Computational Text Analysis for Digital Histories

Developed by Colleen Nugent & Milan Skobic
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

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.

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.

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 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.

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

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 texts you'd like to use.
  - 2. Save the texts in a folder, with consistent naming conventions, where you can easily retrieve them.
  - 3. Open a plain text editor (Notepad for PC, TextEdit for Mac)
  - 4. Copy-paste the contents into individual text files (ex. Antin_Promised_Land.txt)
  - 5. Create a spreadsheet for metadata

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
- Create a metadata spreadsheet in the same folder with useful info
- 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.

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

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!

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

Exploratory Tools

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**

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

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!

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

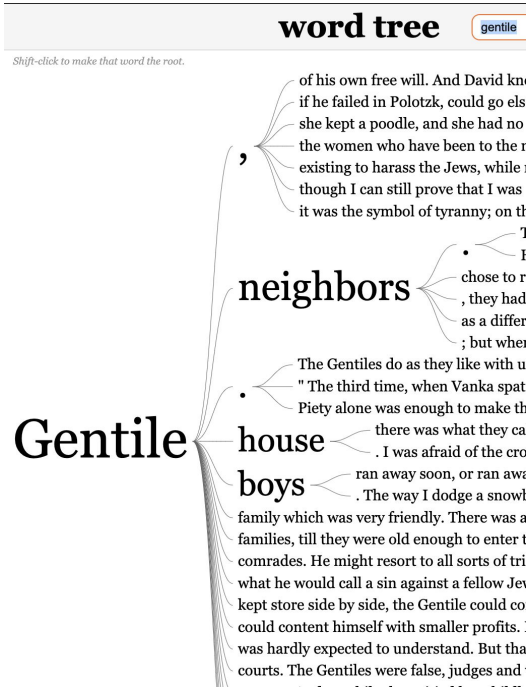
Feel free to ask questions at any point during the presentation!

Word Tree 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.



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.

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

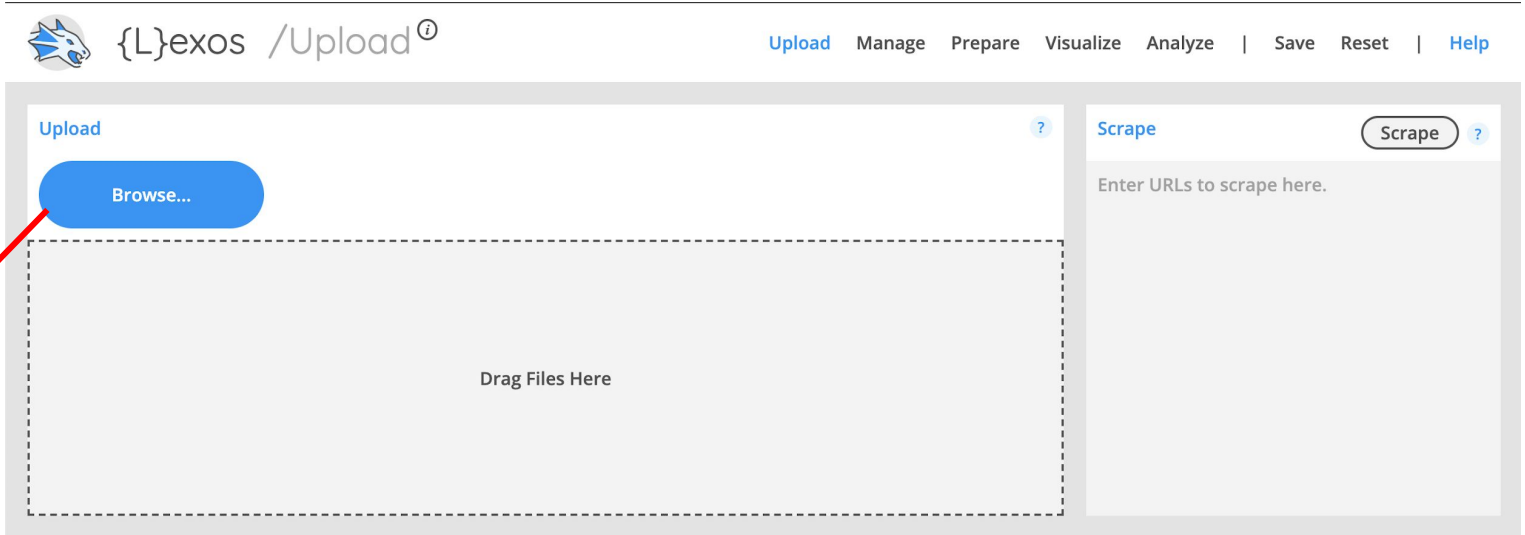
Lexos provides a step-by-step guide for text uploading, preparation, and analysis.

