Political Science
Kirsten Rodine-Hardy
Introduction to Computational Text Analysis

Key Words

- **Computational Text Analysis**: Text analysis is making inferences based on textual data. Computational text analysis (CTA) involves a computer drawing out patterns in a text, and then the researcher interprets those patterns. CTA includes methods such as word count frequency, nGrams, and sentiment analysis. CTA is similar to statistical analysis, but the data are texts.
- **Corpus (plural–corpora)**: A collection of texts or data used for analysis and research purposes.
- **Stop words**: words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for basic computational analysis. Some English stopwords include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
    - In some projects, however, these words are included because stop words have been found to identify writers. For example, when J.K. Rowling published under a pseudonym, researchers were able to determine her identity based on her use of stopwords.
- **Word Count Frequency:** Counting the total times a word appears in a text or corpus or the percentage of how often it appears.
- **Sentiment Analysis**: Measuring the sentiment of a text based on a binary scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.

Our Corpus

The data are several national political party platforms written by Democrats and Republicans during presidential nominations. These platforms are attempts for each party to share the issues that party cares about and how their party plans to tackle these issues; they are written to persuade voters to elect their presidential candidate.

- Data retrieved from https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/national-political-party-platforms
- Data available at **http://bit.ly/dti-fall2019-KRH** under the "data" file

Building a Corpus

When building a corpus, especially in the context of a smaller project, follow these steps:

Questions? Contact us!
Cara Marta Messina, messina.c@husky.neu.edu
Laura Johnson, johnson.lau@husky.neu.edu
Jeff Sternberg, sternberg.je@husky.neu.edu

- Choose the texts you would like to include in your corpus.
    - Remember, these texts are not necessarily representative of a larger body of writing. Also, in your argument and analysis, you want to specifically address and analyze the *context* of these texts.
- Create a folder on your computer or cloud storage where you will store that corpus.
- Open a text editor (for example, Notepad on PCs and TextEdit on Macs) and make *one* file for *each* text. For example, if you have five texts you are working with in your corpus, create five files.
    - Make sure each file name ends with .txt – this is a plain text file and most web-browser tools will accept these.
    - Use filenames to indicate the data inside (ex: '2012obama.txt')
- Copy and paste the text into your text editor. The text editor will remove any pesky formatting (line breaks, bold, italics, etc) that could get in the way of the computational analysis.

Web-Browser Computational Text Analysis Tools

These browser GUI (Graphical User Interface) text analysis tools are not particularly powerful or sophisticated , but they can show word frequencies and patterns in language. While using coding languages like Python and R can open up other types of analysis (such as word embedding models and topic modeling), these GUI tools allow you to do more basic analysis to help begin examining your texts computationally.

We will be working with and trying out several web-browser computational text analysis programs.

- **Voyant**: https://voyant-tools.org/
- **DataBasic.io**: https://databasic.io/en/
- **SameDiff** (part of DataBasic): https://databasic.io/en/samediff/
- **WordTree:** https://www.jasondavies.com/wordtree/
- **Story Bench Sentiment Analysis:** https://storybench.shinyapps.io/textanalysis/

Questions? Contact us!
Cara Marta Messina, messina.c@husky.neu.edu
Laura Johnson, johnson.lau@husky.neu.edu
Jeff Sternberg, sternberg.je@husky.neu.edu