MapReduce

This exercise has 4 parts. In this exercise, you will be writing and implementing two MapReduce programs. Both are a bit challenging, but they will help you to have a better understanding about the MapReduce implementation. After you write the programs, you will need to answer some questions about them.

Remember that neither problem is case sensitive, so transform words to lowercase or uppercase. Also remember to use the StringTokenizer to find the correct answers.

Part 1: Write a program that processes the FirstInputFile `http://www.gutenberg.org/cache/epub/100/pg100.txt` and the SecondInputFile `http://www.gutenberg.org/files/3399/3399.txt`. This program should count the number of words with a specific amount of letters in these files - for example, the number of words with 4 letters, 5 letters and so on. If one word is repeated 20 times in the text, count it individually 20 times.

Part 2:

- Q1: How many words are there with length 10 in FirstInputFile?
- Q2: How many words are there with length 4 in FirstInputFile?
- Q3: What is the longest length between words and what is its frequency in FirstInputFile?
- Q4: How many words are there with length 2 in SecondInputFile?
- Q5: How many words are there with length 5 in SecondInputFile?
- Q6: What is the most frequent length and what is its frequency in SecondInputFile?

Part 3: Write a second program that again processes the FirstInputFile `http://www.gutenberg.org/cache/epub/100/pg100.txt` and the SecondInputFile `http://www.gutenberg.org/files/3399/3399.txt`. However, in addition to counting the number of words with a specific amount of letters, if one word is repeated several times, count it only once. So, your output should be the frequency of words with same length, but count a repeated word only once. Note: You may need to use 2 MapReduce jobs.

Part 4: Answer Questions 7-12.

- Q7: How many words are there with length 10 in FirstInputFile?

- Q8: How many words are there with length 4 in FirstInputFile?
- Q9: What is the most frequent length and what is its frequency in FirstInputFile?
- Q10: How many words are there with length 5 in SecondInputFile?
- Q11: How many words are there with length 2 in SecondInputFile?
- Q12: What is the second-most frequent length and what is its frequency in SecondInputFile?