Computational Text Analysis: Tools, Tricks, and Methods for Large Corpus Text Analysis

Cara Marta Messina, Laura Johnson, and Jeff Sternberg
Dr. Kirsten Rodine-Hardy

Fall 2019

Northeastern University
NULab for Texts, Maps, and Networks

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

Tell me your ads!

Thanks to our activities on our web-browsers, phones, social media accounts, and other forms of digital interaction, companies **collect**, **analyze**, and **sell** our data to make more targeted advertisements.

- What are some ads you've seen recently while you were on Instagram, YouTube, Facebook, etc.?
- Why do you think you're seeing these ads?

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

Digital Data: What Can We Learn?

One form of **big data** is the collection and analysis of mass amounts of data from a large amount of people to determine patterns of behavior and assess:

- Advertising preferences
- Risk assessment for insurance
- Credit scores
- Whether someone is a “good” citizen

Workshop Agenda

- Computational text analysis project and introduction
- Our corpus: National political party statements
- Demonstration of web-based text analysis tools
- Your turn!

Slides, handouts, and data available at

http://bit.ly/dti-fall2019-KRH

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

Workshop Objectives

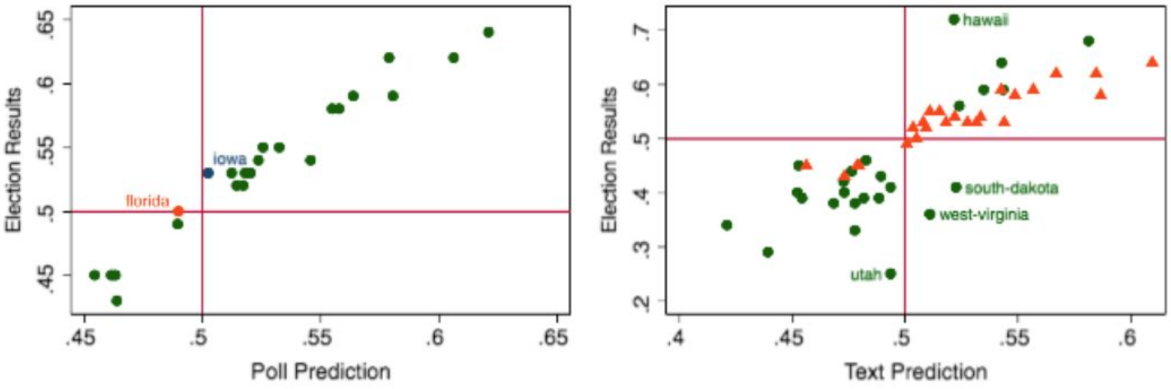
- Understand best practices for collecting and storing textual data when performing basic computational text analysis
- Understanding how web-based computational text analysis programs work, such as in their behind-the-scenes data preparation
- Understand how to interpret the results from your text analysis

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

Example

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science, 61*(2), 490-503.

FIGURE 2 **Polls at 11/4/12 vs. Election Results (Left) and Pure Text-Based Prediction on 11/4/12 from M1 (Right)**



Note: Triangles are training states; circles are other states.

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.

Why Computational Text Analysis?

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

Political scientists and politicians care **deeply** about the language used in political 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 national political party platforms written by Democrats and Republicans during presidential nominations in previous elections. Our files are a series of plain text (.txt) files.

Download this corpus from our email or from the "data" folder: **http://bit.ly/dti-fall2019-KRH**

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

Text Analysis Tools Links

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

Voyant https://voyant-tools.org/

SameDiff https://databasic.io/en/samediff/

Story Bench Sentiment Analysis
https://storybench.shinyapps.io/textanalysis/

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

SameDiff: https://databasic.io/en/samediff/

SameDiff compares **two texts**—only .txt files—and displays the words that appear in both texts and words that only appear in each individual text.

Let's compare the party platforms from the 2016 election (2016clintonH.txt and 2016trump.txt).

