Patrick Kelly

CS446

9/18/21

Part A: 72-96

Part B: 8-59

The first thing I did was open the stopwords file and read through the file line by line adding each unique word into a list. Then I opened the file that was provided by the user and read through the file line by line. I then would go through each line word by word and identify the strings that were abbreviations using the Regular expressions standard python library. Once the strings that were abbreviations were identified I replaced all the "." with an empty char. If a string was not considered to be an abbreviation it was sent to be tokenized. The first check the main tokenization function would do is turn all characters lowercase. Then the stopword removal process took place where I would check if the word was in the stopword list I made earlier. If it wasn't I would continue on with the tokenization process if it was the word would not be added to the final result. Then the stemming process would take place. I followed Step 1a and Step 1b from the textbook. After the stemming process the word was added to a final_words list. This list is iterated through and each word in it is written to tokenized.txt. In order to record most frequent terms for Part B a histogram was created. The histogram is iterated through and each word and its count is written to terms.txt. The first change I would make to the stemming process would be to filter out more 1 letter terms. In my opinion the 1 letter terms add little to understanding what the input files are about. The second change I would make would be to check if some words are actually part of the english language and if they are not to not add them to the final result. One hindrance with this idea though is that you need to still include proper nouns and our program turns all characters lowercase. Based on the top 200 terms from Moby Dick I would say they are relevant for the most part. The most recorded term is “s” which is not relevant to the book but all the other top terms such as “ahab”, “ship”, and “whale” are very relevant to the story. There are no top terms from the stopwords list and I think that there can be a standard stopwords list that can be used for all documents that filters out recurring useless words.

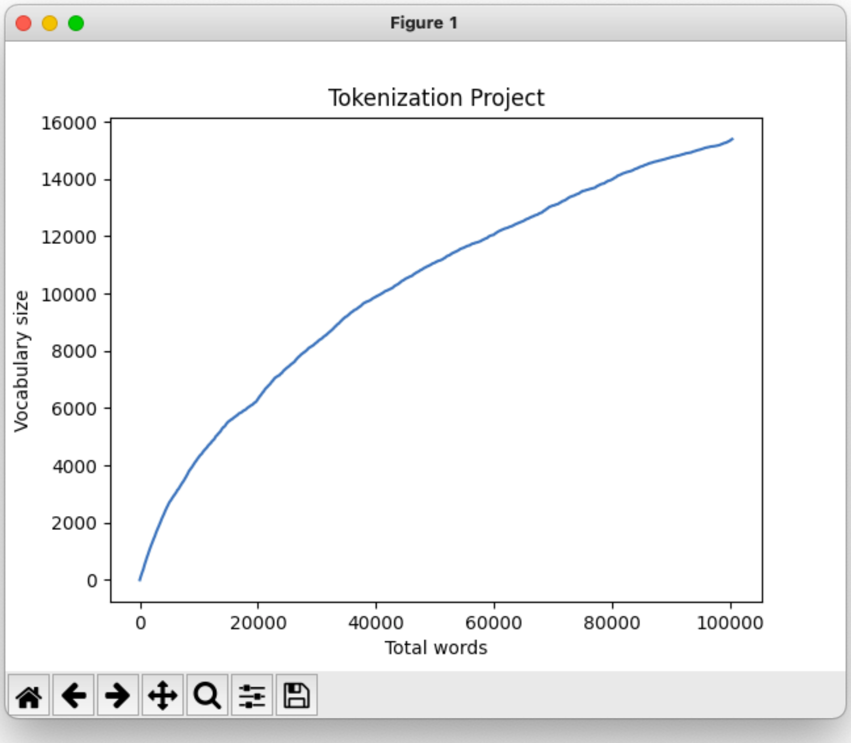
Figure 1
Tokenization Project
Vocabulary size
Total words