

Data Ethics and Text Analysis

Cara Marta Messina and Garrett Morrow
Punishment in the Age of Mass Incarceration
Megan Denver
Fall 2019



Workshop Agenda

- Introduce 'Big Data' Concepts
- Algorithmic Bias and the Criminal Justice System
- Introducing Text Analysis
- Web-browser Text Analysis Tools:
 - Lexos for corpus preparation and analysis
 - Voyant for analysis

Slides, handouts, and data available at bit.ly/diti-fall2019-denver



Workshop Objectives

- Understand the ways in which technologies reflect cultural, social, and political biases.
- Explore the basic process for machine learning algorithms
- Understand the ways data is being used in society as well as how algorithms impact and shape our daily lives and the criminal justice system
- Learn the basics of computational text analysis and use web-browser text analysis tools, like Lexos and Voyant
- Understand corpus preparation as a form of analysis



Big Data, Data Ethics, and Surveillance



Northeastern University
NULab for Texts, Maps, and Networks

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

What is 'Big Data'?

Big data has been called the 'new oil' by some. Shoshana Zuboff argues that we now live in an era of 'surveillance capitalism,' in which large amounts of information—usually our personal information—are being analyzed quickly and typically used for profit.

The four components of big data are: **volume**, **variety**, **velocity** and **veracity**



Big Data: What is it and why should we care?

- Big data is characterized by its **scale**
- Big data **sources** include: digitized records, social media/internet activity, or sensors from the physical environment.
- Big data is often **privately owned**
 - Example: an insurance company purchasing social media activity from facebook in order to make specific insurance sales decisions.



Google's File on You is 10 Times Bigger Than Facebook's — Here's How to View It

Google, Amazon, Apple, and Microsoft are all central players in “surveillance capitalism” and prey on our data.



Example: If you have **location services** turned on for Google (like if you use Google maps), Google can track your every move. Go to:

<https://www.google.com/maps/timeline>



Ethical Implications

- Cambridge Analytica Controversy
- Big data also raises questions of autonomy, anonymity, privacy, discrimination, and bias.
- Questions to consider:
 - How are we being represented online?
 - How is our data being used?
 - Who is using it and for what purposes?
 - How might it be used in the future?



DIY Cybersecurity and Tightening your Privacy

Want to make your life more private? Follow this “DIY Guide to Feminist Cybersecurity”

<https://hackblossom.org/cybersecurity/>



Algorithmic Bias and the Justice System



Northeastern University
NULab for Texts, Maps, and Networks

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

Risk Assessment: Algorithmic Bias

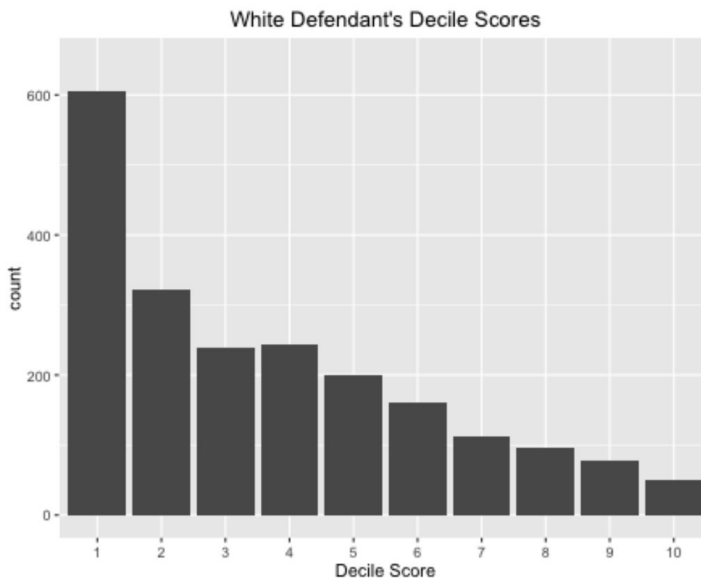
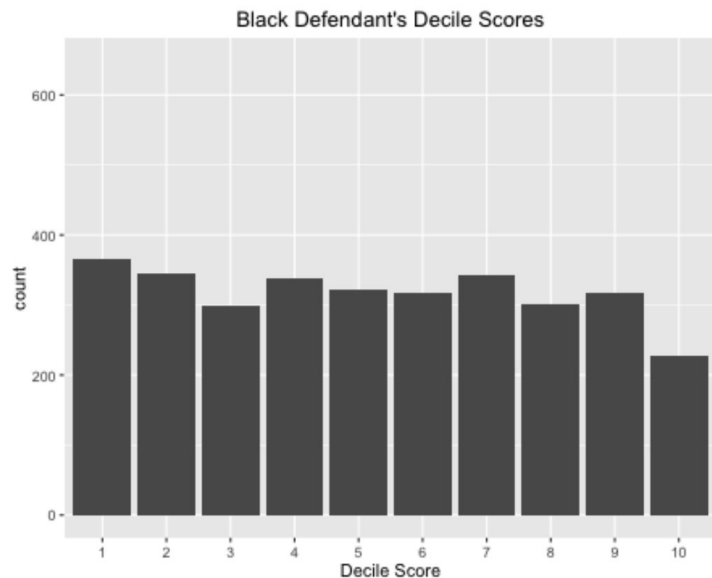
Risk assessment: used to determine the likelihood that someone will reoffend, not appear for trial, etc..

