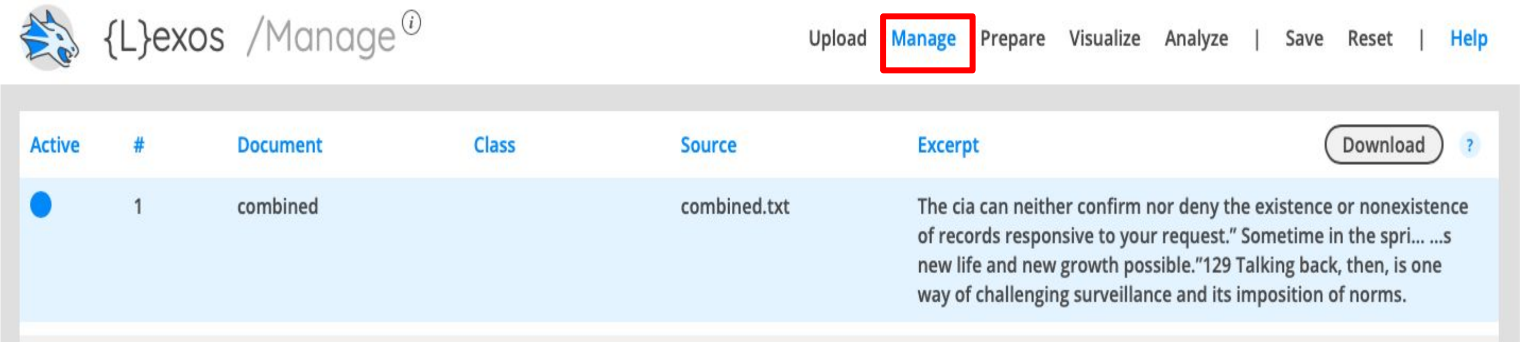


Click Browse and select your entire text (or drag file into the “Drag Files Here” area)

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)



Active	#	Document	Class	Source	Excerpt
●	1	combined		combined.txt	The cia can neither confirm nor deny the existence or nonexistence of records responsive to your request." Sometime in the spri... ...s new life and new growth possible."129 Talking back, then, is one way of challenging surveillance and its imposition of norms.

Lexos: Prepare (scrub)

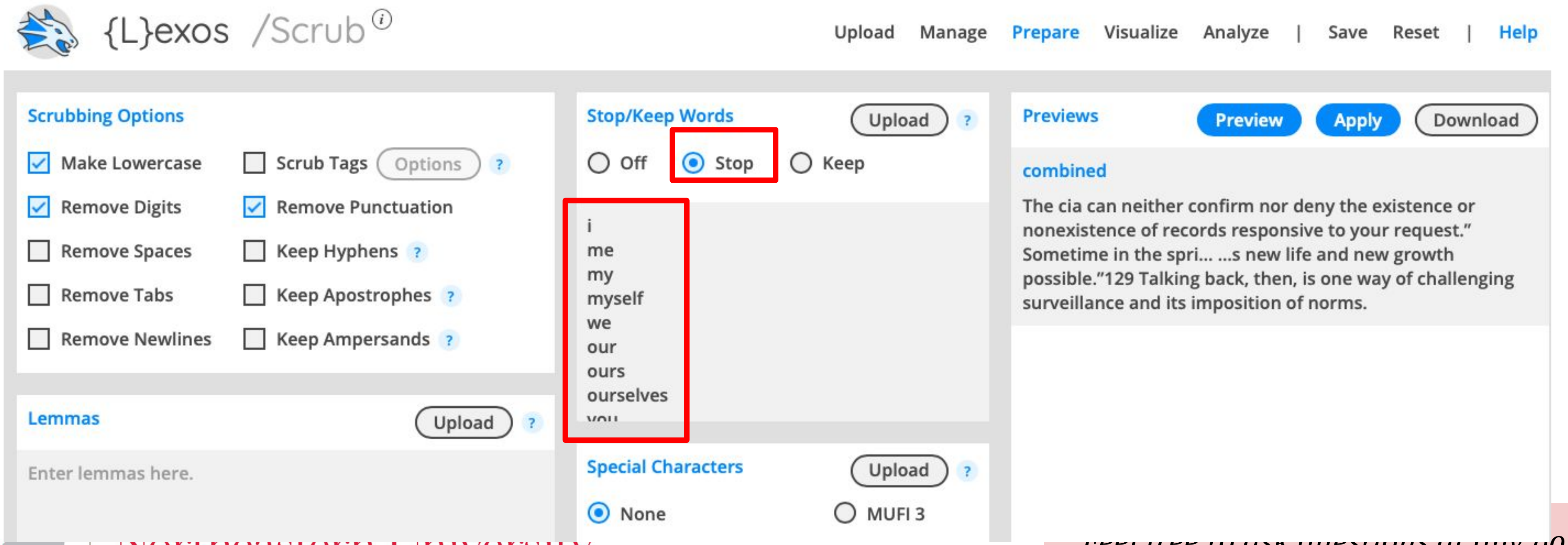
Lexos demonstrates the different options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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"



