**Assgnment 2: Map Reduce (10 Marks)**

**Aim:** This assignment aims to assess your understanding of Map Reduce framework and programming a distributed program using this framework.

**Description**

This assignment consist of 3 parts: 2 Basic and 1 Advance task.

**Part 1: Basic (5 Marks)**

In first task, your goal is to write a Map Reduce program that analyse multiple text files and find out for each unique word containing in these files the document name containing the word maximum number of times. The input to your program will be several text files. The ouptut should have rows with three fields: *Word DocumentName WordCount*. *WordCount is the number of times ‘Word’ apprears in the document.*  For example a sample output file will look like:

BigData Kit318.txt 20

Data Kit712.txt 21

**Part 2 and Part 3(2+3 Marks)**

Here, you need to implement a Map Reduce code for Hadoop that analyses given weather data. This part of the assignment consists of two further sub-tasks: Basic level (2%) and Advance level (3%).

**Data**

Input data will be several **.csv files** for different years. Each file contains several rows giving information about weather conditions at different weather stations on different days of the year. The data is from *ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by\_year/*. There are at least two measurements each day, one for the maximum temperature (TMAX) and one for the minimum temperature (TMIN), and sometimes one for the precipitation (PRCP). Each row contains following relevant information:

1. The weather station id
2. The date in format yyyymmdd
3. Type of measurement (for this homework we care about the maximum temperature TMAX and TMIN)
4. Temperature in tens of degrees (e.g. -90 = -9.0 deg. C., -184 = -18.4 deg. C.)

**Outline of Tasks**

* **Basic level: Finding Average (2%):** In first task, your goal is to write a Map Reduce program that can find the average maximum temperature at each station in different years. The input to your program will be the csv files for different years provided to you. The ouptut should have rows with three fields: *Stationid Year AverageTemp*. For example a sample output file will look like:
* ITE00100554 1789, -63
* ITE00100555 1789 -90
* GM000010962 1789 4
* EZE00100082 1789 -103
* **Advanced level: finding similarity between different stations (3%):** The goal of this task is to implement a MapReduce program that can find similarity between different weather stations. Similarity between two stations is calculated based on the following:

Similarity(p,q)=

where *nY*=number of years, *i* indicates year number, *minT* indicates average yearly minimum teperature, p and q are weather station ids.

You can assume output from the previous task as input to this task. Output for this task will be in following format:

weatherStationID1 weatherStationID2 SimilarityScore.

**Submission**

1. Source Codes of 3 Tasks
2. A report explaining map/reduce program. If any optimisation such as using combiner to reduce number of keys, is done to improve the performance, please also specify with that explaination. If you have taken inspiration from some MapReduce programs to complete these tasks, please give their reference

**Deadline**

31 August 2018 11.55pm