What happens when machine learning algorithms are used to help determine risk assessment?



COMPAS Algorithm & ProPublica's Analysis

The COMPAS recidivism algorithm does not “see” race. Yet...



<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>



Northeastern University
NULab for Texts, Maps, and Networks

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

Initiatives for Justice

Code for America is an organization that works with the mass amount of undigitized, unorganized government documents to help previously incarcerated people.

One project they have, titled “Clear My Record,” attempts to parse through the mass data of governmental records to help clear criminal records, particularly for people who were arrested for marijuana use/distribution.

<https://www.codeforamerica.org/programs/clear-my-record>



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

Computational text analysis is 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 larger **corpus**, or a collection of multiple texts, and provides a glimpse into *patterns* across the texts. Some people also use text analysis on larger documents, like novels.



Why Computational Text Analysis?

Computational text analysis can help us analyze a **ton** of data and discover **patterns** in texts.

Researchers care **deeply** about the language used in judicial & public discourse and how this language may reach intended audiences. Text analysis provides another method for approaching these discourses.



Our Corpus

Our corpus (a collection of texts) collects several State of the Unions from presidents over the years—beginning from Bush Senior to Trump. Our files are a series of plain text (.txt) files.

Download this corpus from our email or from the “data” folder:
LINK



Text Analysis Tools Links

We will be going through several different tools. Links are available on the handout.