- **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

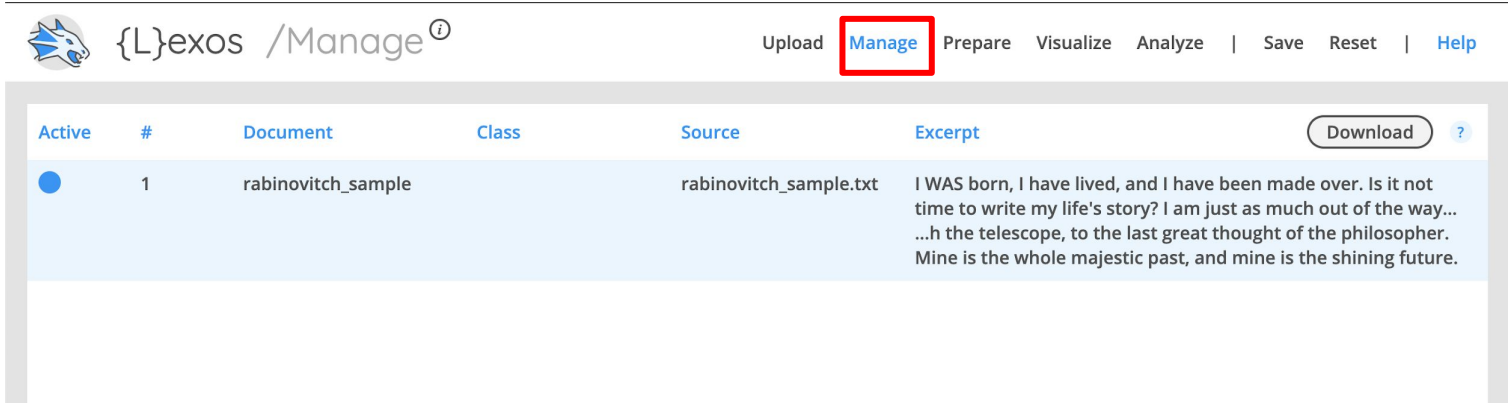
Click Browse and select your entire text (or drag file into the "Drag Files Here" area)



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)



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

Lexos: Prepare (scrub)

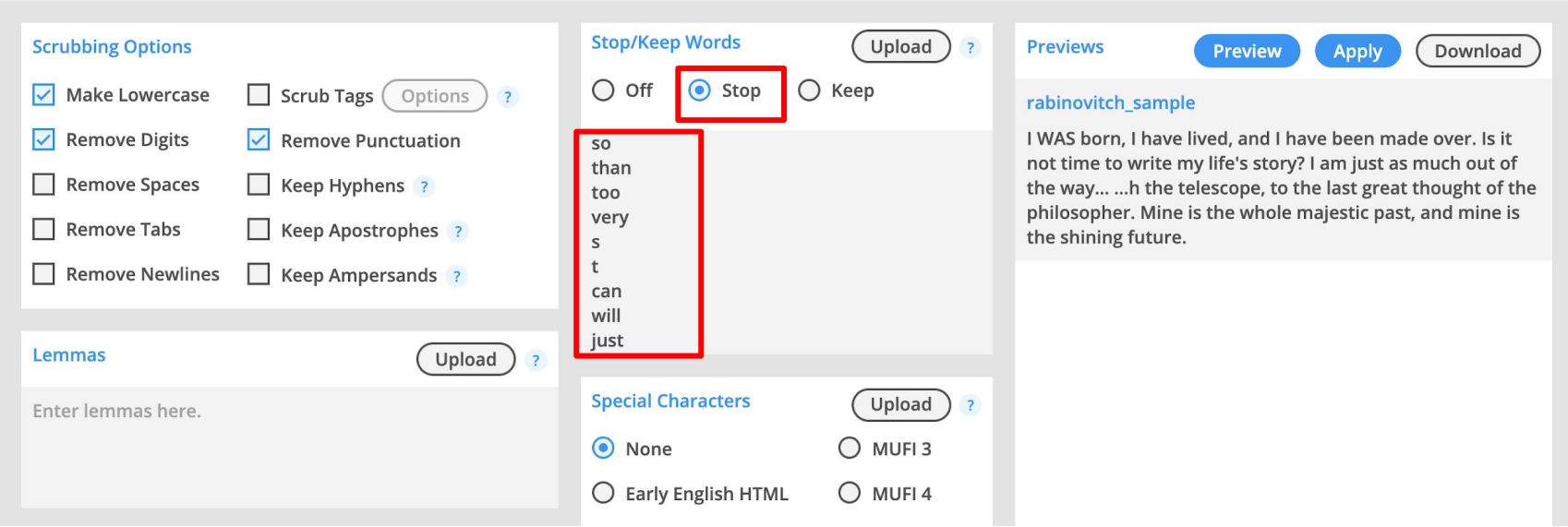
Lexos demonstrates the different options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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