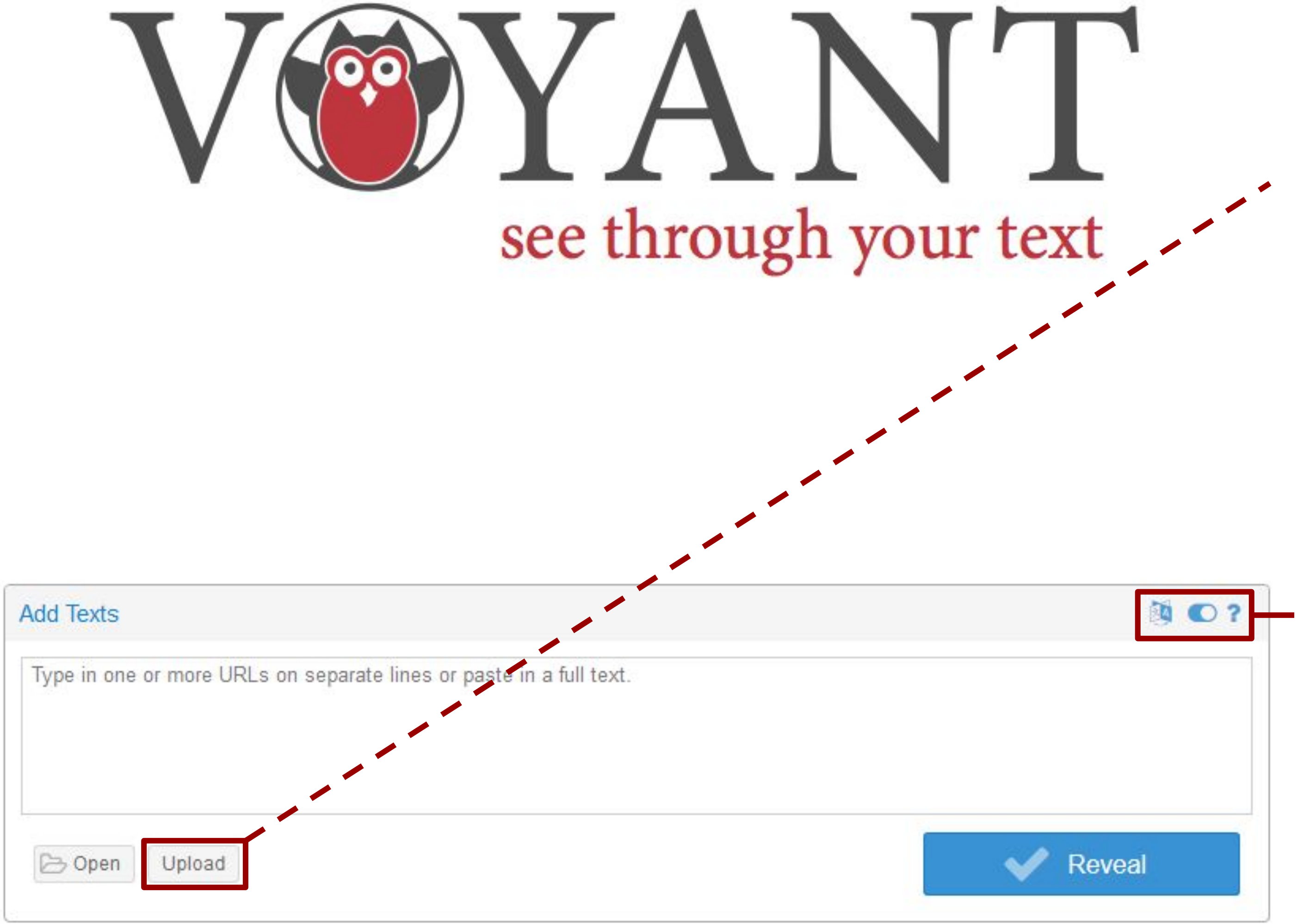
What do we notice? How can we interpret our results?

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), and word frequency distributions.

Click "Upload" and choose all the texts from 2000–2016 to be analyzed.

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



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 documents you wish to analyze.

Alternatively, insert URLs or full text into textbox.

