

Computational Text Analysis for Content Analysis

Developed by **Colleen Nugent & Milan Skobic** for
HIST 2430 Digital Histories of Ethnic Boston
Simon Rabinovitch
Fall 2020



Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Defining computational text analysis
- Demonstration of web-based text analysis tools
 - Word Counter, Word Trees, Lexos, Voyant
- Your turn!

Slides, handouts, and data available at

<https://bit.ly/diti-fall2020-rabinovitch2>



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



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 used and how this language may reach intended audiences. Text analysis provides another method for approaching these discourses.



Preparing Your Text

1. Navigate to
<https://digital.library.upenn.edu/women/antin/land/land.html>
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!



Our Text

Our text is a plain text (.txt file) of *The Promised Land* by Mary Antin, 1912. This is an autobiography of Mary Antin, describing her early life in Belarus and immigration to the United States.

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.



Exploratory Tools



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



Word Counter Examples

TOP WORDS ⬇

Word	Frequency
could	356
one	314
would	308
us	292
little	269
time	247
life	218
father	210

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	885
in the	726
i was	497
it was	371
to the	319
on the	309
and the	296
i had	272

TRIGRAMS ⬇

trigram [®]	Frequency
i did not	89
that i was	61
i could not	47
i do not	45
it was a	40
and i was	40
there was no	38
it was not	35

It is interesting how many of the trigrams are negations!



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

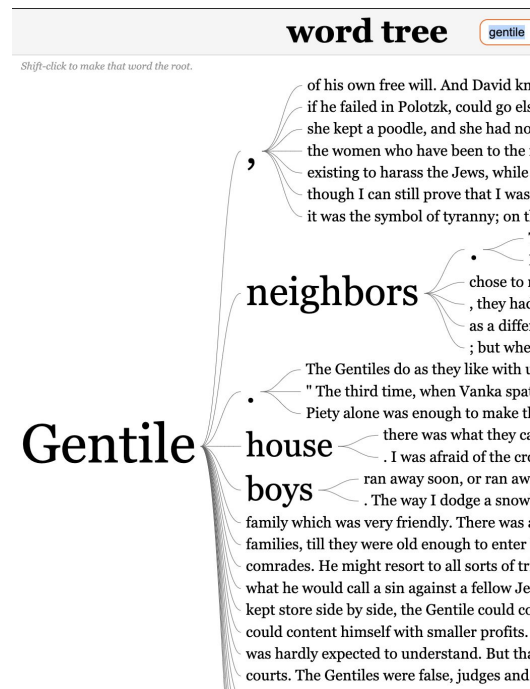


Word Tree Examples

Reflects the focus of the book as on the Jewish community, while Gentiles are more likely neighbors.

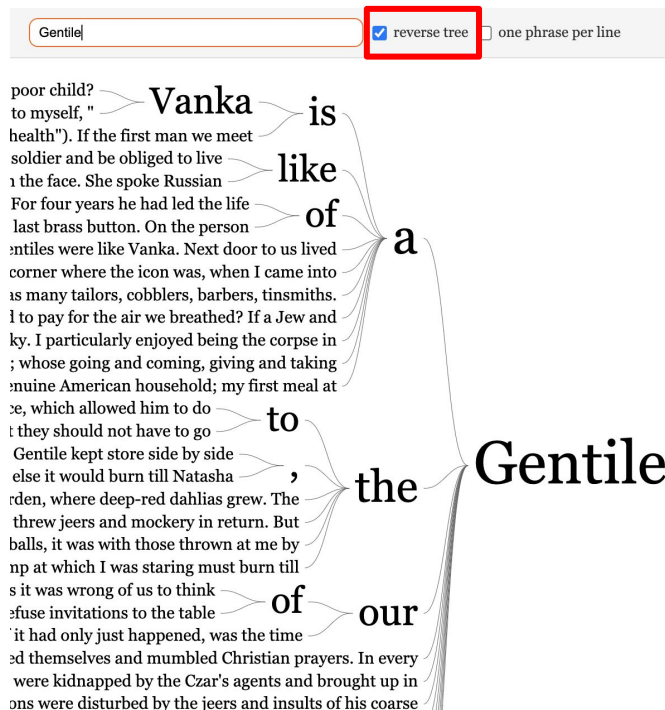
Even though written by a woman, the book references Jewish and Gentile boys more than girls.

