CRIM 3300: Punishment in the Age of Mass Incarceration
Megan Denver
Corpus Building

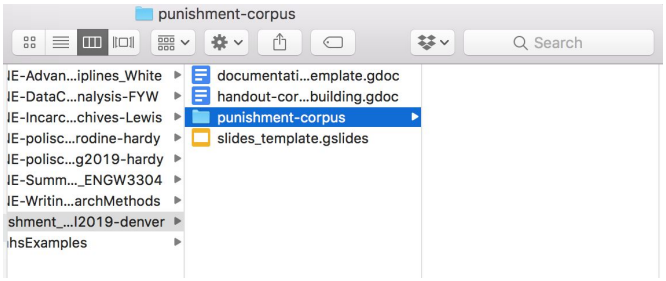
About

Before using text analysis tools, first you will need to build a corpus. A **corpus** (plural "corpora") is a collection of texts, and corpora are often created and collected for research. This handout will cover the basics of building a corpus for computational text analysis, or work with computational tools that "read" texts.

Steps for Building a Corpus

When building a corpus, especially in the context of a smaller project like the ones you will be conducting for this class, follow these steps:

1. Find and **choose** the texts you would like to include in your corpus
    a. There are many databases online that you can use to find texts. Some databases are available at https://web.northeastern.edu/nulab/resources/
    b. Remember, these texts are not necessarily representative of a larger body of writing, but rather a contextualized collection of texts.
    c. In your argument and analysis, you should specifically address and analyze the contexts of these texts.
2. Create a folder on your computer or cloud storage where you will store that corpus. As the screenshot demonstrates, I have created a folder titled "punishment-corpus"

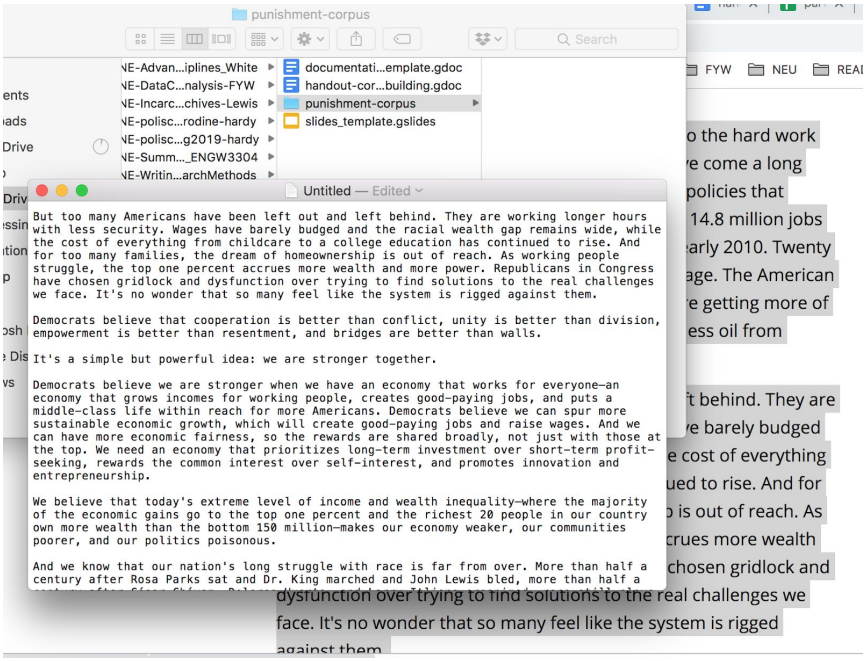


3. Once you have chosen the texts you will include, open a text editor (for example, Notepad on PCs and TextEdit on Macs)
4. To add the actual text, simply **copy and paste** the text into the text editor. Often, texts that are on websites can easily be copied and pasted. Only copy **one text** into

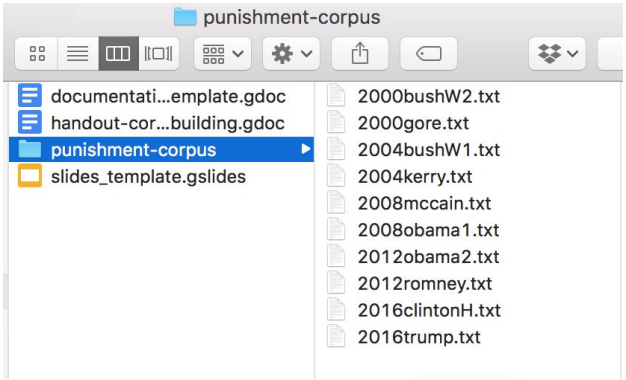
Questions? Contact us!
Cara Marta Messina, messina.c@husky.neu.edu
Garrett Morrow, morrow.g@husky.neu.edu

each new text editor file. As the screenshot below shows, I have copied the web browser text in the background, and copy-pasted it into a plain text editor.

a. If you cannot copy and paste the text (if it's a PDF or an image), either find a text that you *can* copy and paste, or transcribe the text.



5. Do steps 4 and 5 for each text in your corpus. See the screenshot below of my final corpus.
   a. For example, if you have five texts in your corpus, create five files.
   b. Make sure each file name ends with .txt – this is a plain text file and most web-browser tools will accept these.
   c. Use **filenames** to indicate the data inside (ex: '2012obama.txt')



Cara Marta Messina, messina.c@husky.neu.edu
Garrett Morrow, morrow.g@husky.neu.edu