Assignment #5

1. **[MapReduce] (40)** This set of questions test the understanding and application of MapReduce framework.

   a. (20) Facebook updates the "common friends" of you and response to hundreds of millions of requests every day. The friendship information is stored as a pair (Person, [List of Friends]) for every user in the social network. Write a MapReduce program to return a dictionary of common friends of the form ((User i, User j), [List of Common Friends of User i and User j]) for all pairs of i and j who are friends. The order if i and j you returned should be the same as the lexicographical order of their names. You need to give the pseudo-code of a main function, and both Map() and Reduce() function. Specify the key/value pair and their semantics (what are they referring to?).

   b. (20) Top-10 Keywords. Search engine companies like Google maintains hot webpages in a set R for keyword search. Each record $r \in R$ is an article, stored as a sequence of keywords. Write a MapReduce program to report the top 10 most frequent keywords appeared in the webpages in R. Give the pseudo-code of your MR program.

2. **[Graph Parallel Models] (40)** This sets of questions relate to MR for graph processing

   a. (20) Consider the common friends problem in Problem 1.a. We study a "2-hop common contact problem", where a list should be returned for any pair of friends i and j, such that the list contains all the users that can reach both i and j within 2 hops. Write a MR algorithm to solve the problem and give the pseudo code.

b. (20) We described how to compute distances with mapReduce. Consider a class of d-bounded reachability queries as follows. Given a graph G, two nodes u and v and an integer d, it returns a Boolean answer YES, if the two nodes can be connected by a path of length no greater than d. Otherwise, it returns NO. Write an MR program to compute the query $Q(G, u, v, d)$ and give the pseudo code. Provide necessary correctness and complexity analysis

3. **[Hadoop] (30)**

a. (30) Hadoop Program:

The attached CSV file contains hourly normal recordings for temperature and dew point temperature at Asheville Regional Airport, NC, USA. The unit of measurement is tenth of a degree Fahrenheit. So, 344 is 34.4 F.

Write a program using Hadoop to compute and output daily average measurements for temperature and dew point temperature. The daily average measurements should include measurements for 24-hour period, for example from 20100101 00:00 (2010, January 1st, 00:00) to 20100101 23:00 (2010, January 1st, 23:00). Output the result in the format shown below - the columns are date and the combined result (separated by comma) of daily temperature and daily dew point temperature:

```
20100101    377.04, 285.58
20100102    378.67, 286.92
....        ...., ....
```

You may write the application in Java, C/C++ or Python language. Provide both source code and compiled code, if applicable, for your program.