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



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.



Lexos



Northeastern University
NULab for Texts, Maps, and Networks

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

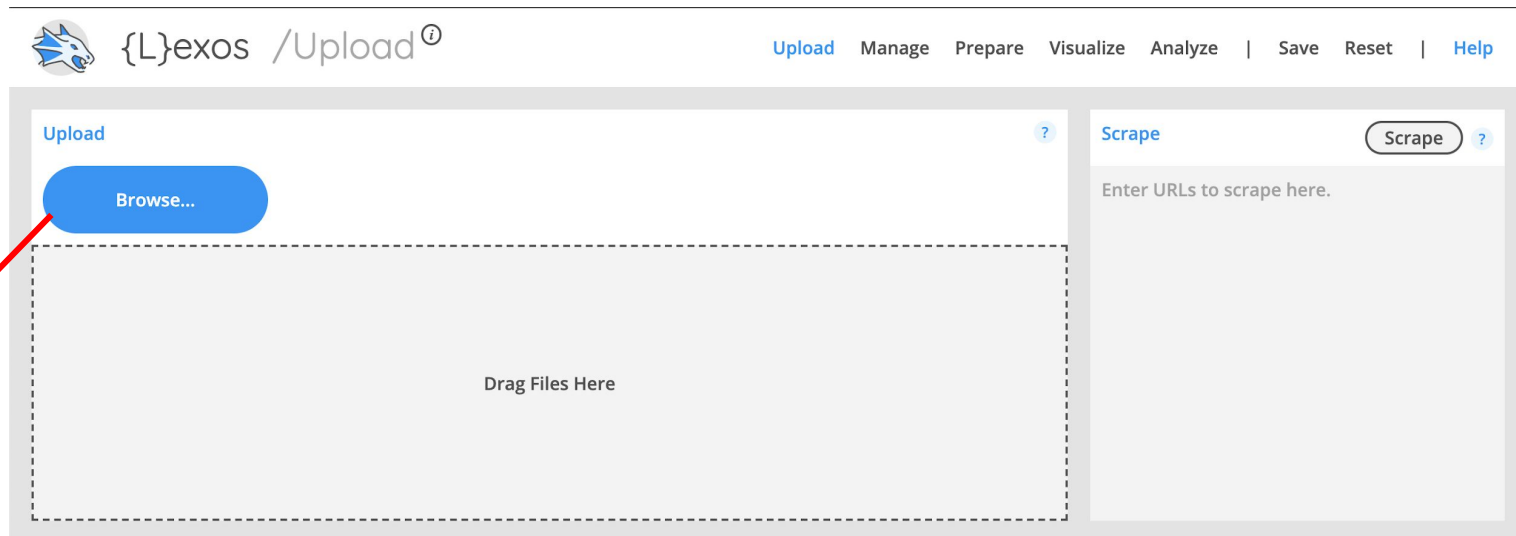
Lexos: <http://lexos.wheatoncollege.edu/upload>

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



Lexos: Upload



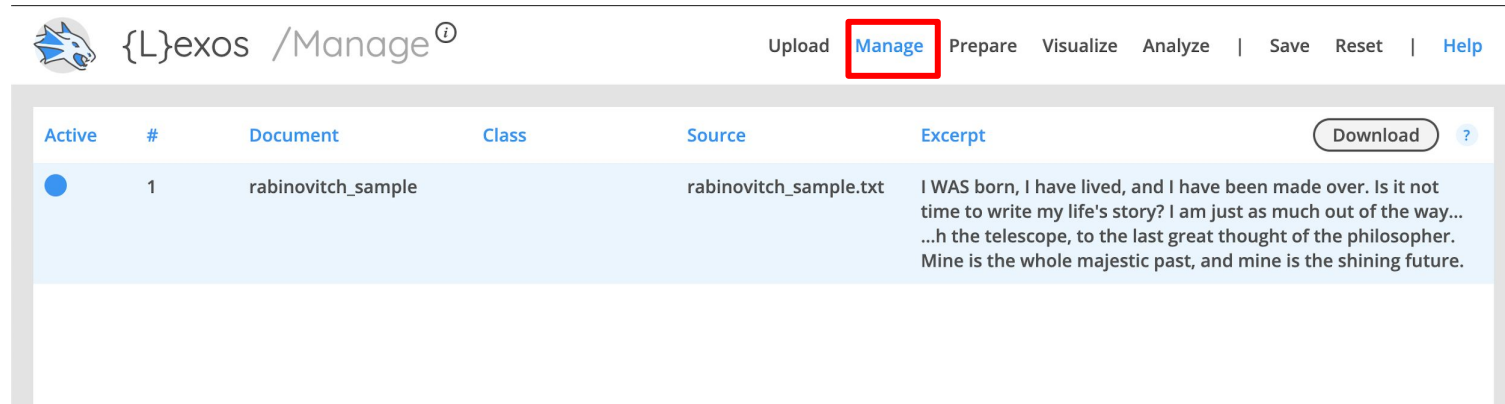
The screenshot shows the Lexos web application interface. At the top, there is a navigation bar with the Lexos logo (a blue fox head) and the text "{L}exos /Upload". To the right of the navigation bar are several links: "Upload", "Manage", "Prepare", "Visualize", "Analyze", "Save", "Reset", and "Help". The main content area is divided into two panels. The left panel is titled "Upload" and contains a blue "Browse..." button and a large dashed box labeled "Drag Files Here". A red arrow points from the "Browse..." button to the text box on the left. The right panel is titled "Scrape" and contains a "Scrape" button and a text area labeled "Enter URLs to scrape here."

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



Lexos: Manage

Make sure the document you want to use is selected (blue = selected, gray = not selected)



The screenshot shows the Lexos Manage interface. At the top, there is a navigation bar with a logo, the text "{L}exos /Manage", and a series of buttons: Upload, Manage (highlighted with a red box), Prepare, Visualize, Analyze, Save, Reset, and Help. Below the navigation bar is a table with the following columns: Active, #, Document, Class, Source, and Excerpt. The first row of the table is highlighted in blue, indicating it is selected. The row contains a blue circle in the Active column, the number 1 in the # column, the text "rabinovitch_sample" in the Document column, an empty cell in the Class column, the text "rabinovitch_sample.txt" in the Source column, and an excerpt of text in the Excerpt column. A "Download" button is located in the top right corner of the table area.

Active	#	Document	Class	Source	Excerpt
<input checked="" type="radio"/>	1	rabinovitch_sample		rabinovitch_sample.txt	I WAS born, I have lived, and I have been made over. Is it not time to write my life's story? I am just as much out of the way... ...h the telescope, to the last great thought of the philosopher. Mine is the whole majestic past, and mine is the shining future.



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”



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”

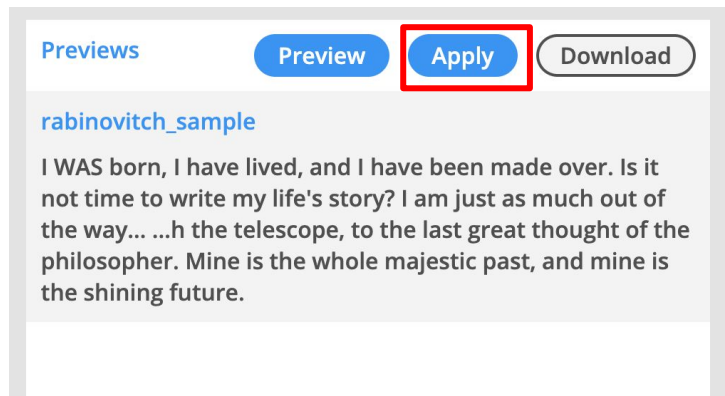
The screenshot shows the Lexos web interface with several sections:

- Scrubbing Options:** Includes checkboxes for 'Make Lowercase', 'Remove Digits', 'Remove Spaces', 'Remove Tabs', 'Remove Newlines', 'Scrub Tags', 'Remove Punctuation', 'Keep Hyphens', 'Keep Apostrophes', and 'Keep Ampersands'. There are also 'Options' and 'Upload' buttons.
- Stop/Keep Words:** Features radio buttons for 'Off', 'Stop' (selected and highlighted with a red box), and 'Keep'. Below the buttons is a text area containing a list of stopwords: 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', and 'just', which is also highlighted with a red box. An 'Upload' button is present.
- Special Characters:** Includes radio buttons for 'None' (selected), 'Early English HTML', 'MUFI 3', and 'MUFI 4'. An 'Upload' button is also present.
- Previews:** Includes buttons for 'Preview', 'Apply', and 'Download'. Below these is a preview of the text: 'rabinovitch_sample' followed by the paragraph: 'I WAS born, I have lived, and I have been made over. Is it not time to write my life's story? I am just as much out of the way... ..h the telescope, to the last great thought of the philosopher. Mine is the whole majestic past, and mine is the shining future.'



