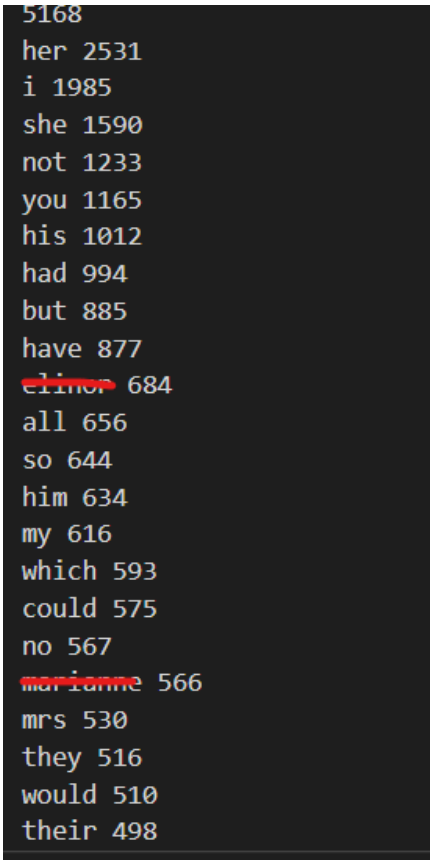
CS 446 Project 1 Report

Your Name

February 21, 2023

1 First Analysis question

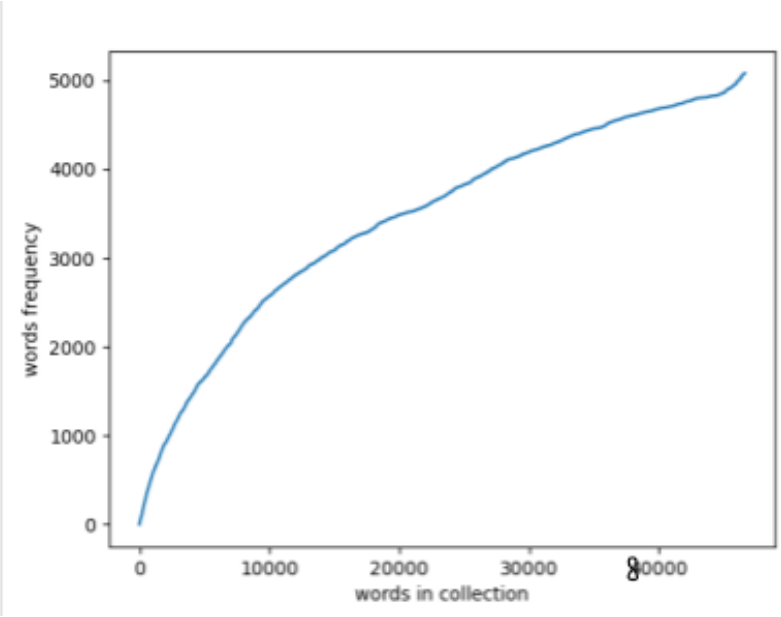
The top terms look like they are everyday words that aren't particular to the novel because words like 'her', 'i', 'she', 'you', 'had', 'but' are the top terms and they are generally used in almost every other sentence to make more sense. Here are some examples of the top words:



2 Analysis question

A lot of these terms can be in the stop words list because the general grammar difference in would' and 'will' or 'can' and 'may' (these 2 are just examples) would not bring much difference to the meaning of the sentence and the point is still understood, so, it would make much more sense to expand the stop words list to a lot more connecting words than what we used in the project.

3 Analysis question



4 Analysis Question

It does follow the heap's law. Heaps' law predicts that the number of new words will increase very rapidly when the corpus is small and will continue to increase indefinitely, but at a slower rate for larger corpora. And as we can see from the data that we gathered, the total number of tokens is a huge number while the number of distinct words increases very slowly. In the heaps.txt the number of unique tokens increase rapidly but when we reach towards the end of the novel, the unique tokens completely slows down while the total number of words keep increasing.