Duc Nguyen
CS446

Report 1

1. My source code for part A contains in Tokenizer.java, Stoper.java, processingChar.java, and they are executed in Main.java. For part B, mostly everything is written in Main.java bellow Part B comment, except I used WordFrequency.java to store the strings and their occurrences, and one method mostFrequentTerm() in Tokenizer.java.

2. For Tokenizer, I use Scanner and Regular Expressions to manipulate the data. First, the tokens are split by all punctuations except '.', then if they are not abbreviations, they are split by '.' and lowercased. If they are abbreviations, then all the '.' are removed and each abbreviation is made into a token. Everything then stored in an ArrayList tokens for the next step.

   For Stoper, input datas are stored into an ArrayList of String stoplist. Stoper has an Arraylist of tokens which are words that have been tokenized previously form part 1 and an ArrayList of String stopResult to store the result of the process. The algorithm is to add all tokens that don't appear in the input stoplist and store them into stopResult.

   For PorterStemmer, processingChar is used to apply the rules 1ab to a word (or an array of character), and PorterStemmer is used to read the input data and call functions from processingChar to apply the rules to the whole document.

3. I used File to read input and to output data, Scanner and Reg Ex to match the tokenization requirements, ArrayList to store and manipulate datas. I also use PriorityQueue, Collectors, Comparator to rank and find top 200 most frequent terms.

4. Not splitting at decimals for example 3.14, or 3'2" (3 feet two inches), would improve the effectiveness of search. For stemming, there are still some special cases that cause both false negatives and false positive, such as past/paste, or spare/sparsity, so making a list of these special cases and adding more rules would also help to improve the effectiveness of search.

5. They seem just like everyday words. There are words such as “the”,”a”,”of”,”and”, or “an” should be in the list of stop words because they are too common but have little meaning. I don’t think that there is a complete list for all documents because it can vary greatly depend on the context or the length of each document.