Quiz02: Apache Hadoop

Complete each of the following requirements by including to your submission a folder of Java source codes, a runnable JAR file, and a short text description.

1. Write the program `WordCount-1` that maps (extracts) words from an input source and reduces (summarizes) the results, returning a count of each word. [1]
2. Modify the above program to obtain the second version, called `WordCount-2`. You remove case sensitivity so that both lowercase and capitalized versions of your words are included in a single count. [2]
3. Develop the third version, `WordCount-3`, by creating a list of stop words and punctuation, and then having the application skip them at run time. [3]
   Store your list of stop words to a text file named `stop_words.txt`. Let your program refer to this text file during runtime.
4. Finally, for the fourth version `WordCount-4`, instead of outputting all distinct words with their associated counts, output only one word that has the highest frequency (if there are multiple words with the same highest frequency, just choose one randomly).

Please prepare your code such that the grader can replace the following files with his own files.

- The input and output text files, as well as their locations
- The list of stop words, `stop_words.txt`, as well as its locations

Also please prepare a careful guideline how to run your JAR files with argument specifications on Apache Hadoop.

References

[1] Example: WordCount v1.0
https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1.html

[2] Example: WordCount v2.0
https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount2.html

[3] Example: WordCount v3.0
https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount3.html

Citations are mandatory in your documentation and source codes.

Rubric

Requirements	1	2	3	4
Score	2	2	3	3