Created by Cara Marta Messina
Digital Teaching Integration
Assistant Director, 2019–2020
Northeastern University College of Social Sciences and Humanities

Video Link: https://youtu.be/4PTI1jmt25M

Use these times to follow particular information:

- Introduction [0:00]
- Important vocabulary [0:40]
- Sample corpus information [4:17]
- Voyant [6:04]
- DataBasic [11:51]
- Jason Davies' Word Tree [15:27]
- Interpreting results [17:54]

Links to the web-browser text analysis tools:

- Voyant: https://voyant-tools.org/
- Data Basic: https://databasic.io/en/
- Jason Davies' Word Tree: https://www.jasondavies.com/wordtree/

Introduction

This video will go over the basics for using web-browser text analysis tools to analyze a corpus of texts. This video is created by Northeastern University's Digital Teaching Integration, part of the College of Social Sciences and Humanities' initiative to better integrate digital proficiencies, skills, methods, and tools across the disciplines.

This video will:

- provide important vocabulary words and definitions
- briefly walk through the steps for building a corpus
- show how to use some web-browser text analysis tools
- finally provide questions to help you interpret results.

If you're interested in jumping to a particular section, you can find time stamps in the description box

Important vocabulary

First I will go over some important vocabulary words and terms.

- Computational text analysis: Computational text analysis uses computational methods and digital tools to provide information about patterns in a text. Some methods include word & phrase counts, collocation, and correlation.
- Web-browser text analysis tools: Tools that can be accessed as long as you have computer and internet access. These tools have limits for how much data they can handle, but are usually useful for small projects.
- Voyant: A web-browser text analysis and visualization tool that can analyze different texts and file types

    - For example, it can analyze .txt files, word document files, and PDFs. Just be careful, though, because some file types may have hidden formatting that could skew your results. I recommend almost always saving your texts as plain text files, or .txt files.
- Plain Text File: A plain text file, which ends in .txt, removes all hidden formatting. You can create .txt files by using plain text editors such as NotePad on Windows or TextEdit on Macs. I recommend using plain text files when doing computational text analysis, especially when using web-browser tools. While Voyant may accept Word Doc files, other web-browser tools may not.
- Corpus/corpora: a collection of texts that you will be analyzing. For this video, I will be using a sample corpus I created based on Democratic and Republican Party Platforms from the past few years.
- Data about your data, or information about the individual texts. Metadata about texts may include:
    - Title of the text
    - Author of the text
    - Where the text was published
    - When the text was published
    - A URL of the place where you got the data originally, if you collected it from somewhere online
    - The name of the .txt file (so you can remember which metadata belongs with which text)
- Data Preparation: The steps you take before you input your data into a text analysis tool. Textual data contains letters, numbers, white space, characters, and punctuation. Most computational text analysis tools like Voyant will do data preparation work for you.
- Data preparation includes removing stopwords as well as getting rid of punctuation or unnecessary white space.
    - Stopwords are the most common words that appear in a language. For example, some English stopwords are: the, a, an, I, we, she, he, they, and, & or.
    - Stopwords are typically removed in the data preparation stage because they can clutter the results for word frequency counts.
    - Sometimes, though, a researcher can choose to keep stopwords because stopwords, like prepositions, may help us identify particular information about the author and the authors' patterns as they write.

Corpus

First, you will need to create your corpus. If you are analyzing several different texts, I recommend keeping each text in an individual file and organizing these files into one folder. In my corpus, I have 10 individual text files, and you can tell they're text files because they end with '.txt'

As you can see here, my corpus is a collection of the Democratic and Republican Party Platforms from different presidential election years. I collected these texts from *The American Presidency Project* archive (https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/national-political-party-platforms).

I simply copy and pasted each individual platform into individual plain text files and named the files so I could tell what information was in each text. So, when I see 2000gore, I know this was the 2000 Democratic Party Platform written when Al Gore was running for president.

I also recommend creating a separate document, like a spreadsheet, that contains metadata about the texts in your corpus, such as the author of the text, the title, the name of the file the text is stored in, and the site where you got the text from.

Voyant

Now that I have my corpus, I want to use Voyant–a web-browsed text analysis tool–to begin analyzing my text. I go to voyant-tools.org and click "Upload." Then I navigate to my folder that contains my corpus. Make sure to select only the texts that you want to analyze–for example, I don't really want my metadata file to influence my results–then you click "Open."

This is the Voyant dashboard. Voyant provides several different visualizations and ways to analyze a corpus, while also giving you interesting information about your corpus.

This is a word cloud. It shows you the most frequent words (with the stop words removed).you can hover over each word and see the actual frequency the words appear in your document. For example, America appears 1011, which is unsurprising.

The summary is a basic summary of your corpus. Because I have my corpus broken up into individual text documents, Voyant can tell me information about individual documents. For example, the longest document is the 2000 Republican Party platform, while the shortest is the 2004 Democratic Party Platform. It also tells you which documents have the largest average words per sentence, the most frequent words, and distinctive words for each document. These are the unique words that appear in each document. This may be a great place for you to begin to analyze particular individual patterns in each texts as well as larger patterns across the corpus.

The reader gives you the actual text. You can go through to different texts and hover over words to see the document frequency. Here we can see stopwords and punctuation are still included, as this is important to reading the whole texts.

One of my favorite tools is the trends visualization. This shows how often a particular word appears across the texts. You can see the names of the individual text files at the bottom and then the frequency each word appears. For example you can trace how often the word health appears in each document and make inferences about each partys investment at the time. This is why its important to know the context of your data though. When and how is the word "health" used? Let's say I'm also interested in which party cares about crime. I can type crime here and see. But now I may be interested in how crime is being used. Is it about criminal reform? Or about being tougher on crime?

