CS 446 Project Report

Kanav Talwar
10th Oct 2023

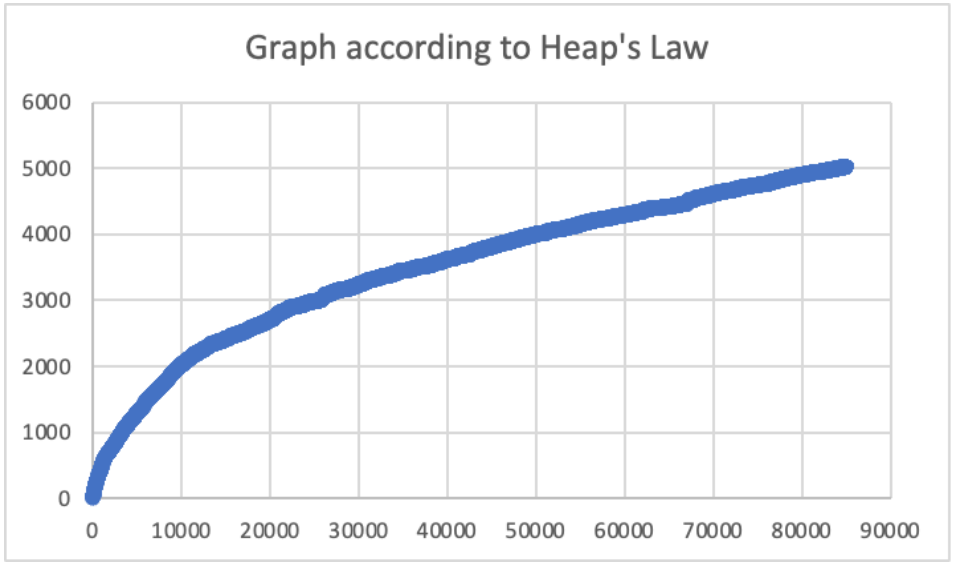
First Analysis Question:

No, the words with the most frequency do not have much relevance to the story, and are just everyday words used according to proper grammar. For example, the most frequent word is ‘her’ with a frequency of around 2561 and the second most frequent word is ‘i’ with a frequency of 1999. The only thing we can supposedly decode is that the protagonist of the book might be a woman as the third most relevant word is ‘she’ with a frequency of 1613. The first relevant word is ‘elinor’ with a frequency of 685

Second Analysis Question:

Yes I feel there are a few words which could be used as stopwords instead. The most obvious one can be ‘i’. ‘I’ is one of the most frequently used words in the English language. It appears in nearly every sentence or document, making it less informative in many contexts. Removing such high-frequency words can reduce the size of the dataset and improve the efficiency of text processing. Other ones can include ‘you’ and ‘so’ which are also used very frequently in the english language

Third Analysis Question:



Fourth Analysis Question:

Yes, I believe "Sense and Sensibility" adheres to Heap's Law. When we look at the graph that represents the evolution of the vocabulary size in the text, we can see a clear pattern emerge. In particular, when we begin to explore the text, it becomes clear that the more terms we encounter in the first sections of the text, the more unique words we uncover. This observation is consistent with Heap's Law ideas.