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

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



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>

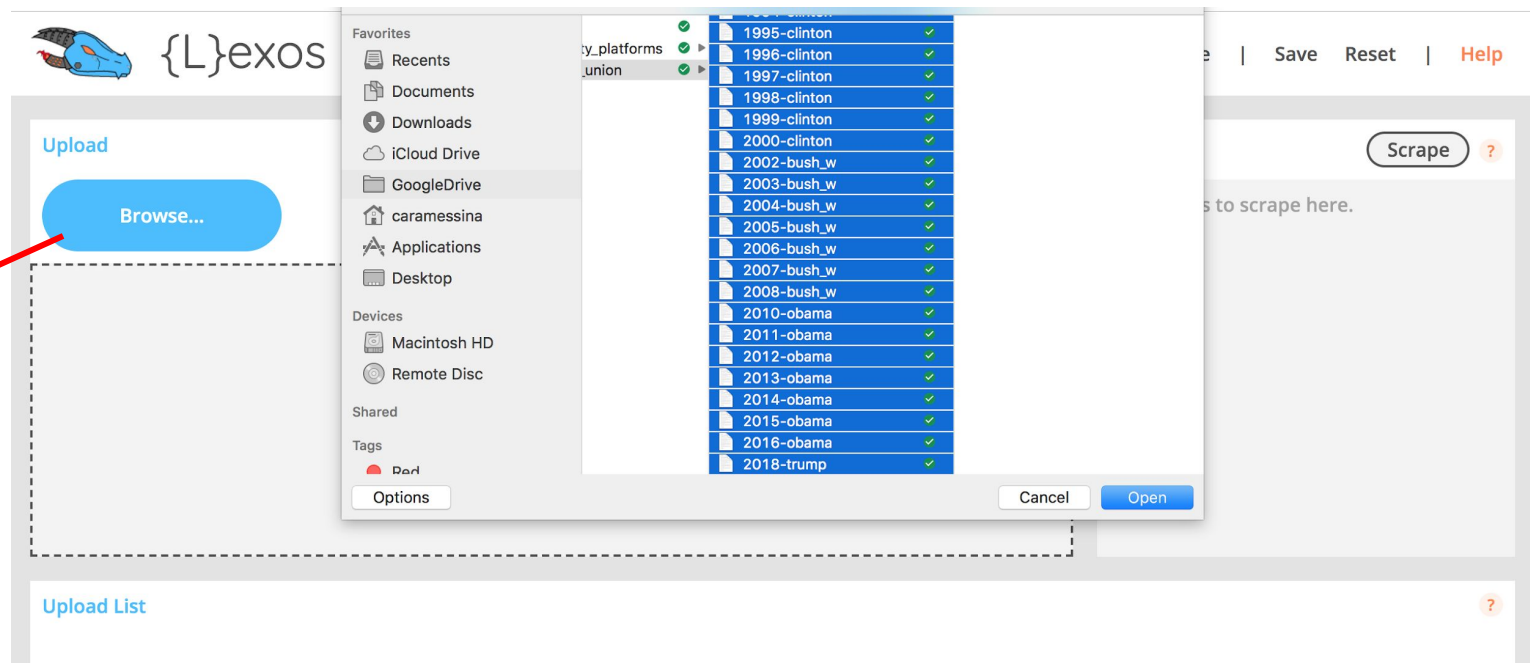
Lexos provides a step-by-step guide for corpus uploading, preparation, and analysis.

- **Upload:** upload your corpus (your separate .txt files)
- **Manage:** select the files you want to prepare and analyze
- **Prepare:** prepare your corpus to be analyzed
- **Visualize:** create visualizations of patterns across your corpus or in one text
- **Analyze:** analyze your corpus, including comparing texts



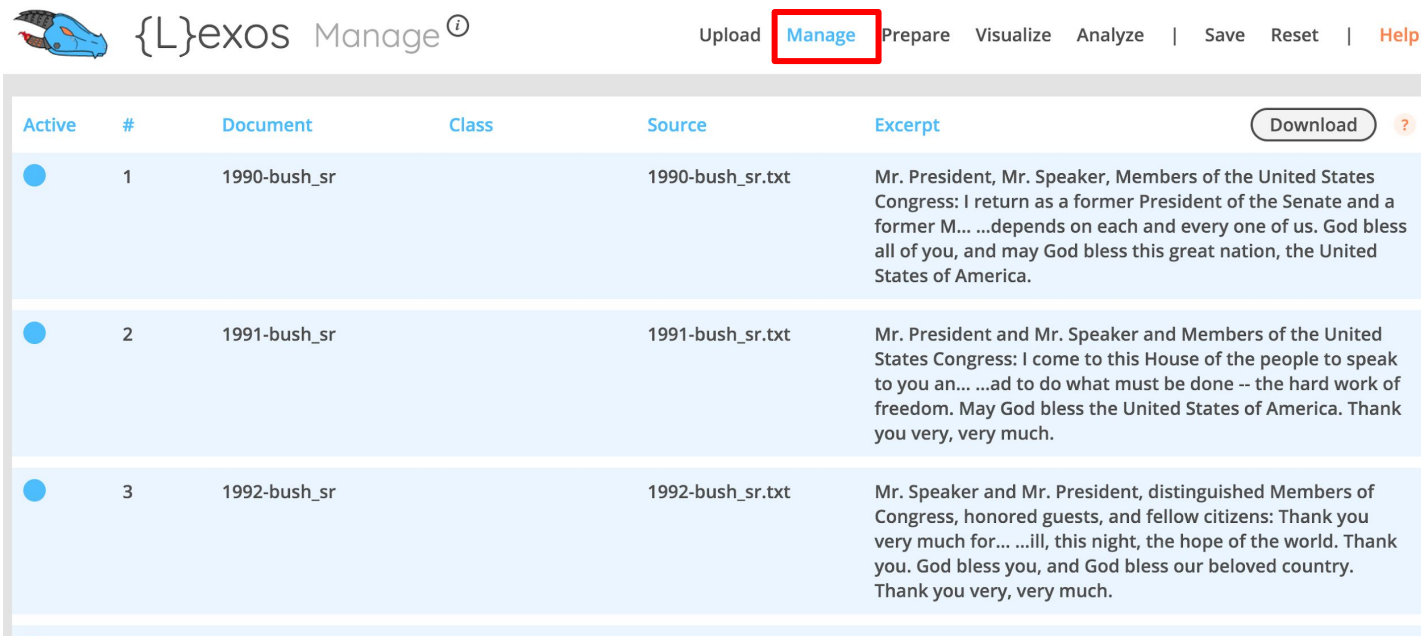
Lexos: Upload

Click browse
and select your
entire corpus
(or drag and
drop)



Lexos: Manage

Make sure all the documents in your corpus you want to use are selected (blue = selected, gray = not selected)



The image shows the Lexos Manage interface. At the top, there is a navigation bar with the Lexos logo (a blue dragon head) and the text "{L}exos Manage". To the right of the logo are several 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", "Excerpt", and a "Download" button with a help icon. The table contains three rows of data, all of which are selected (indicated by blue circles in the "Active" column).

Active	#	Document	Class	Source	Excerpt	Download
<input checked="" type="radio"/>	1	1990-bush_sr		1990-bush_sr.txt	Mr. President, Mr. Speaker, Members of the United States Congress: I return as a former President of the Senate and a former M... ...depends on each and every one of us. God bless all of you, and may God bless this great nation, the United States of America.	
<input checked="" type="radio"/>	2	1991-bush_sr		1991-bush_sr.txt	Mr. President and Mr. Speaker and Members of the United States Congress: I come to this House of the people to speak to you an... ...ad to do what must be done -- the hard work of freedom. May God bless the United States of America. Thank you very, very much.	
<input checked="" type="radio"/>	3	1992-bush_sr		1992-bush_sr.txt	Mr. Speaker and Mr. President, distinguished Members of Congress, honored guests, and fellow citizens: Thank you very much for... ...ill, this night, the hope of the world. Thank you. God bless you, and God bless our beloved country. Thank you very, very much.	



Lexos: Prepare (scrub)