Thankfully Voyant has a tool to help with that! Contexts, which show collections or the words that surround a particular word, can show you the contexts for particular words. As you can see, the

Democratic party in 2000 cards a LOT about crime, so you might go through the different contexts to get an understanding for why crime is such a concern for the 2000 Democratic party.

Voyant also has a bunch of other tools, which you can find more out on the documentations page. Simply go to the top right of the browser screen and click the four panel icon.

You can also save your results! Let's say you want to post your results on your website or save them for an essay. You can simply go to the visualization or results you're interested in saving and click the Export button. This provides you with a bunch of ways to export, including saving as an image or even just saving the data and results. For example, I can save this visualization as an image or an HTML snippet. I can also save these context results as a JSON file, or as data you can navigate

Data Basic.io
Data Basic.io has several different types of text analysis tools. Today, we will look at SameDiff and Word Counter. Both SameDiff and Word Counter only accept .txt files or text copy-and-pasted.

Word Counter
Voyant also has a word counting mechanism, as we saw with the word cloud, but I want to show you Word Counter anyway. Word counter counts the most frequent words used in the text. Unlike Voyant, though, you can only select *one* .txt file. Let's say I'm interested in looking at the 2000 Democratic Party Platform. I click "upload file" and find my text. Then, I can choose to keep stopwords as well as keep capitalization. For the sake of this tutorial, I will remove stopwords and have all the letters lowercase.

Here we see another wordcloud, only this is a word cloud of *just* the Democratic party platform. We can also look at the top words and the top bigrams and trigrams. If you're interested, by scrolling to the bottom of the document, you can download individual counts, such as the word counts or bigram count.

SameDiff
SameDiff compares **two** .txt files, looking at the most frequent words that appear across both documents and then looking at the unique words that appear in each document. Let's say I'm interested in comparing the political party platforms in 2012. I choose my files and click upload.

Once you click your results, you can see the cosine similarity, which tells you how similar your documents are based on the words used. We can also see the words that appear most frequently in both documents, such as "president" and "american." Then, we can see the unique words for each document. In the Democratic platform, the word "worked" comes in, which may be because Obama was already president, so the party was discussing the work he had already done. We also see this interesting 've' here, which is the second half of a conjugation, like "would've". This happens when you remove punctuation.

In the Republican party platform, we see healthcare is used uniquely, which may be because Republicans were challenging Obamacare.

You can also save these results by scrolling down to the bottom.

Word Tree

Next we'll look at Jason Davies' WordTree. Voyant has a similar function, but this tool is created for the sole purpose of finding contextual patterns of a particular word.

Unlike Voyant and the Data Basic tools, this Word Tree works by copy-and-pasting text into this text file. Let's say I'm looking at contextual patterns in the 2000 Democratic Party platform. I open my txt file and copy the text. Then, I paste it into the "Paste Text" box and click "Generate Word Tree." Unlike Voyant, Word Count, and SameDiff, there is no data preparation done to the text; the stopwords, captial letters, and punctuation are all there.

When we were looking at Voyant results, the word "crime" was used quite often in this text. Let's see what are some linguistic patterns before and after the word crime.

As we can see, "is" comes after crime several times. In fact, we see "crime is down" is used **four** times, which may be because Clinton, a Democrat, was president during this time, so the party may be arguing for why their platforms are more effective. Then we see "crime before it…" which may suggest the party was concerned with targeting the causes of crime.

If you're interested in seeing the patterns *before* the word crime, simply click "reverse tree." There are less patterns here, but we can see adjectives like violent and organized as well as verbs like stop.

Interpreting results

I have shown you a few different web-browser text analysis tools you can use to conduct computational text analysis on a corpus. One of the most important things to remember, though, is that getting these results isn't necessarily enough.

Computational text analysis can be a form of discovery, and provides another method for exploring your corpus rather than just reading the texts. This may be particularly important if you have a huge corpus, say a million words. computational text analysis may help you summarize patterns in your text. However, you don't want to stop there.

Whether you're doing a larger or a smaller scale project, like looking at the 10 Party platforms since 2000, I recommend you **interpret** those results and use them to present some kind of argument or idea about your corpus. For example, we may look at the way the presidential party platforms all talk about crime to get a better understanding of different values among the different parties as well as how those values may shift over time. If you compare how "crime" is used in the 2000s vs how "crime" is used in 2018, you may get a better sense of the values being perpetuated in different periods and by different parties. Let's say were talking about the 2000 Democratic party. The text analysis results suggested there was a focus on crime partially because crime rates dropped under the Clinton administration. What we

might be able to argue, though, are how parties use the current administrations' choices and impacts to advocate for their nominee. And sometimes how ideas of crime get perpetuated in those

When looking at results, there are a **ton** of questions you can ask. And your questions will be much more specific depending on your corpus, which is why it's *always* good to be familiar with your data. Here are some general questions you might find useful to apply to your corpus:

- Why do some words appear in texts but not appear in other texts?
- Why might some words appear more frequently than other words?
- How does word usage change over time?
- What values might be upheld in individual texts or across the corpus when you look at word usage? How might values differ between individual texts?

And remember, while computational text analysis quantifies information from your corpus, your corpus is just a small representation of a much larger wealth of information. Keep your arguments focused and contextualized. And how you interpret your results and make your argument is up to your framework and perspective. As always, be critical and mindful when you make arguments.

For more information, visit our website. The link is down in the description box