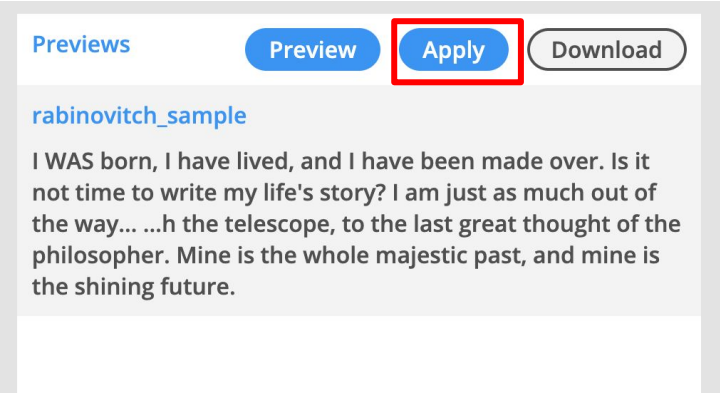
Get a list of English stopwords here: https://gist.github.com/sebleier/554280 (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the "Stop/Keep Words" box then select "Stop"



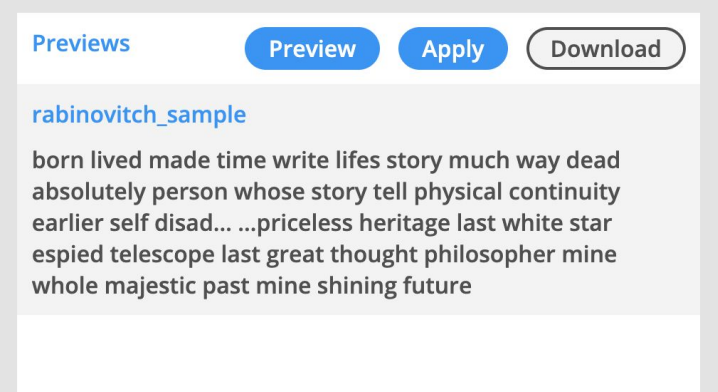
Feel free to ask questions at any point during the presentation!

Lexos: Applying your Preparations

BEFORE PREP



AFTER PREP

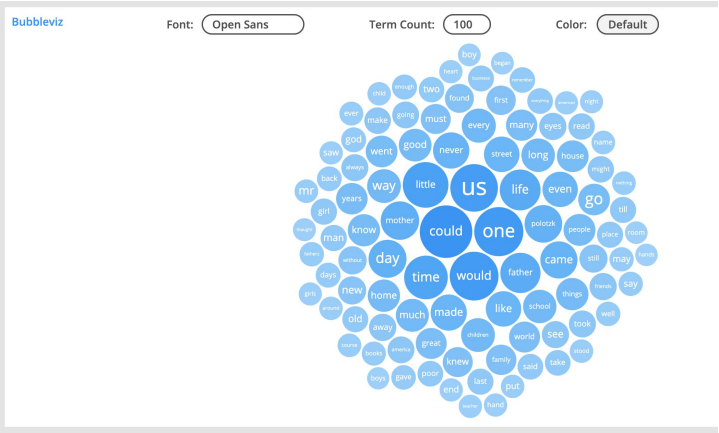


Once you have made decisions about your preparations, click "**Apply**" and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your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

