CS 446 Project 1 Report

Siddharth Jain

March 14, 2023

1. Run your program with fancy tokenization, stopping, and Porter stemming on sense-and-sensibility.gz and look at the -stats.txt file to see the most frequent terms from Sense and Sensibility. Are the top terms relevant to the story or do they seem like everyday words that aren't particularly to the novel? Support your answer with examples.

   The majority of the terms in SAS-stats.txt appear to be everyday words that are not particularly relevant to the story - "Sense and Sensibility". However, some of the terms are character names such as Elinor, Marianne, Dashwood, Willoughby, Colonel, and Lady, which are crucial to the plot of the novel as given on Wikipedia.

   Additionally, some of the words that appear frequently in the novel include "sister," "mother," "love," "marriage," and "wealth," which are also significant themes in the story but we can't make much out of them from the list.

   Overall, while some of the terms are relevant to the story, the majority are not particularly specific to "Sense and Sensibility."

2. Are there any of those top terms that should have been stopwords? Do the top terms or the list of all tokens suggest any changes to the tokenizing or stemming rules? What are they and why should they be made?

   There are so many words in the top terms list that should have been stopwords because they seem to be just everyday words and don't add much context to the story. Some examples are: i, her, you, his, had, but, have, all, so, him, etc.

   Yes, there are some changes that could be made to the tokenizer and the stemmer.

   - For the tokenizer, we should consider <Title>. <Name> as one token instead of multiple tokens. The title is almost as important as the name itself.

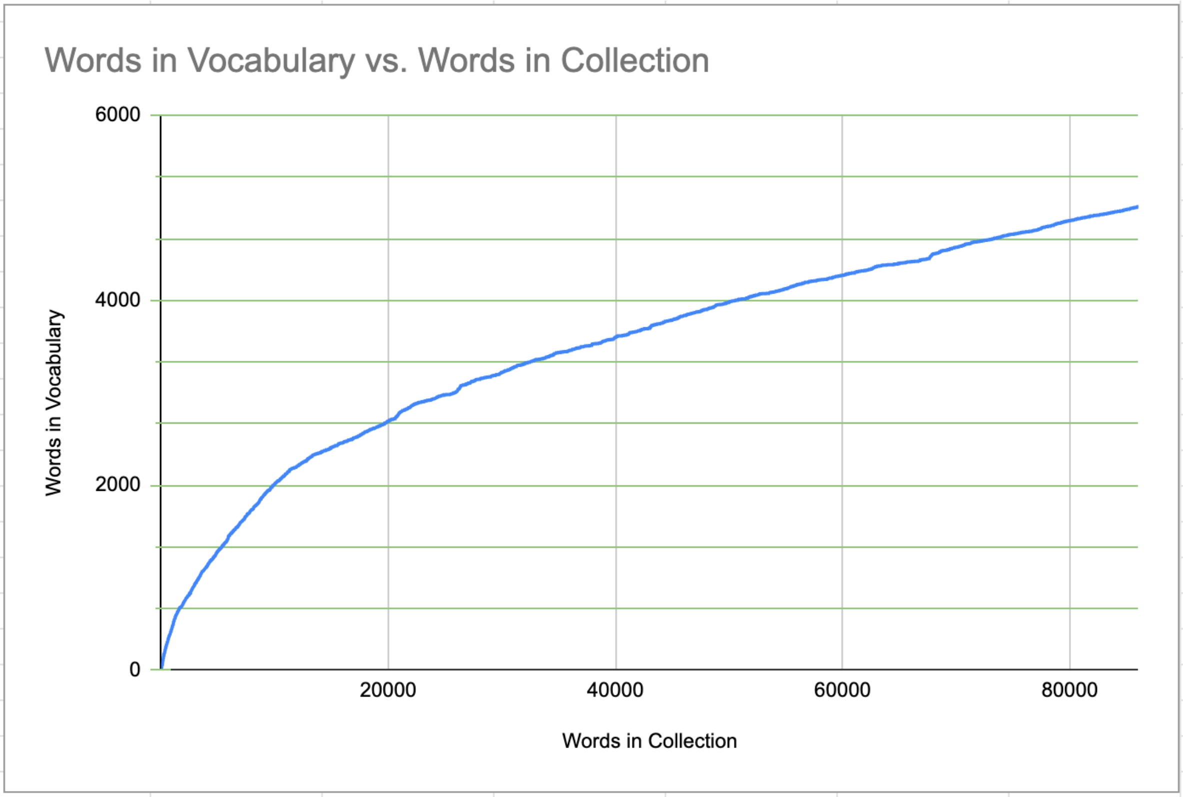
   - In the original text, we should identify names and other pronouns so that we don't stem them as the names of characters, places, etc. is important information and must not be altered.

   - Step 1c also seems very unnecessary as it stems a lot of words to meaningless words and also affected some character names:
     Ex: very > veri, every > everi, only > onli, lucy > luci,
     willoughby > willoughbi, lady > ladi, etc.

3. Figure 4.4. in the textbook (p. 82) displays a graph of vocabulary growth for the TREC GOV2 collection. Create a similar graph for Sense and Sensibility and upload an image of your graph.

   Note that you should be able to use the -heaps.txt file to generate the graph.

   Included in the Compressed version too as it was throwing a "MISSING FILES" error



4. Does Sense and Sensibility follow Heaps Law? Why do you think so or think not?

   Yes, Sense and Sensibility follow's Heaps Law. The curve we generated is clearly similar to the Heaps curve given in the book.

   It can also be verified mathematically:

   v = k * (n)^b

   5011 = 10 * (86061)^b

   501.1 = (86061)^b

   log(501.1) = b * log(86061)

   b = log(501.1) / log(86061)

   b = 2.69992440274 / 4.93480638813

   b = 0.54711860818 = approx. 0.5

   Therefore, as Beta = approx. 0.5, the data follows heaps law