Project 1

This project is to create a WordCount Hadoop application. WordCount is a simple application that counts the number of occurrences of each word in a given input file(s). If you google “Hadoop WordCount,” you will find a lot of information including source codes on the Internet. It is also included as a MapReduce Tutorial example by Apache Hadoop: <https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>

Your task is to 1) create three java files: WordCount.java, WordCountMapper.java, and WordCountReducer.java, 2) convert them into a Jar file, and 3) run the Jar file in Hadoop to count the words in the given file titled *wordcountsample.txt*.

You need to turn in 1) the three java files, 2) the commands from converting them into a Jar file to running the Jar file in Hadoop, and 3) the screenshot of the final output showing the results of counting the words.

The source code used for this project is forked from Apache Hadoop MapReduce Tutorial, example: WordCount v2.0, <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v2.0>, and use CSUEB Hadoop system to finish this project:

1. Create folder named as “project1” in HDFS:

hdfs dfs -mkdir /home/student8/project1

1. Create sub-folder named as “input” under folder “project1” in HDFS:

hdfs dfs -mkdir /home/student8/project1/input

1. Create a file include a list of stop words and punctuation I want to omit from the output, the stop words include: “a an and but is or the to \, \. \; \: ”
2. copy file from local to HDFS:

hdfs dfs -copyFromLocal wordcountsample.txt /home/student8/project1/input

hdfs dfs -copyFromLocal patterns.txt /home/student8/project1/input

1. compile Java file:

javac -classpath /home/student8/hadoop-common-2.6.1.jar:/home/student8/hadoop-mapreduce-client-core-2.6.1.jar:/home/student8/commons-cli-2.0.jar -d . WordCount.java WordCountMapper.java WordCountReducer.java

1. Create a Jar file:

jar -cvf wordcount.jar ./WordCount\*.class

1. Run a Jar file on Hadoop:

hadoop jar wordcount.jar WordCount -Dwordcount.case.sensitive=false /home/student8/project1/input/wordcountsample.txt /home/student8/project1/output -skip /home/student8/project1/input/patterns.txt

1. Display the output on screen:

hadoop fs -cat /home/student8/project1/output/part-r-00000

1. Screenshot of final result in CSUEB Hadoop:

