CS 446 Report for P1- Dev Ahuja

Question 1: There are several changes that I observed in the project, which if implemented, can greatly improve the tokenization project:

1. Since almost all word separators are applied to the list in a way that there is no distinction between the application of those delimiters, there might be a **case for numbers**, as has been mentioned in the description as well, that two numbers are split into something very different. For example, 200,000 was separated into [“200”, “000”] where both numbers hold completely different meanings from the original number. This may tamper with the search results and produce inaccurate outputs. It can be changed by keeping numbers together instead of splitting them.
2. Converting all words to lowercase may be harmful in cases where we are dealing with **proper nouns**. Although modern search engines account for both the cases, the one we have implemented can change the meaning of the word. For example, “Bush” and “bush” will hold different meanings when the case is changed. This can be accounted for by either keeping both meanings or using advanced techniques to identify the context of the search.
3. The description for the word being **“short” in the Porter Stemmer** step 1b, can come off as a vague approach for classifying words when stemming. This occurred in the case of the word “fish” where it was made “fishe” because it adhered to the “short” rule but was not actually meant to be. A potential change for this is to make use of hybrid stemmers that are able to map the outputs of the porter stemmer to some possible related words.

Question 2:

- Based on the nature of the book, it follows a story of a family the characters of which occur at the top of the terms-B file. It can be understood by observing the file that the story revolves around “Elizabeth” (most recent) and the Bennet family along with Mr. and Mrs. Darcy. Following the relation of the main characters, it is also expected that relational names such

as “father” and “daughter” also occur at a large frequency. Most of them are surely not everyday words and pertain most to the context of the story.

- Although all words from the stopword list were removed, I believe that words like “chapter” and “sure” do not hold any direct meaning to the story. Moreover, “chapter” is only used to tell us the position in the book. I believe, there can be no list that accumulates stopwords for all documents as stopwords can arise based off of context and not a general meaning. Moreover, in different contexts, stopwords can also be meaningful words.

Question 3: Graph attached below: