ENGL 3161 Composite Aesthetics: Race as Technology
Eunsong Kim
Corpus-building Handout for Computational Text Analysis

Key Words

- **Computational Text Analysis**: Text analysis is making inferences based on textual data. Computational text analysis (CTA) involves a computer drawing out patterns in a text, and a researcher interpreting those patterns. CTA includes methods such as word count frequency, nGrams, and sentiment analysis. CTA is similar to statistical analysis, but the data are texts.
- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stopwords include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.

Web-Browser Computational Text Analysis Tools

These browser-based GUI (Graphical User Interface) text analysis tools can show word frequencies and patterns in language. While using coding languages like Python and R can open up other types of analysis (such as word embedding models and topic modeling), these GUI tools allow you to do more basic analysis to begin examining your texts computationally. We will be working with:

- Lexos: http://lexos.wheatoncollege.edu/upload
- Voyant: https://voyant-tools.org/
- Word Counter: https://databasic.io/en/wordcounter/
- Word Tree: https://www.jasondavies.com/wordtree/

Our Sample Text

Data, *Dark Matters : On the Surveillance of Blackness*, Duke University Press, 2015. *ProQuest Ebook Central*, https://ebookcentral.proquest.com/lib/northeastern-ebooks/detail.action?docID=2194890.

The DITI has provided the .txt files you will use during this session. For future reference, we included detailed instructions and typical responsibilities for creating and cleaning a corpus. Feel free to practice when you are able.

How to Build a Corpus

When building a corpus, especially in the context of a smaller project, follow these steps:

1. Choose the texts you would like to include in your corpus.
    - Remember, these texts are not necessarily representative of a larger body of writing, and that the texts you select will have a significant impact on your results.
    - In any argument and analysis of your results, you should specifically address and analyze the contexts of these texts and consider any possible limitations in their ability to serve as proxies for the phenomena you wish to study.
2. Create a folder on your computer or cloud storage where you will store your corpus. You might title this folder "browne_corpus".
3. Once you have chosen the texts you will include, open a plain text editor (for example, Notepad on PCs and TextEdit on Macs).
    - TextEdit on Macs: You must make sure it is configured to work with plain text files. To do this, open Text Edit and go to "Preferences" and make sure "plain text editor" is selected. Then, restart TextEdit.
4. Making plain-text files:
    - The individual plain text (.txt) files that make up your corpus are stripped-down and machine readable versions of the documents (PDFs, .doc, .docx, etc.) you chose to include in your corpus.
    - To add the actual text, you would simply copy and paste the contents of each document into the text editor. Step-by-step instructions for this process are included in the slides.
        - Often, texts that are on websites can easily be copied and pasted; however, copying and pasting the text from documents can take a bit more time and require more extensive data cleaning.
        - **Only copy one text** into each new plain text file (unless you are combining texts from similar resources for research purposes).
    - Some articles might have HTML/web-browser versions that will be easier to copy-and-paste than PDFs.
    - If you cannot copy and paste the text (if it is a PDF or an image), either find a text that you can copy and paste, or transcribe the text.
    - Make sure each file name ends with .txt – this is a plain text file and most GUI tools will accept these.
    - Use filenames to indicate the data inside (ex: "2012obama.txt")
    - Make sure not to put any spaces in the names of the files as you save them. Use underscores or hyphens to mark spaces between words instead.
5. Repeat steps 4 and 5 for each text in your corpus.
    - For example, if you have five texts in your corpus, create five files.