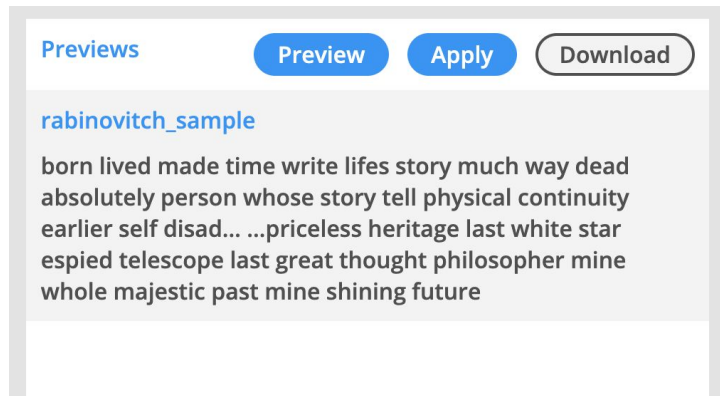
Lexos: Applying your Preparations

BEFORE PREP



The interface shows a 'Previews' tab with three buttons: 'Preview' (blue), 'Apply' (blue and highlighted with a red border), and 'Download' (grey). Below the buttons, the text 'rabinovitch_sample' is displayed in blue. The main text area contains the following paragraph: 'I WAS born, I have lived, and I have been made over. Is it not time to write my life's story? I am just as much out of the way... ..h the telescope, to the last great thought of the philosopher. Mine is the whole majestic past, and mine is the shining future.'

AFTER PREP



The interface shows the same 'Previews' tab with 'Preview', 'Apply', and 'Download' buttons. The text 'rabinovitch_sample' is displayed in blue. The main text area now shows the processed output: 'born lived made time write lifes story much way dead absolutely person whose story tell physical continuity earlier self disad... ..priceless heritage last white star espied telescope last great thought philosopher mine whole majestic past mine shining future'.

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.

