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** (plural: corpora): a text or collection of texts that is used for analysis.
    - Example: One could create a corpus of all of Frederick Douglass's speeches to trace his language use over time.
- **nGram**: A continuous sequence of *n* items in a text.
    - For example, in Frederick Douglass's speeches, a bigram (or 2 continuous tokens) could be 'United States' and a trigram (3 tokens) could be 'justice for all'.
- **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.
- **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.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.

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 downloaded 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.
2. Click on count.
    a. Note that 'ignore case' and 'ignore stopwords' are selected by default.
    b. See how your results vary if you turn these off!
3. WordCounter outputs a word cloud and a list of top words, bigrams, and trigrams.
4. You can export files with the top words, bigrams, and trigrams by scrolling to the bottom of the page and selecting the export option you want.

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 the 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. You can scroll to the bottom to find options for exporting your results.