SOCL4600 Senior Seminar
Professor Ineke Marshall
Text Analysis using the DataBasic Suite

DataBasic website: https://databasic.io/en/
Data Basic is an easy-to-use package of data analysis tools for beginners.
Data basic includes four tools:

1. WordCounter (text analysis)
2. WTFcsv (data visualization and analysis)
3. SameDiff (text analysis)
4. ConnectTheDots (network analysis)

Important Terminology

- Corpus/corpora: a text or collection of multiple texts that is used for analysis.
    - Example: One could could create corpus of all of Barack obama's presidential speeches to trace his language over the course of his presidency
- N-gram: A continuous sequence of n items in a text. For example, in Barack Obama's speeches, a bigram (or 2 continuous words) could be 'United States,' while a tripgram (3 words) could be 'yes we can.'
- Stopwords: commonly used words that are part of natural language usage, but typically do not add meaning to a sentence but can add context. Stopwords are commonly ignored in text analysis, but this depends on the questions being asked.
    - Examples: the, but, this, that

WordCounter website: https://databasic.io/en/wordcounter/

What is WordCounter?
WordCounter analyzes a corpus to count words and n-grams.

Word counts, bigram, and trigram data can then be download as a .csv for further analysis.

Step-by-step WordCounter guide:

1. To use your own text, select: paste text, upload a file, or paste a link.
    a. For paste text, copy text into the text box.
    b. For upload a file click on upload and navigate to the text on your local device.
    c. For pasting a link, copy the URL into the text box.

2. Click on count.
    a. Note that 'ignore case' and 'ignore stopwords' is selected by default.
3. WordCounter outputs as a word cloud and a list of top words, bigrams, and trigrams.
4. The researcher can output a .csv of top words, bigrams, and trigrams accessed by scrolling down the page below the word cloud and lists.

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

What is SameDiff?

SameDiff compares one corpus or text to another corpus or text and tells the user how similar they are based upon a cosine similarity algorithm.

Step-by-Step SameDiff Guide:

1. Select 'upload files.'
2. Click on browse file 1 and navigate to the first text.
3. Repeat step 2 for second text.
4. Click on compare.
5. SameDiff outputs a similarity score, total word counts, and the specific words that are similar and the words that differentiate the two documents.
6. The researcher can output a .csv of word counts accessed by scrolling down to below the results page.