Click here for help and advanced options

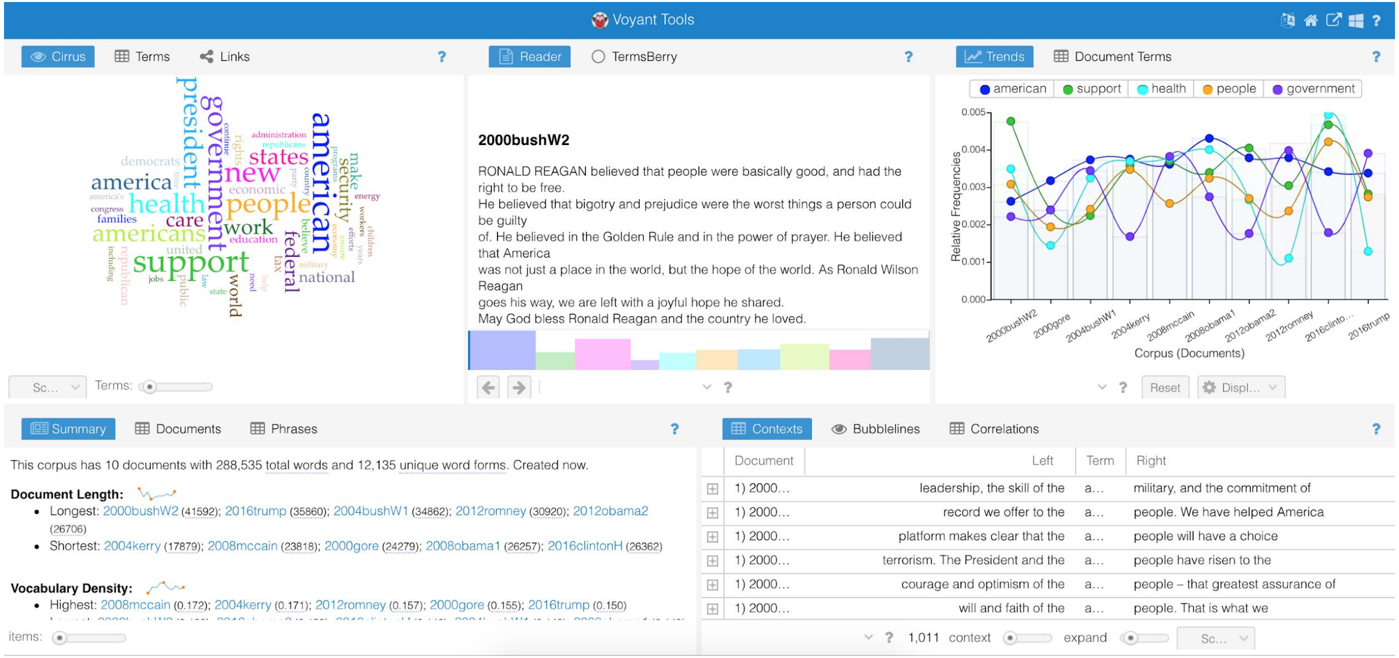
Voyant: Understanding the Dashboard

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!



2000bushW2

RONALD REAGAN believed that people were basically good, and had the right to be free.
He believed that bigotry and prejudice were the worst things a person could be guilty
of. He believed in the Golden Rule and in the power of prayer. He believed that America
was not just a place in the world, but the hope of the world. As Ronald Wilson Reagan
goes his way, we are left with a joyful hope he shared.
May God bless Ronald Reagan and the country he loved.

This corpus has 10 documents with 288,535 total words and 12,135 unique word forms. Created now.

Document Length:

- Longest: 2000bushW2 (41592); 2016trump (35860); 2004bushW1 (34862); 2012romney (30920); 2012obama2 (26706)
- Shortest: 2004kerry (17879); 2008mccain (23818); 2000gore (24279); 2008obama1 (26257); 2016clintonH (26362)

Vocabulary Density:

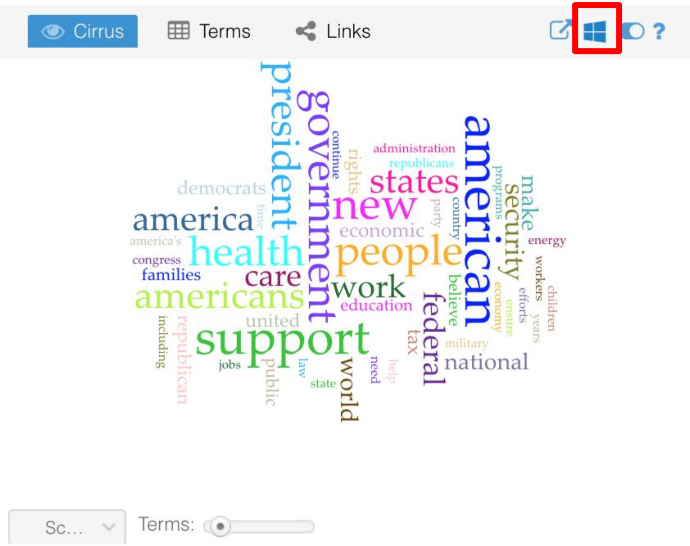
- Highest: 2008mccain (0.172); 2004kerry (0.171); 2012romney (0.157); 2000gore (0.155); 2016trump (0.150)

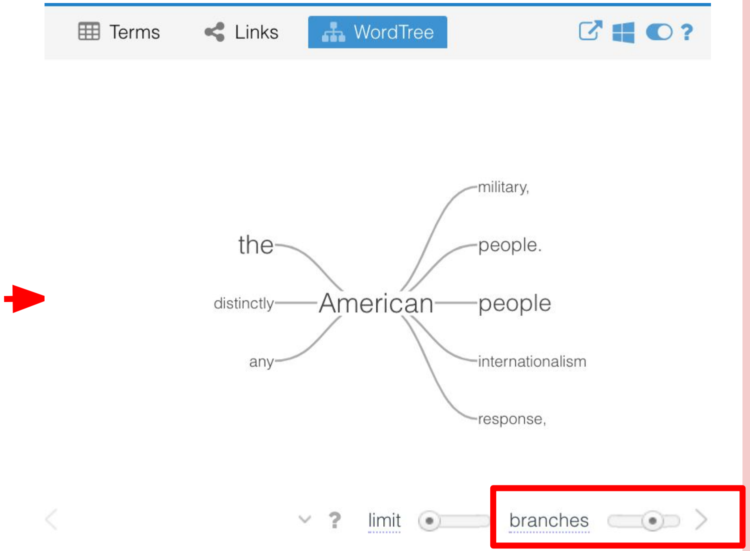
	Document	Left	Term	Right
⊞	1) 2000…	leadership, the skill of the	a…	military, and the commitment of
⊞	1) 2000…	record we offer to the	a…	people. We have helped America
⊞	1) 2000…	platform makes clear that the	a…	people will have a choice
⊞	1) 2000…	terrorism. The President and the	a…	people have risen to the
⊞	1) 2000…	courage and optimism of the	a…	people – that greatest assurance of
⊞	1) 2000…	will and faith of the	a…	people. That is what we

1,011 context expand

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.

Storybench Sentiment Analysis

https://storybench.shinyapps.io/textanalysis/

This tool only works with **one text** (it works with CSVs as well as .txt files), but provides a sentimental analysis of that text as well as nGram results.

What can we interpret about the 2012 Republican party platform (2012romney.txt) based on the results of the sentiment analysis?

Content warning: there are painful words in the results.

Your Turn!

Take some time to explore the different tools we showed you.

- What do you find challenging or exciting about these tools?
- What results come up?
- How might you interpret those results based on what you know about the contexts surrounding those texts and how language is used to persuade audiences?

Find slides, handout, and data at **http://bit.ly/dti-fall2019-KRH**

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

Thank you!

If you have any questions, contact us at:

Cara Marta Messina
Digital Teaching Integration
Assistant Director
messina.c@husky.neu.edu

Laura Johnson
Digital Teaching Integration
NULab Coordinator
johnson.lau@husky.neu.edu

Jeff Sternberg
Digital Teaching Integration
NULab Research Fellow
sternberg.je@husky.neu.edu

Slides, handouts, and data available at **http://bit.ly/dti-fall2019-KRH**

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

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