Homework 6
STAT 4410/8416 Section 001
Fall 2014
Due: December 11, 2014 by midnight

1. **Working with HDFS:** The Hadoop Distributed File System (HDFS) allows us to manipulate massive amount of data using scalable computing power. For this question we will use the cloudera virtual machine version CDH 5.2X to manipulate data in HDFS. You can take help from others, but **submit your own work**.

   Please answer the questions below. The first question is answered so you can have an idea how to write answers for this homework.

   (a) In your virtual machine, create two HDFS folders using the following commands. Display the output of the command `hadoop fs -ls`.

   ```
   hadoop fs -mkdir wordcount
   hadoop fs -mkdir wordcount/input
   ```

   Answer:

   ```
   hadoop fs -ls

   Found 2 items
   drwxr-xr-x   - cloudera cloudera          0 2014-12-04 13:20 hdfs
   drwxr-xr-x   - cloudera cloudera          0 2014-12-04 22:08 wordcount
   ```

   (b) Download the file `words.txt` from blackboard and save it to your home directory (`/home/cloudera`). Then copy that file to the newly created HDFS folder `wordcount/input`. Please don't worry if the file is not clean. Just present the output of the following command.

   ```
   hadoop fs -ls wordcount/input
   ```

   (c) Now download the JAVA source codes `WordCount.java` from the blackboard and save it to your home directory. Also create a new folder called `wordcount_classes`. Now provide the output of the following command

   ```
   ls
   ```

   (d) Compile, build and run the `WordCount.java` program similar to what we did in slide 6 of the hadoop-mapreduce lecture. Output of the program should be saved in a folder called `wordcount/output` and present the result of the following command.

   ```
   hadoop fs -ls wordcount/output
   ```

2. **Pig in action:** For this exercise we will use `Pig` to manipulate data in HDFS. Please answer the following questions, provide all the `Pig` commands you used.

   (a) Launch the grunt shell using the command `pig`. Show the output of the following command and comment if the output is same as what you have seen in question (1d).

   ```
   grunt> ls wordcount/output
   ```

   (b) LOAD the data file `part-00000` you just created in folder `wordcount/output`. While doing this name the first column words and the second column count. Also, first column should be chararray and the second column should be int. Display first 10 rows.

   (c) ORDER the data by the count of words and display the top 10 most frequent words.