Scrubbing Options

☑ Make Lowercase □ Scrub Tags (Options) ?

☑ Remove Digits ☑ Remove Punctuation

□ Remove Spaces □ Keep Hyphens ?

□ Remove Tabs □ Keep Apostrophes ?

□ Remove Newlines □ Keep Ampersands ?

Lemmas (Upload) ?

Enter lemmas here.

Stop/Keep Words (Upload) ?

○ Off ● Stop ○ Keep

i
me
my
myself
we
our
ours
ourselves
you

Special Characters (Upload) ?

● None ○ MUFI 3

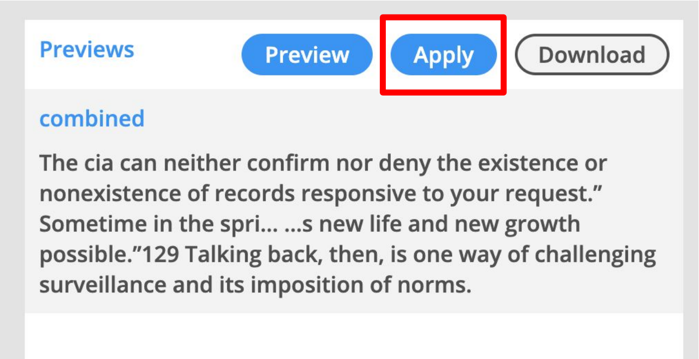
Previews (Preview) (Apply) (Download)

combined

The cia can neither confirm nor deny the existence or nonexistence of records responsive to your request." Sometime in the spri... ...s new life and new growth possible."129 Talking back, then, is one way of challenging surveillance and its imposition of norms.

Lexos: Applying your Preparations

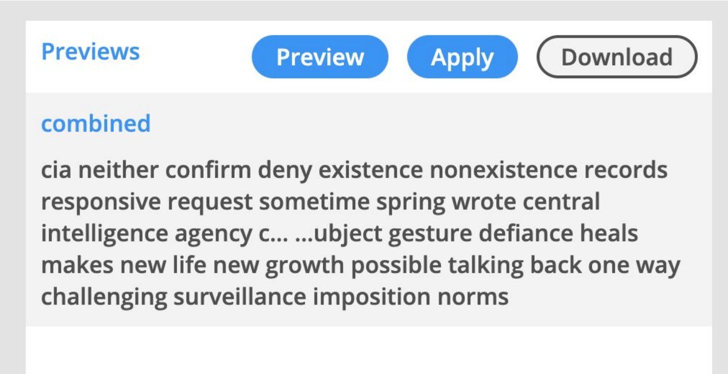
BEFORE PREP



combined

The cia can neither confirm nor deny the existence or nonexistence of records responsive to your request." Sometime in the spri... ...s new life and new growth possible."129 Talking back, then, is one way of challenging surveillance and its imposition of norms.

AFTER PREP



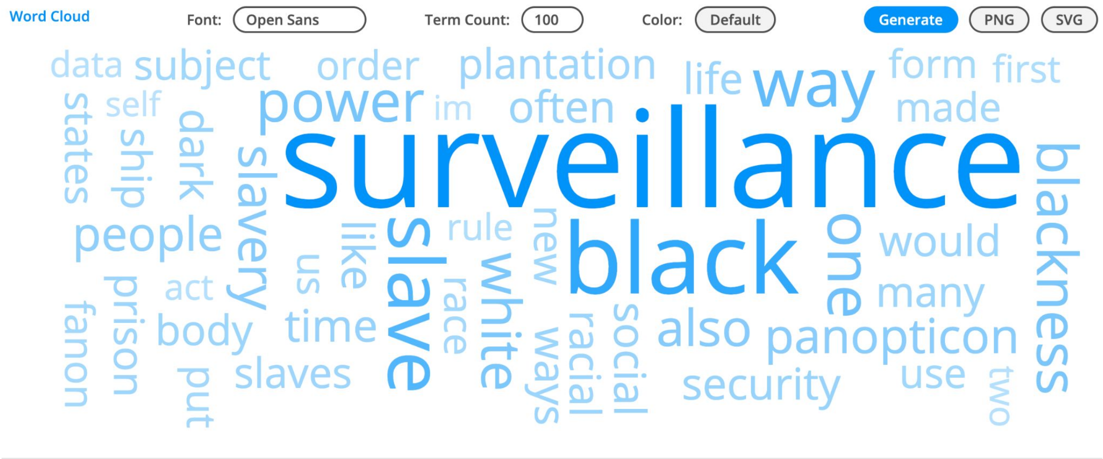
combined

cia neither confirm deny existence nonexistence records responsive request sometime spring wrote central intelligence agency c... ...ubject gesture defiance heals makes new life new growth possible talking back one way challenging surveillance imposition norms

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

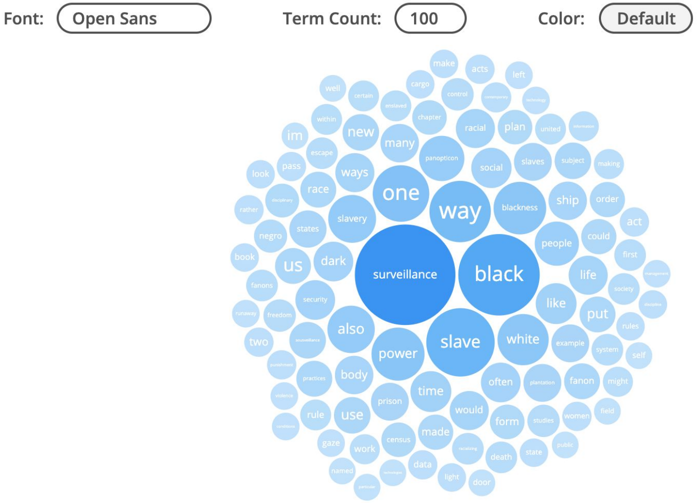
Font: Open Sans

Term Count: 100

Color: Default

Word Cloud: visualize a wordcloud across the entire text. Note the similarity to the wordcloud generated by the Word Counter tool!

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



Font: Open Sans

Term Count: 100

Color: Default

Lexos: Rolling Window

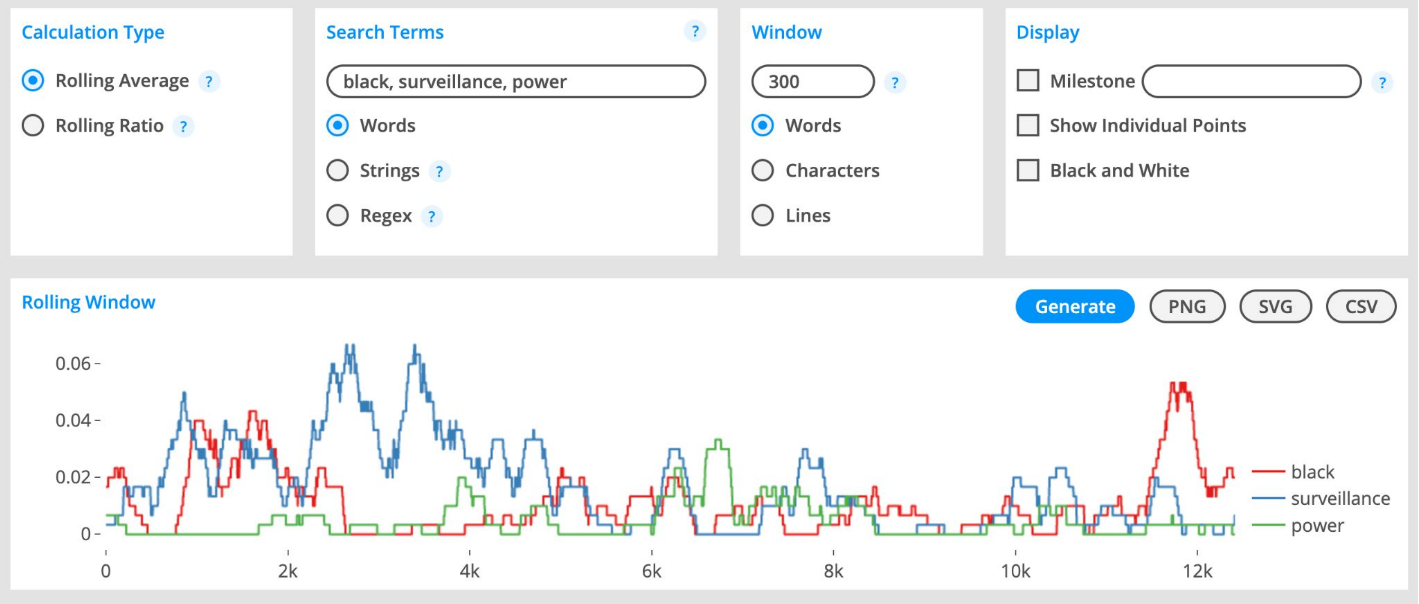
Rolling windows allow you to look at word trends across **one** document. To use a rolling window:

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 (jewish, russia, america)
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 *Dark Matters*, and searching for the strings “Black, surveillance, power” with a window of 300, we can get an idea of how different terms work together in the book. You may also be interested in **contrasting** terms to see how they’re used across a text.



Calculation Type

● Rolling Average ?
○ Rolling Ratio ?

Search Terms ?

black, surveillance, power

● Words
○ Strings ?
○ Regex ?

Window

300 ?

● Words
○ Characters
○ Lines

Display

□ Milestone ?
□ Show Individual Points
□ Black and White

Rolling Window

Lexos: Dendrogram

The dendrogram demonstrates similarity between the different documents.

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

Once you have more of your corpus built, you can analyze your texts further by using the tools in the “**Analyze**” tab.

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 save individual visualizations as images (PNGs).

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

Voyant

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

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

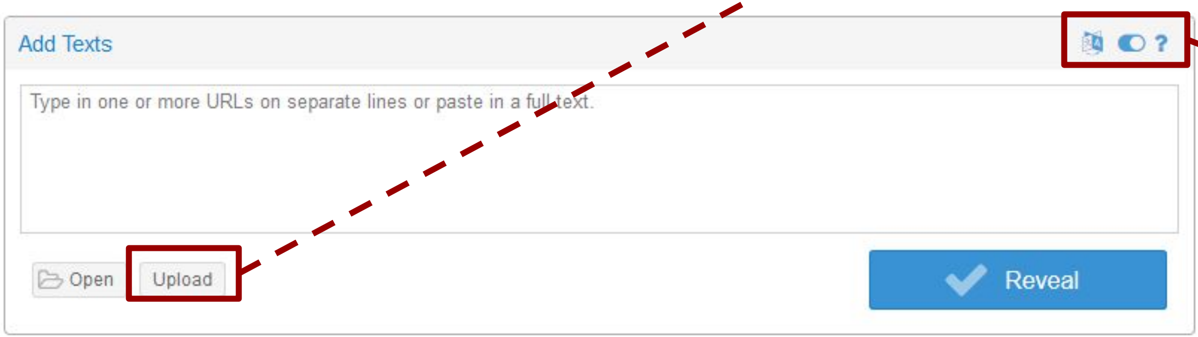
Voyant makes it possible to perform analyses on one or multiple files in many ways, including word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances. It also makes nice visualizations!

Click "Upload" and choose all the texts you want to analyze.

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

VOYANT

see through your text



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

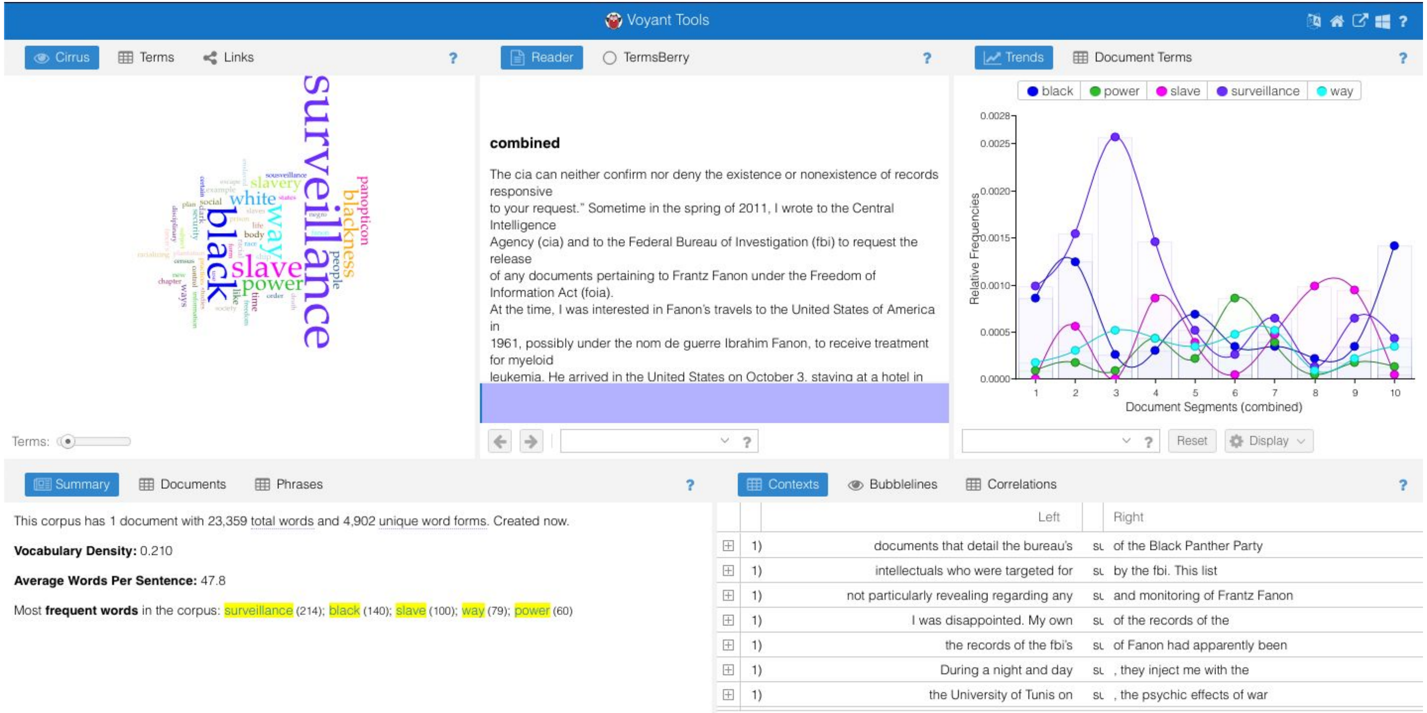
Voyant: Understanding the Dashboard

Results:

From Browne's monograph you can see the default results page with multiple panes:

- A wordcloud
- Reader section
- Trends
- Document Summary
- Word Contexts

These boxes can all be changed!



combined

The cia can neither confirm nor deny the existence or nonexistence of records responsive
to your request." Sometime in the spring of 2011, I wrote to the Central Intelligence
Agency (cia) and to the Federal Bureau of Investigation (fbi) to request the release
of any documents pertaining to Frantz Fanon under the Freedom of Information Act (foia).
At the time, I was interested in Fanon's travels to the United States of America in
1961, possibly under the nom de guerre Ibrahim Fanon, to receive treatment for myeloid
leukemia. He arrived in the United States on October 3, staying at a hotel in

This corpus has 1 document with 23,359 total words and 4,902 unique word forms. Created now.

Vocabulary Density: 0.210

Average Words Per Sentence: 47.8

Most **frequent words** in the corpus: surveillance (214); black (140); slave (100); way (79); power (60)

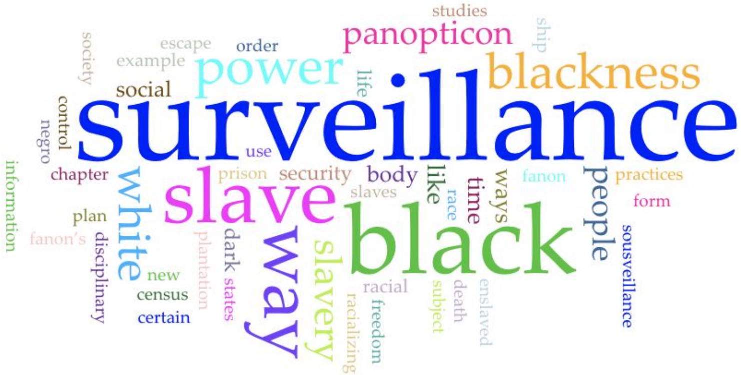
		Left		Right
⊞	1)	documents that detail the bureau's	su	of the Black Panther Party
⊞	1)	intellectuals who were targeted for	su	by the fbi. This list
⊞	1)	not particularly revealing regarding any	su	and monitoring of Frantz Fanon
⊞	1)	I was disappointed. My own	su	of the records of the
⊞	1)	the records of the fbi's	su	of Fanon had apparently been
⊞	1)	During a night and day	su	, they inject me with the
⊞	1)	the University of Tunis on	su	, the psychic effects of war

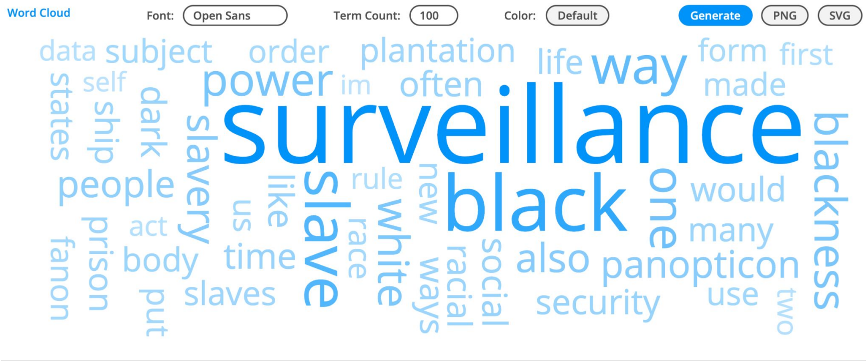
Feel free to ask questions at any point during the presentation!

Voyant vs. Lexos: Wordclouds

Voyant enables a wider array of color options which can help with readability.

Lexos wordcloud




Word Cloud

Font: Open Sans Term Count: 100 Color: Default

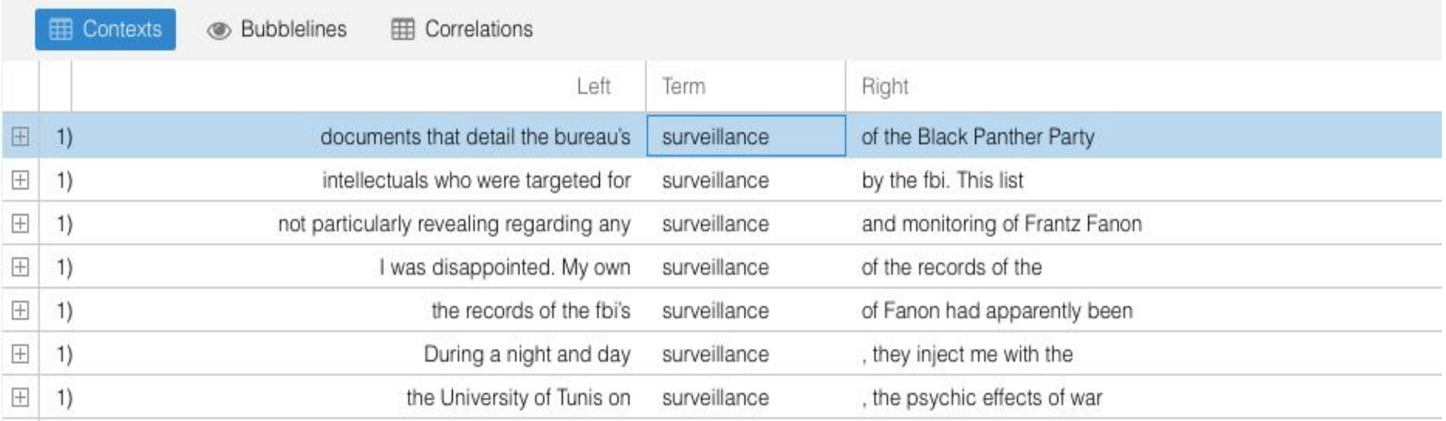
Some key terms are weighted heavier in Voyant than in Lexos. What could be causing this distinction?