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

Lexos: Visualize

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



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

Northeastern University
NULab for Texts, Maps, and Networks

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

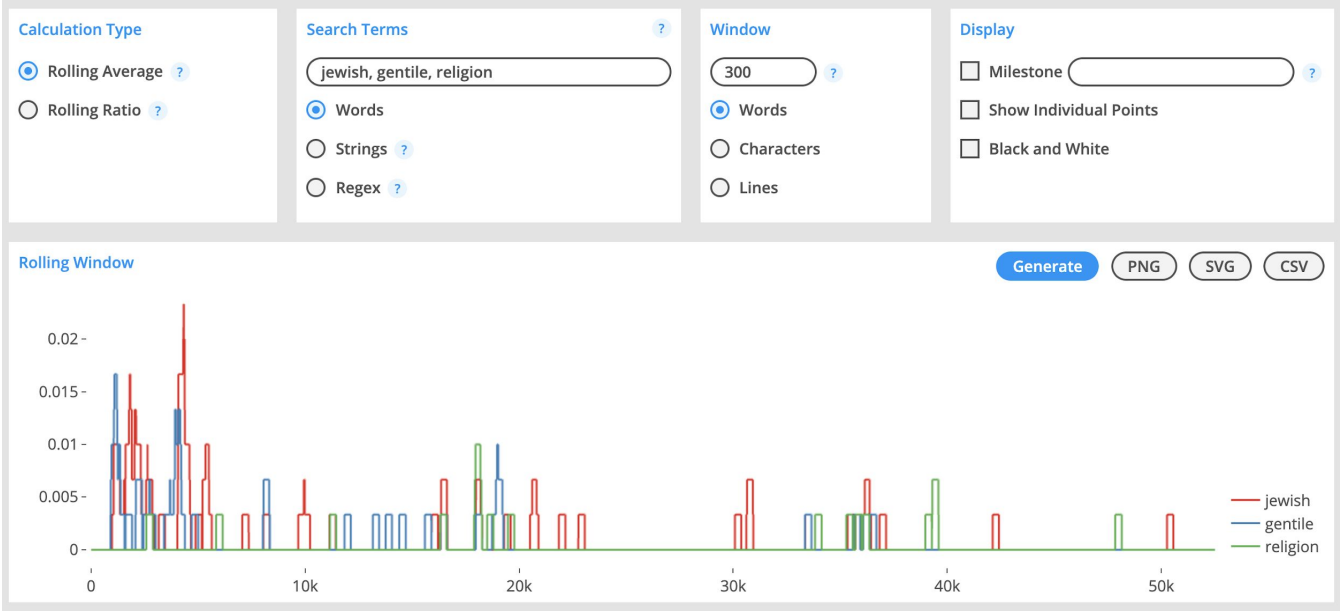
Lexos: Rolling Window

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 (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.



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

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.

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 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!