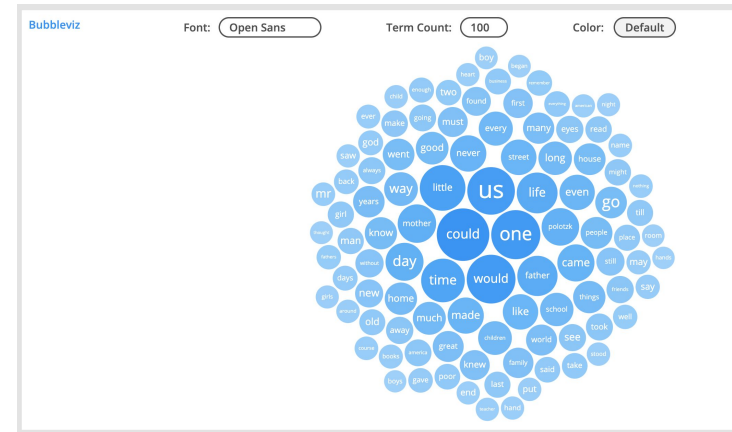


Lexos: Visualize



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

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



Lexos: Rolling Window

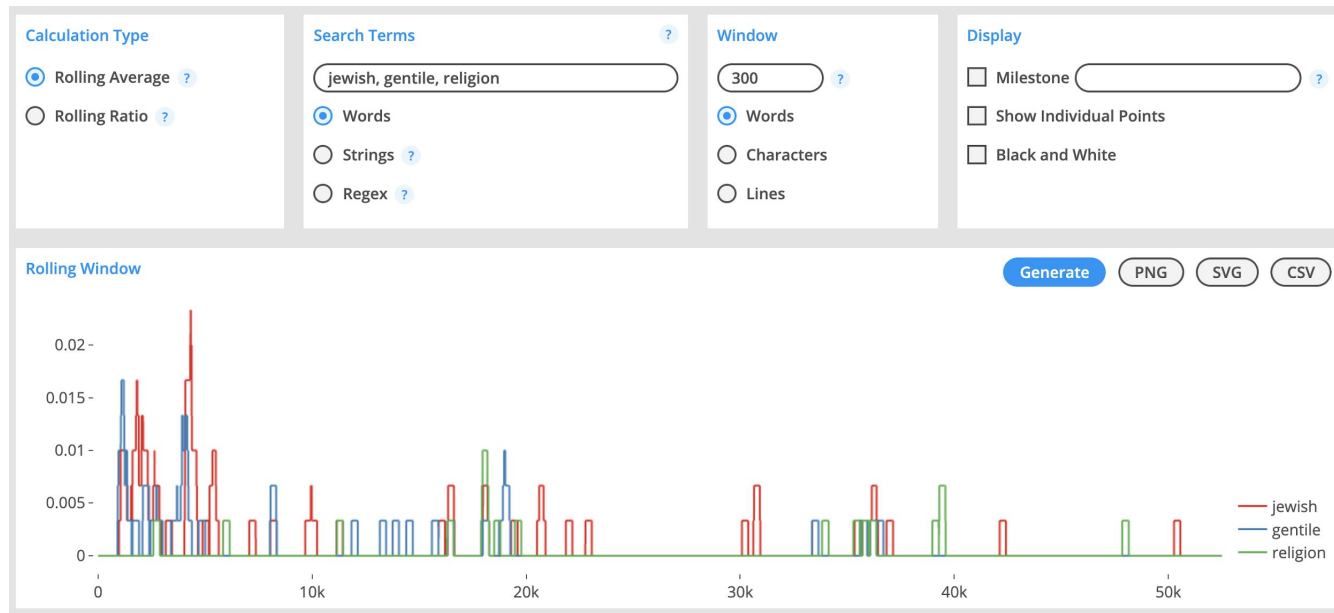
Rolling windows allow you to look at word trends across **one** document. To use a rolling window:

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”



Lexos: Rolling Window Results

Using *The Promised Land*, and searching for the strings “jewish, gentile, religion” 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.



Lexos: Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms are able to show the hierarchical 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

Lexos offers a variety of tools to explore, including Dendrograms. 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



Northeastern University
NULab for Texts, Maps, and Networks

*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.



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

Add Texts

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

Open Upload Reveal



Results:

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

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

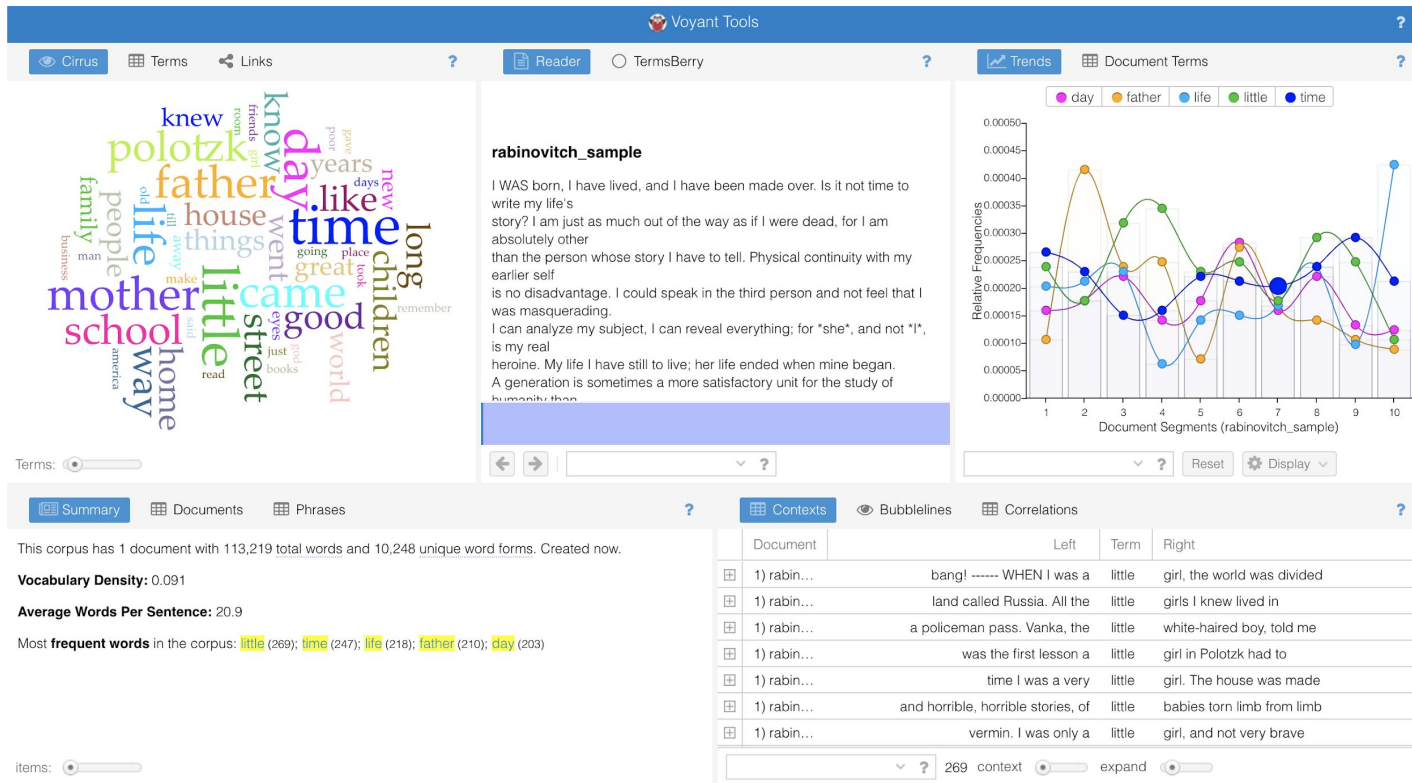
These boxes can all be changed!

Results:

From Antin'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!



Northeastern University
NULab for Texts, Maps, and Networks

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 “little” appears in the text and the contexts in which it appears.

Contexts

Bubblelines

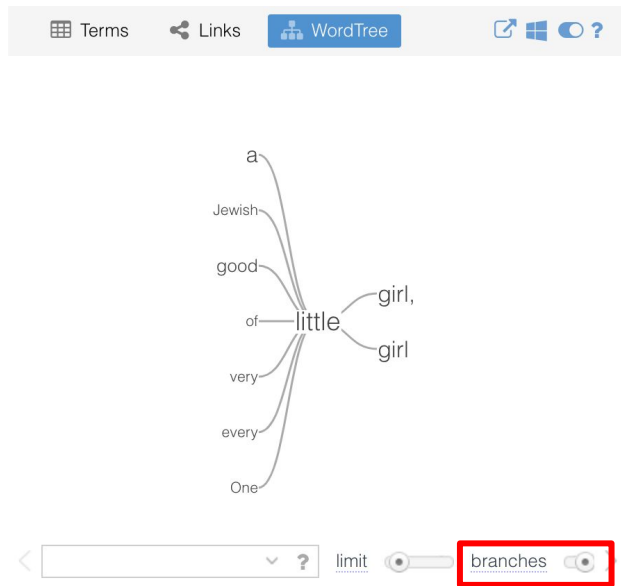
Correlations

	Document	Left	Term	Right
+	1) rabin...	bang! ----- WHEN I was a	little	girl, the world was divided
+	1) rabin...	land called Russia. All the	little	girls I knew lived in
+	1) rabin...	a policeman pass. Vanka, the	little	white-haired boy, told me
+	1) rabin...	was the first lesson a	little	girl in Polotzk had to
+	1) rabin...	time I was a very	little	girl. The house was made
+	1) rabin...	and horrible, horrible stories, of	little	babies torn limb from limb
+	1) rabin...	vermin. I was only a	little	girl, and not very brave



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.



Your Turn!

Using the text prepared from *The Promised Land*, begin practicing web-browser text analysis

- Follow the “Preparing Your Text” steps to get your .txt file
- Prep your text using Lexos. 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?
- Explore different Lexos features.
- Explore different Voyant features.

Slides, handout, and data: <https://bit.ly/diti-fall2020-cilke>



Post-Exploration Discussion

- 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?
 - What language is often used in these documents?
 - What values might be reflected by this language?



Thank you!

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

Developed by Colleen Nugent

Digital Integration Teaching Initiative

DITI Research Fellow

Slides, handouts, and data available at

<https://bit.ly/diti-fall2020-rabinovitch2>

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



Northeastern University
NULab for Texts, Maps, and Networks

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