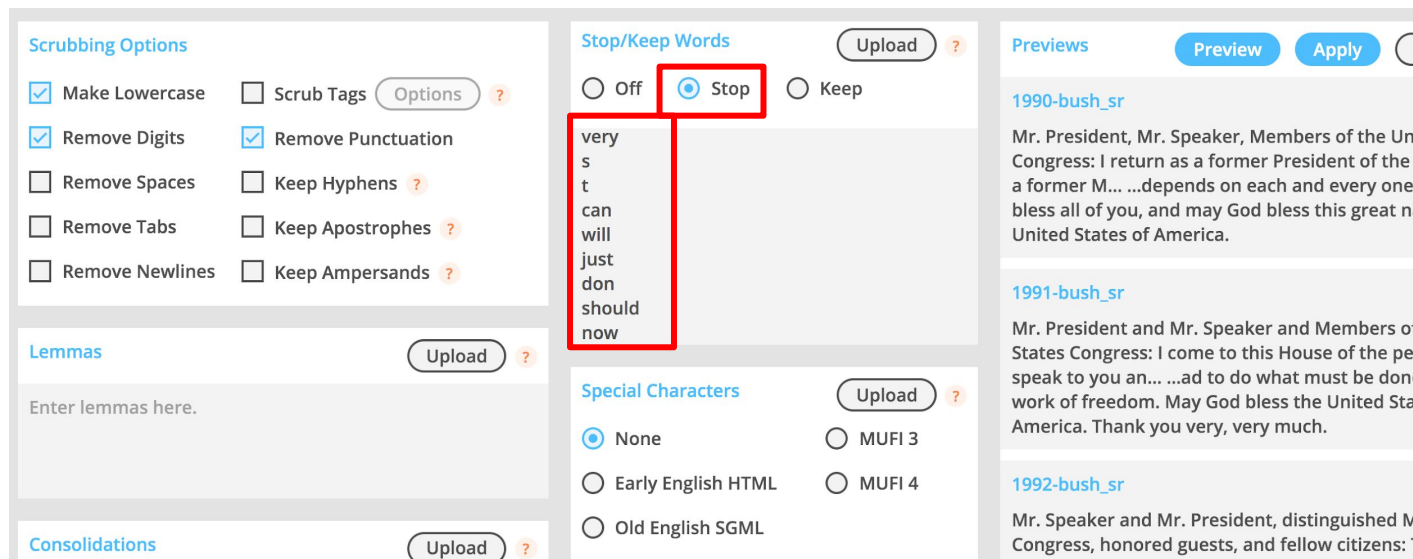
Lexos demonstrates the different choices you can make to prepare your corpus. By “scrubbing” you are transforming the text in your corpus and making choices that will impact your results.

- **Make lowercase:** make all your letters lowercase. Even though you know “A” and “a” are the same, the computer recognizes these as two separate characters. Lowercasing removes this.
- **Remove punctuation:** removes 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 the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to “talk”



Lexos: Removing Stop Words

Get a list of English stopwords here: <https://gist.github.com/sebleier/554280> (we also sent you a .txt file). Copy and paste the stopwords (or upload the .txt file) into the “Stop/Keep Words” box then click “Stop”



The screenshot displays 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.
- Lemmas:** A section with a text input field labeled 'Enter lemmas here.' and an 'Upload' button.
- Consolidations:** A section with an 'Upload' button.
- Stop/Keep Words:** This section is highlighted with a red box. It contains radio buttons for 'Off', 'Stop' (which is selected), and 'Keep'. Below these is a text input field containing a list of stopwords: 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', and 'now'. There is an 'Upload' button with a question mark icon.
- Special Characters:** Includes radio buttons for 'None' (selected), 'Early English HTML', and 'Old English SGML', along with 'MUFI 3' and 'MUFI 4' options. There is an 'Upload' button with a question mark icon.
- Previews:** A section on the right showing three preview cards for '1990-bush_sr', '1991-bush_sr', and '1992-bush_sr', each with a 'Preview' and 'Apply' button.



