


Writing Studies Research: Working with Data

Cara Marta Messina
Digital Teaching Integration
Prof. Neal Lerner



Workshop Objectives

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

- Learn how digital tools can be useful to analyze data
- Identify data file formats
- Understand the differences between data file formats
- 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 with your group work

Open files

Prof Lerner will have sent you a zip file titled
“files_for_lerner.”

When you unzip this file, you will find two different folders:
“practice_data” and “sample_types.” Open “**sample_types**”

Trivia: Question 1

In the “sample_types” folder, take a few minutes to go through the files. What are the differences between them, especially for how the data is presented and organized?

- “csv.csv” is a csv-or comma separated value-file
- “text” is a .txt file, which is unformatted
- “doc” is a .docx file and is formatted
- “excel.xlsx” is an Excel file

Trivia: Question 2

What are some indicators of the file type?

- The letters after the period (.csv, .txt, .doc, etc)
- The icon (a Word Doc icon looks like versus the Excel icon)
- What programs can open the files
- What the data actually looks like when you open the file

Trivia: Question 3

Why is it important to know your file types?

- To figure out what files can open with what programs. For example, you can open a .txt file with Word, but you cannot open a .doc with a Text Editor. You can open an .xlsx with Excel), but not a text editor; you can open a .csv file in both Excel and in a text editor!
- To decide **methods** to analyze and manipulate that data

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?

Example: Writing Program Committee Reflection

The Writing Program Committee is currently studying student reflections. One of the goals is to articulate genre conventions that make for successful reflections.

I am currently using text analysis methods on over 300 student-written reflections. In order to do this work, though, I had to know a lot about file structures and how to re-organize data for it to be analyzed.

Using Python, I was able to reformat the data into a structure that was easier for Python to analyze.

Case Study: Basic Methods

Using Python, I have conducted several different types of text analysis so far:

- Natural Language Processing
- Topic Modeling
- Word Frequency
- nGrams

Don't worry if these large corpora text analysis methods are unfamiliar! We will talk about some of them later today.

Case Study: One Interesting Preliminary Result

Natural Language Processing (parts of speech analysis) showed that reflections with higher scores used nouns like “genre,” “audience,” and “review,” while reflections with lower scores focused on the content of the essay. This suggests that reflections rated strongly demonstrated the student writer thinking through writing process, while reflections that were rated less strongly showed the student thinking more about the *content* that they learned, rather than what they learned through the process.

Large Discussion

Let's talk a bit about the data you are collecting and working with!

Since not every group is working directly with data, let's have a large group discussion about the different data types you have run into so far with your data. What have you collected so far? What is the data type? How is your data formatted? What information is in your data?

The Data Provided

I am giving you two groups of data, which can be called datasets or corpora:

- Marvel and DC comics Wiki data about the characters in the universes. Both sets of data are .csv files and begin with “comics”
 - Downloaded from Kaggle (<https://www.kaggle.com>), an opensource place to download free data.
- National Political Platforms from the 2012 and 2016 elections; four .txt files that begin with “pres.” These platforms are written by each party while presidential candidates are campaigning to show what the party values and convince voters why they should choose a particular side.
 - Copy and pasted into .txt files from The American Presidency Project (<https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/national-political-party-platforms>)

Let's play!

Using the data I sent you all, we're going to take some time to start working with different digital tools to help us begin analyzing our “**sample_data**”. The platforms we will be using are:

- [Voyant](#)
- [SameDiff](#)
- [WordTree](#)
- Excel/Google Sheets
- Story Bench Sentiment Analysis
 - [Text files](#)
 - [CSV files](#)

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.

In the “practice data” folder, you will see several .txt files with the label “pres” at the beginning. Highlight these and drop them into Voyant.

FYI: a lot of the Voyant tool features 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.

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 information about all the tools, visit Voyant's Tool Index:

<http://docs.voyant-tools.org/tools/>

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 from the 2016 election?

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 “pres_2012obama.txt” in. What happens when you search the word “people?” You can either look at the patterns of words that come before (by clicking ‘reverse tree’) or after the word people.

Excel/Google Sheets

Open the two “comics” .csv files in Excel. Remember, .csv files can be opened and used in Excel, but are better for data analysis than .xlsx files because they have less hidden formatting *and* they are the standard data format used for research.

Excel/Google Sheets is **great** at organizing, counting, and visualizing. Let’s see what we can do with these “comics” datasets!

Thanks!

For the rest of the class time, I will be here to answer any questions you have as you work in your groups.

Contact

Cara Marta Messina

PhD Candidate, Writing
and Rhetoric

NULab Coordinator

messina.c@husky.neu.edu

@cara_messina