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!

Voyant: Contexts (Concordances)

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

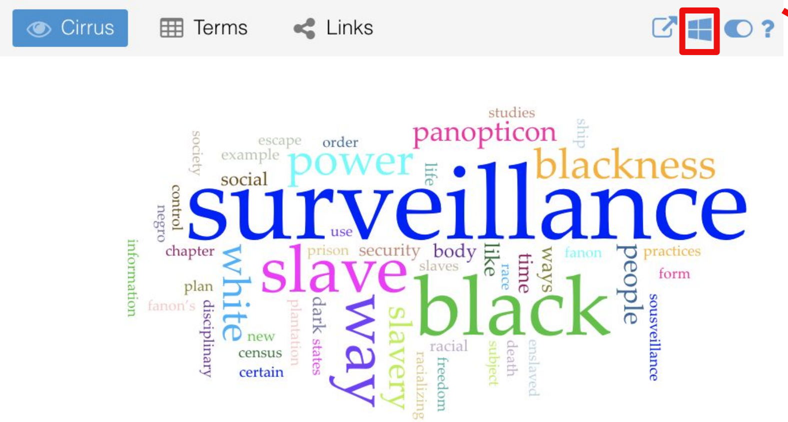


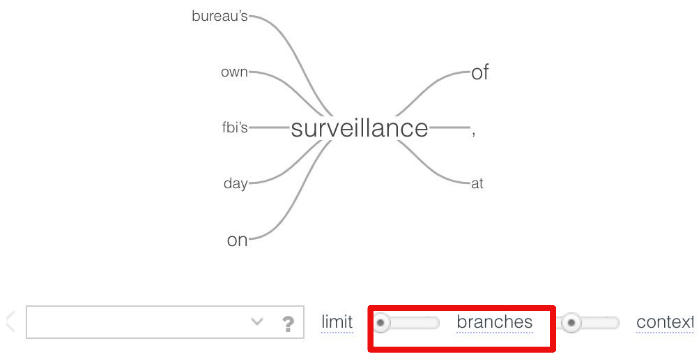
		Left	Term	Right
⊞	1)	documents that detail the bureau's	surveillance	of the Black Panther Party
⊞	1)	intellectuals who were targeted for	surveillance	by the fbi. This list
⊞	1)	not particularly revealing regarding any	surveillance	and monitoring of Frantz Fanon
⊞	1)	I was disappointed. My own	surveillance	of the records of the
⊞	1)	the records of the fbi's	surveillance	of Fanon had apparently been
⊞	1)	During a night and day	surveillance	, they inject me with the
⊞	1)	the University of Tunis on	surveillance	, the psychic effects of war

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

Voyant: Changing Displayed Results

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!

Your Turn!

Using the text prepared from *Dark Matters: On the Surveillance of Blackness,* begin practicing web-browser text analysis

- Follow the “Preparing Your Text” steps to get your .txt file
- Prep your text using any of the four interfaces. Which preparation steps did you choose and why?
    - See what happens if you keep the stopwords. What are some of the most-used verbs and pronouns?

Slides, handout, and data: **http://bit.ly/diti-fall2021-kim-textanalysis**

Post-Exploration Discussion

- What other kinds of sources besides the Browne monograph would be useful with these tools?

- What interesting or surprising results came up in your own explorations?

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 Colleen Nugent and Milan Skobic
Digital Integration Teaching Initiative
DITI Research Fellow

Taught by Tieanna Graphenreed and Colleen Nugent
Digital Integration Teaching Initiative
NULab Research Fellow

Slides, handouts, and data available at:

http://bit.ly/diti-fall2021-kim-textanalysis

Schedule an appointment with us! **http://bit.ly/diti-office-hours**

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