ENGL 3370: Summer I
How to Create a Corpus: Tools and Methodologies

Key Words

- **Data:** the structured information that you will research. In this case, your data are the texts in your corpus and the metadata (the information about the texts)
- **Corpus (plural, corpora)**: the texts you have aggregated for your research
- **Computational Text Analysis:** One method for conducting research on a corpus. This method looks at linguistic and semantic patterns in the corpus.

The Basics of Conducting Research

1. Choose the **content** which you want to study.
2. Select a **theoretical framework**. Your framework will determine your research questions, your data selection, and how you analyze your data.
3. Develop **research questions** (2-3). These questions should be able to be answered/explored with your chosen content.
4. Choose the **research methods** for which you will tackle your research questions. One of the methods you will use will be computational text analysis, so one of your research questions should reflect an interest larger patterns in the corpus.
5. Start collecting your data. In the case for computational text analysis, you will **build a corpus**, or a group of texts.
6. **Analyze** your data. For this assignment, you will be analyzing the texts using computational text analysis means.
7. Present your **findings** and include background on your research process (ie: the steps above).

Questions to Consider

- Which texts can help me answer my larger research question?
- Whose voices do I want represented in my corpus?
- What biases might appear while crafting my corpus and how should I address them in my larger research project?

Creating a Corpus

- Choose your archive or source from where you will retrieve data.
- Begin choosing the texts that you will put in your corpus. The more texts you have, the more broad your computational text analysis results will be.

    - You may choose texts that **relate directly** to your research question.
    - You may also choose a **random assortment** of texts as a more exploratory and less biased process.
- In order to properly build a corpus, you have to **store your data** in an accessible and systematic way.

Storing Data

- Create a **folder** titled "corpus", "data", or whatever you choose
- Create an **individual .txt file** for each text. A .txt file is a plain format file; it's important to use .txt because certain programs will only accept .txt files, it standardizes the texts, and it removes any hidden formatting that could mess up your results.
    - Copy and paste the text into the .txt file
    - If you cannot copy and paste the text (photograph, header...etc) you might need to transcribe your text.
- Follow a **naming convention** to for each .txt file. This convention is based on your choice.
- Save **metadata** (data about your data) with the file name, author name, links, and other information. You can save it in a spreadsheet or create a new .txt file.

Helpful Resources

- DTI GitHub (the handouts and slides for the in-class modules can be found here)
    - https://github.com/NULabNortheastern/digitalassignmentshowcase
- "How to Build a Corpus", John Sinclair (2004)
    - https://ota.ox.ac.uk/documents/creating/dlc/appendix.htm

Corpora Examples

- Women Writers Online
    - https://www.wwp.northeastern.edu/wwo/
- Oceanic Exchanges
    - https://oceanicexchanges.org/

Contact
If you have questions, free to contact us:
Cara Marta Messina: messina.c@husky.neu.edu
Molly Nebiolo: nebiolo.m@husky.neu.edu