Coding Activity

Language, Power, Computation

College of the Atlantic. January 31, 2023

Figure 1: A unicorn. Just because. Image from `https://freesvg.org/1539642047`

1. Go to Project Gutenberg (`https://www.gutenberg.org/`) and explore choose two books (or whatever) that you'd like to analyze.
2. Start a new colab notebook (or make a copy of one you already have) and import the two books using the `wget` command. Choose the URL for the utf-8 text file.
3. Then do the standard things to your books: Clean up the files, tokenize, lemmatize, remove stopwords.
4. Optional: make a function that takes a filename as input and returns the cleaned file. By clean, I mean with numerals and punctuation and such removed. (This will save you considerable time, since you'll need to clean files many times. If you have a function that works, you can just use that function.)
5. Optional: write a function that takes a list of tokens as input and returns a list with the stopwords removed.
6. Optional: write a function that takes a list of tokens as input and returns a list of lemmatized tokens.
7. Create a counter object from a list of tokens.

   ```
   import collections
   Text_Counted = collections.Counter(Text)
   # Text_Counted is now a Counter object: basically a
   # list of word frequencies
   ```

8. I'll now show you how to make frequency plots, starting from two counters, where you can compare the frequencies of words in the two different texts.