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

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

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

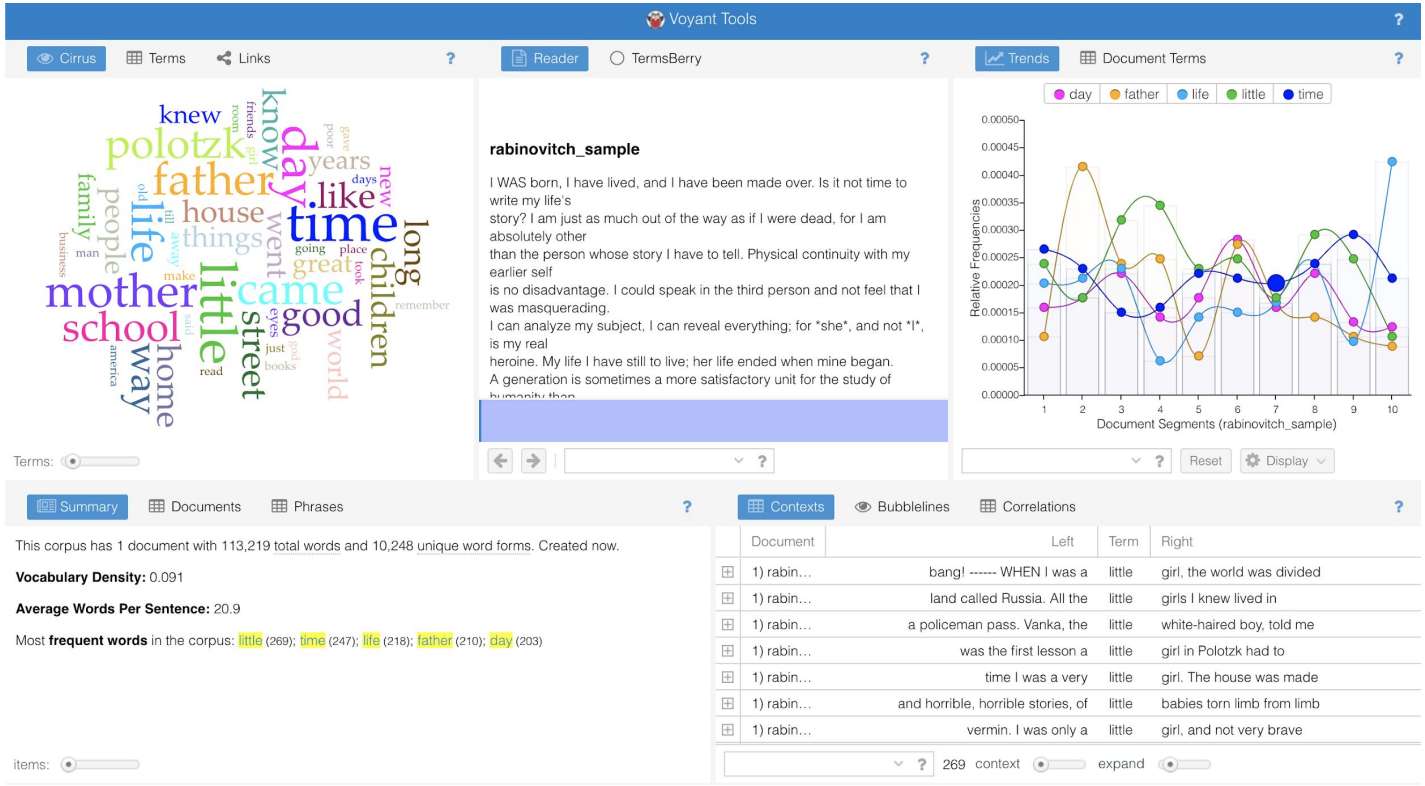
Voyant: Understanding the Dashboard

Results:

From Antin's auto-biography 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!



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 one 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!

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.

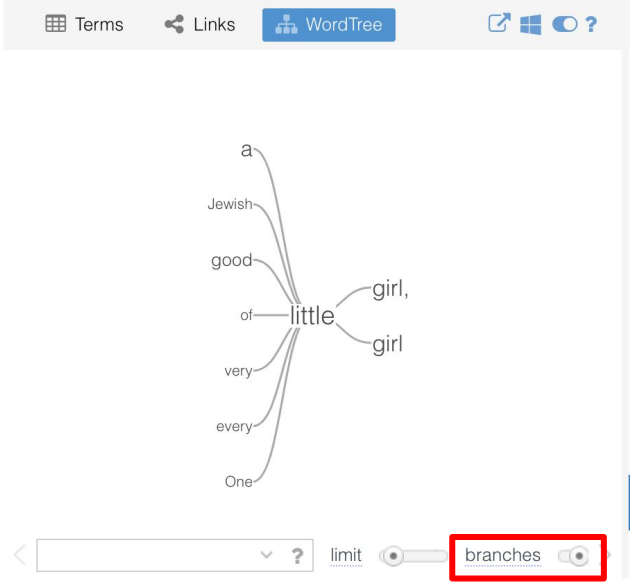
Contexts | Bubblelines | Correlations

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

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.

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

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 any of the four programs. 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: **https://bit.ly/diti-fall2020-rabinovitch2**

Post-Exploration Discussion

- What other kinds of sources besides the Antin book 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
Digital Integration Teaching Initiative
DITI Research Fellow

Taught by Adam Tomasi and Talia Brenner
Digital Integration Teaching Initiative
NULab 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**

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