In this project you will implement a program in Java for a collection of dataset of a textbook by Lewis Carroll

Task. Preprocessing:

Implement a program for preprocessing tokenization and stopword removal. The index terms or words will be all the words left after filtering out punctuation, numbers, stopwords, etc. You can use Java String tokenizer or your own method

- Input: Documents that are read one by one from the corpus
- Output: Words (vocabulary, all words/tokens).

Note: Make sure to show what data structures are used for this task!

Task 2. [20 points] frequency count: Now you have all of the words (Index terms) from Task 1. Now count the frequency of each word in the dataset.

- Input: Words obtained from the preprocessing module
- Output: frequency of each word (you can save in a file)

Task 3. [20 points]: Report top n words and ratio of the documents:

- Find the top n words in corpus and list them. (show in a Table: word, frequency)
- Find the ratio of data: Ratio = #words/(# tokens = all words in dataset after preprocessing))
- **Hint**: You may implement a method/function to do above

Task 4 (10 points): Implement below function in your program:

- How big was the dataset: Number of all tokens (include everything before preprocessing): function name can be sizeOfDataset
- Number of words: must be implemented in your code, function name numberOfWords after preprocessing
- Number of Stop words in dataset: must be implemented in your code, ie.e function name numberOfStopWords
- Number of punctuation: must be implemented in your code a function: i.e. function name numberOfPuncutation
- Your own creativity: Think to implement one method that you think it can be helpful/useful

Task 5: Report (20 points):

- Specify how you use and implement your algorithm and which Java data structure you used and why)

- Explain briefly each of the three steps.
- Complexity of each method, Big-O of each method in your program
- Show in a Table a statistic result about the dataset:
    - How big was the dataset
    - Number of words: must be implemented in your code, function name numberOfWords before and after preprocessing
    - Number of Stop words in dataset: must be implemented in your code, ie.e function name numberOfStopWords
    - Number of punctuation: must be implemented in your code a function: i.e. function name numberOfPuncutation

Task 6: Programming and Design Style (15 Points):

- Your coding style
- Efficiency of your program, methods, etc
- Using appropriate Data structure (make sure to understand where to use for example, Treeset, Hastset, HashTable, HashMap, TreeMap, ArrayList, etc etc) Why you used one rather than another one.
- You can explain in your code briefly
- Commenting and Doc style in your code: for example: what are the input/parameters of a method, return values, etc.
- Briefly describe each method job.
- How readable and modular is your program