Lexos: Applying your Preparations

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

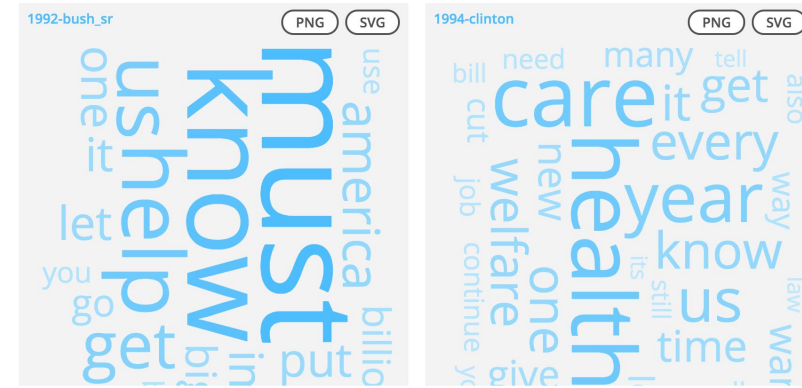
BEFORE PREP	AFTER PREP
<p>Previews Preview Apply Download</p> <p>1990-bush_sr</p> <p>Mr. President, Mr. Speaker, Members of the United States Congress: I return as a former President of the Senate and a former M... ..depends on each and every one of us. God bless all of you, and may God bless this great nation, the United States of America.</p> <p>1991-bush_sr</p> <p>Mr. President and Mr. Speaker and Members of the United States Congress: I come to this House of the people to speak to you an... ..ad to do what must be done -- the hard work of freedom. May God bless the United States of America. Thank you very, very much.</p>	<p>Previews Preview Apply Download</p> <p>1990-bush_sr</p> <p>mr president mr speaker members united states congress return former president senate former member great house now president... ..a call america let us remember state union depends every one us god bless you may god bless great nation united states america</p> <p>1991-bush_sr</p> <p>mr president mr speaker members united states congress come house people speak americans certain stand defining hour halfway a... ..toward next century confident ever home abroad must done hard work freedom may god bless united states america thank very much</p>



Northeastern University
NULab for Texts, Maps, and Networks

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

Word Cloud:
visualize a
wordcloud across
the entire corpus.



Lexos: Rolling Window

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

1. Go to “Manage” and right click one blue dot. Click “Deactivate all”
2. Choose the one **document** you want to analyze with the rolling window
3. Go back to “Visualize-> Rolling Window” and type in a search term you want to visualize. You can also do multiple by clicking “String” and separating words with a comma (president,health,repulican)
4. 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 a window size until you get a visualization that makes sense.
5. Click “Generate”



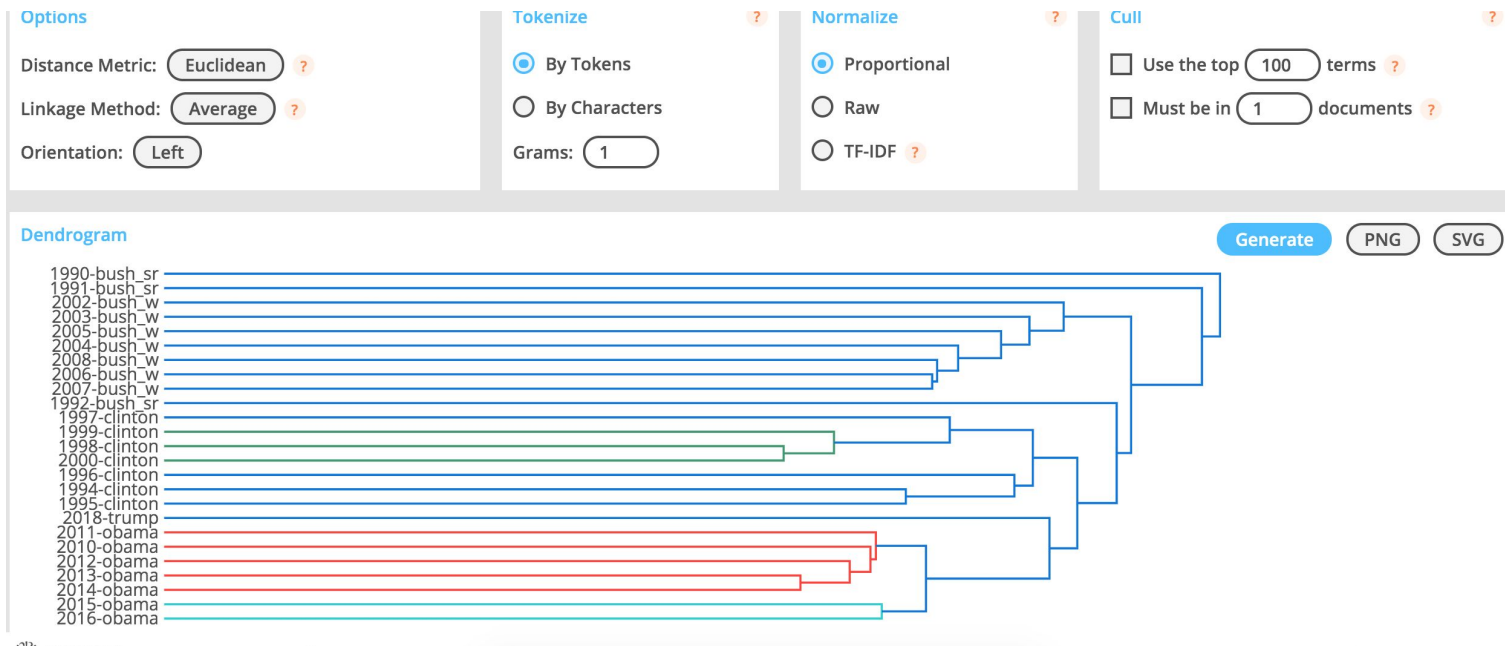
Lexos: Rolling Window Results

Using Trump's State of the Union from 2017, search strings such as “border,job,jobs”, and a window of 300, we can get an idea of how different terms work together in Trump's speech. You may also be interested in **contrasting** terms to see how they're used across the speech.



Lexos: Analyze, Dendrogram

The dendrogram demonstrates similarity between the different documents and connects the documents.



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 anytime 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 from to be analyzed.



VOYANT

see through your text

Click on upload and navigate to the folder with the text documents you wish to analyze.

Alternatively, insert URLs or full text into 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

ask questions at any point during the presentation!

Results:

From a corpus of political party platforms you can see the default results page with multiple panes:

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

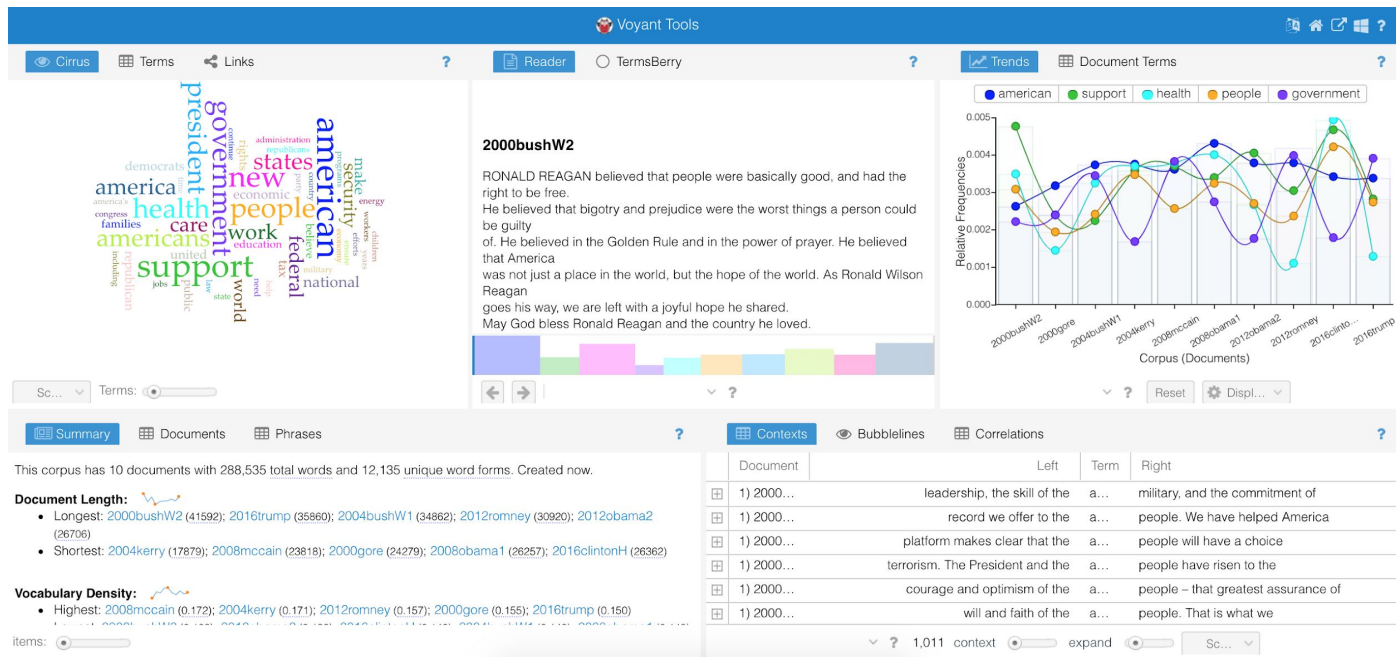
These boxes can all be changed!

