Data Analysis and Visualizations

Cara Marta Messina
NULab Coordinator
messina.c@husky.neu.edu

Northeastern University
NULab for Texts, Maps, and Networks

Northeastern University
NULab for Texts, Maps, and Networks

Workshop Objectives

By the end of this workshop, you should be able to:

- Identify data file formats & understand the differences between these
- Learn how digital tools can be useful to analyze data
- Determine what programs can work with what data file types
- Use basic web-based text analysis tools
- Think about how to incorporate this knowledge as you move forward analyzing your reflections using these tools

Northeastern University
NULab for Texts, Maps, and Networks

Review from last class

Last time I came in, we discussed different data types. Open the zip file titled "files_for_FYW" and find the data types we looked at last class in "files_for_FYW_class1."

Let's review!

1. What are the different file types? Describe similarities and differences.
2. How do you recognize these file types?
3. Why is it important to know your file types?

Northeastern University
NULab for Texts, Maps, and Networks

Know Your Data

This project invites you to **create your own data**. The reflections you have been composing in Excel is now the data that you will work with and analyze.

So what?

A large portion of research is thinking through your data collection, storage, manipulation, and analysis. How will you collect your data? How will you organize it? In what ways will you parse information from your data? What tools will you use to analyze your data?

Northeastern University
NULab for Texts, Maps, and Networks

Text Analysis (one form of data analysis)

At this point, most of you have been exposed to “close reading” as a type of text analysis. For this class and project, we will look at **computational** text analysis methods, such as:

- counting word frequencies
- nGrams
- Collocations
- unique words in texts
- linguistic patterns
- and more!

Northeastern University
NULab for Texts, Maps, and Networks

Structuring your data!

Now that you have your data in an Excel file, we can practice how to transfer information from one structure to another in order to make the text data easier to work with.

From Excel -> .txt files: copy text into .txt files. Your choices when copying these .txt files depends on how you want to analyze them.

When creating multiple files, always think about the **NAMES** of your files. How will names help you remember what is in the files? How will naming keep you organized?

Example: Individual .txt files

In the “class2” folder, there are a series of .txt files all with different names. By looking at the names of these files, can anyone guess what is in the content of each .txt file?

- “All” means all of the textual data, while “all_Q1/Q2” means all of the text from reflection question 1 and then all the text from question 2
- If the file begins with “Q1”, that is text from question 1. The “name” also signifies the name of the student about whom I was writing.
- If the file begins with “Q2”, that is text from question 2. The “name” also signifies the name of the student about whom I was writing.

Why are naming conventions and choosing organizational structures important?

Now when you begin to use your tools for analysis, you can try different approaches! For example, you might want to compare the words you used in your response to Q1 with your response to Q2. You can do this with some of the tools we will discuss next.

Northeastern University
NULab for Texts, Maps, and Networks

Let's play!

Using the .txt files in the "class2" folder, we're going to take some time to start working with different digital tools to help us begin analyzing our textual data. The platforms we will be using are:

- **Voyant**
- **SameDiff**
- **WordTree**
- Excel/Google Sheets
- **Story Bench Sentiment Analysis**–sentiment analysis measures how "positive" or "negative" a text is (we will not go into using this, but this is an awesome tool)
    - Text files
    - CSV files

Northeastern University
NULab for Texts, Maps, and Networks

About Voyant

https://voyant-tools.org/

Voyant can look at one **OR** multiple text files (and recognize them as different files). It can read .pdfs and .docx, although I always recommend using .txt files because it removes messy formatting.

Highlight the .txt files that begin with “Q1” and “Q2” and drop those into Voyant! What do we see? What might this tell us about our data?

FYI: a lot of the Voyant tool features–and a lot of the tools we will use today–remove “stopwords”, or the most popular words used in English (the, a, she, her, of, or in, and, etc..). This is sometimes referred to as “cleaning” data.

Northeastern University
NULab for Texts, Maps, and Networks

Popular Voyant Features

Wordcloud: most frequent words will appear here; the largest words are the most frequent, while smaller words are still frequent, but a bit less

Phrases: the frequency of several words that appear in a row (also called ‘nGrams’)

Contexts: the string of words that appear around one word (also called ‘colocation’)

Correlation: words that appear in similar contexts

For more infromation about all the tools, visit Voyant’s Tool Index:
http://docs.voyant-tools.org/tools/

Northeastern University
NULab for Texts, Maps, and Networks

Same Diff

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

Same Diff compares the unique and similar words used between two texts. Similar to Voyant, stopwords are removed.

In order to use SameDiff, you must have **two .txt** files. For example, what are similar and unique words between “all_Q1” and “all_Q2”? What might this tell us about our data?

Northeastern University
NULab for Texts, Maps, and Networks

Word Tree

https://www.jasondavies.com/wordtree/

Copy and paste the text (or texts) you would like to explore. Word Tree shows linguistic pattern frequencies that appear surrounding a word.

For example, in the pres texts I have provided, copy and paste “all” in. What happens when you search the word “talk?” You can either look at the patterns of words that come before (by clicking ‘reverse tree’) or after the word people.

Excel/Google Sheets

Excel/Google Sheets is great at organizing, counting, and visualizing. We talked a bit about this last time, but we can quickly go over charts and pivot tables.

Creating Charts

https://support.office.com/en-us/article/video-create-a-chart-4d95c6a5-42d2-4cfc-aede-0ebf01d409a8

Pivot Tables

https://support.office.com/en-us/article/Create-a-PivotTable-to-analyze-worksheet-data-A9A84538-BFE9-40A9-A8E9-F99134456576

Your Turn!

For the rest of class, please take some time to organizing your data or using some of the tools we went over today. I highly suggesting organizing your data first, but I also understand wanting to dive right in. I will be here if you have questions!

Northeastern University
NULab for Texts, Maps, and Networks

Contact

Cara Marta Messina

PhD Candidate, Writing and Rhetoric

NULab Coordinator

messina.c@husky.neu.edu

@cara_messina