CSE 482: Big Data Analysis (Spring 2017) Homework 5
Due date: Monday, April 3, 2017 (before 9:00am)

Please make sure you submit the homework via the handin system (go to `https://secure.cse.msu.edu/handin`.

1. Write the corresponding HDFS commands to perform the following tasks. Each of these tasks must be accomplished with a single HDFS command. Hint: type `hadoop fs -help` for the list of commands available. Note the difference between the local (Linux) file directory and HDFS directory. To double-check your answers, you're encouraged to test the commands (on your own HDFS directory) to make sure they work correctly.

   (a) Make a directory on HDFS named `/user/hadoop`.

   (b) Upload a file named `data.txt` from the local filesystem to HDFS.

   (c) Rename the uploaded file on HDFS from `data.txt` to `temp.txt`

   (d) Move the file from its current location at `/user/hadoop/temp.txt` to `/user/hduser/temp.txt`.

   (e) Copy the file from `/user/hduser/temp.txt` to `/user/hadoop/data.txt` (note that the filename is also changed).

   (f) Display the content of the file `/user/hduser/temp.txt`.

   (g) Delete the file `/user/hadoop/data.txt`.

   (h) Download the file named `/user/hduser/temp.txt` to the local filesystem directory `/user/cse482`.

   (i) Merge all the files under the HDFS directory `/user/hadoop/results` into a single file `output.txt` to be stored in your current working directory.

   (j) Changing permission of the HDFS directory `/user/hduser` so that only the user hduser can read and write files stored in the directory (but not other users).

2. For each problem and data set described below, state how would you setup the (key,value) pairs as inputs and outputs for the mappers and reducers. Also, explain the operation performed by the map and reduce functions given their input. Assume your Hadoop program uses TextInputFormat as its input format (where each record corresponds to a line of the input file). Since the inputs for the mappers are the same (byte offset, content of the line), you only need to specify the mappers' outputs. If the problem requires more than one mapreduce jobs, specify the role of each map/reduce function along with their input and output key-value pairs). You should solve the problem with minimum number of mapreduce jobs.

   Example:

   Data set: Collections of text documents.

   Problem: Count the frequency of nouns that appear at least 100 times in the documents.

   Answer:

   - Mapper function: Tokenize each line into a set of terms (words), and filter out terms that are not nouns.

     Mapper output: key is a noun, value is 1.

     Reducer input: key is a word, value is list of 1's.

     Reduce function: sums up the 1's for each key (noun).

     Reducer output: key is a noun, value is frequency of the word (filter the nouns whose frequencies are below 100).

   (a) **Data set:** Student transcript database. Each record consists the following information: student ID, course code, semester enrolled, number of credits, and GPA for the course. For example:

   A123456074 CSE260 FS2013 3 3.5

   GPA is a numeric-valued attribute ranging from 0.0 to 4.0.

   Problem: Compute the average GPA for each student.

   (b) **Data set:** Facebook friendship graph. Each record corresponds to a Facebook user, followed by the list of his/her friends. For example, the graph data may contain the following records:

   john123 mary456 tom312 lee222
   mary456 john123
   tom312 john123 lee222
   lee222 john123 tom312

   Problem: Find pairs of users who are not friends with each other but share one or more common friends. This is known in network science literature as the "friend-of-a-friend" (FOF) problem. For example, mary456 and tom312 share a common friend, john123, but they are not friends with each other. In this case, the hadoop program should output the pair (mary456, tom312) only once, i.e., it should not output the duplicate pair (tom312, mary456).

(c) **Data set:** Online music streaming data. Each line in the data file has 4 columns (user_id, song_title, artist, vote), where vote is an integer value with the following values: +1 (the user liked the song), 0 (the user did not rate the song), -1 (the user disliked the song).

Problem: For each user, list his/her favorite artist, i.e., the artist that received the highest net vote. Do not output anything if the net vote is non-positive (0 or negative).

(d) **Data set:** Measurements of total daily precipitation for weather stations from around the world. Each line in the data files has 3 columns (station_id, date, total_precipitation).

Problem: Find the station_id and date of anomalous precipitation values. A precipitation value is considered anomalous if it is negative-valued or is 3 standard deviations or more away from its mean.

(e) **Data set:** Online 2-player gaming data. Each line in the data file has the following information:

`gameID, player1, player2, player1Score, player2Score`

For example, the line

`123311 Wchampion lansing233 50 24`

suggests that `Wchampion` won the game by a score of 50 to 24 (the winner is the player who achieves a score of 50 or higher). Note that a player can play against the same opponent more than once. Assume each player in the dataset has played more than 100 games.

Problem: For each player, find the toughest opponent to play against (i.e., the opponent to whom they have lowest winning percentage). For example, if `Wchampion` has a winning percentage of only 0.20 when playing against `msu4life` and has a winning percentage higher than 0.2 when playing against other opponents, the program should output the pair (playerID, toughestOpponent):

`Wchampion`, `msu4life`

For tie-breaking cases (i.e., if a player has more than one toughest opponent with lowest winning percentage), the program may randomly select one of the toughest opponents as its output. If a player has won all of his/her games, then the program should output the pair (playerID, NoToughestOpponent).

3. Write a Hadoop program that computes the correlation between pairs of stocks from the stock market dataset `stocks.csv` provided on the class webpage. Every line in the data file has the following information:

`date`, `AAPL`, `BAC`, `C`, `F`, `GOOG`, `HMC`, `MSFT`, `TYM`, `WFC`

where the first column is the trading date, followed by the closing prices for 9 different stocks (same data was used in homework 3). An example of the output reducer file may look as follows:

(`AAPL`, `BAC`) 0.42344

Deliverable: Your hadoop source code (*.java), the archived (jar) file, and the output result file (part-r-00000) that contains the pairwise correlation values.

4. Use the Best Buy online search query data `bbuy.csv` from the class webpage for this question. Each line in the query file has the following comma-separated attribute values:

```
user,sku,category,query,click_time,query_time
```

where user is the id of user, sku is the id of the product page clicked on by user, category is the product category, and query is the search string. For this exercise, you need to write a Hadoop program that identifies the most popular product page clicked on for each query. Specifically, the output of the reducer must have 2 **tab-separated** columns: (query, sku), where sku is the id of the product page most frequently clicked by users for a given query.

Deliverable: Your hadoop source code (*.java), the archived (jar) files, and the reducer output.