Results:

From a corpus of political party platforms you can see the default results page with multiple panes:

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

These boxes can all be changed!



Voyant: Contexts (concordances)

Contexts:

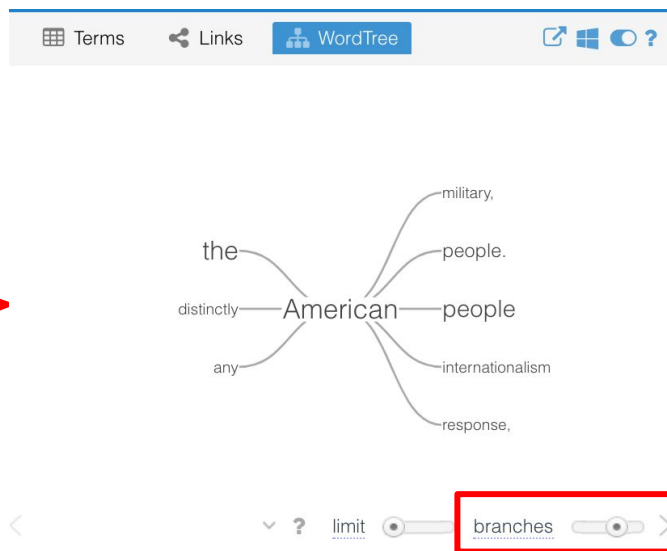
Contexts, or concordances, looks at the different contexts around particular search terms. For example, you can see all the times the word “president” appears in the corpus, which documents it appears in, and the contexts in which it appears.

<div>Contexts</div> <div>Bubblelines</div> <div>Correlations</div> <div>?</div>				
	Document	Left	Term	Right
+	1) 1990-...	you and to the American	people	about the state of the
+	1) 1990-...	or oppression for millions of	people	around the world. Nineteen forty
+	1) 1990-...	year -- one year ago, the	people	of Panama lived in fear
+	1) 1990-...	held hopes of the Americ...	people	; events that validate the long...
+	1) 1990-...	alive in the minds of	people	everywhere. As this new world
+	1) 1990-...	know this about the Ame...	people	: We welcome competition. W...
+	1) 1990-...	of the Congress: The Am...	people	did not send us here



Voyant: Changing displayed results

Select the panes button and select 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!

Take some time to explore Lexos and Voyant with your corpora.

- Prep your corpora using Lexos. Which preparation steps did you choose and why?
- Use Lexos' Rolling Window to visualize how particular terms are used across a document. What did you find? What do the different linguistic uses across the document tell you about the document?
- Use Voyant to explore contexts of a particular search term.

Find slides, handout, and data at <https://bit.ly/diti-fall2019-denver>



Thank you!

If you have any questions, contact us at:

Cara Marta Messina

Digital Integration Teaching Initiative

Assistant Director

messina.c@husky.neu.edu

Garrett Morrow

Digital Integration Teaching Initiative

Research Fellow

morrow.g@husky.neu.edu

Slides, handouts, and data available at <https://bit.ly/diti-fall2019-denver>

DITI Office Hours: Tuesdays, 1–3PM in 409 Nightingale Hall



Northeastern University

NULab for Texts, Maps, and Networks

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