**CS 6350.001 Big Data Management and Analytics**

**Instructor: Prof. Latifur Khan**

**Homework 1**

Submitted By:

**Shyam Patharla (sxp178231)**

Given below are the steps for running the code for all the questions.

**Running Code for Question 1**

Jar File: MapReduce.jar

Class: Part1

1. Create a directory on HDFS and put the input files:

> hdfs dfs -mkdir /test

> hdfs dfs -put <location of soc-LiveJournal1Adj.txt on PC> /test

> hdfs dfs -put <location of userdata.txt PC> /test

2. Delete the output directory if it already exists:

> hdfs dfs -rm -r /test/out

3. Run the jar file:

> hadoop jar <Location of MapReduce.jar on PC> Part1 /test/soc-LiveJournal1Adj.txt /test/out

4. Read the output file:

> hdfs dfs -cat /test/out/part-r-00000

5. To get output for specific pairs, run:

> hdfs dfs -cat /test/out/part-r-00000 | grep "<userid\_1>,<userid\_2> <press ctrl+v> <press tab>"

**Output:**

0,1 5,20

20,28193 1

1,29826

6222,19272

28041,28056

**Running Code for Question 2**

Input Files:

1. soc-LiveJournal1Adj.txt

Jar File: MapReduce.jar

Class: Part2

1. Delete the output directories if they already exist:

> hdfs dfs -rm -r /test/out1

> hdfs dfs -rm -r /test/out2

2. Run the jar file:

> hadoop jar <Location-of-MapReduce.jar-on-local-PC> Part2 /test/soc-LiveJournal1Adj.txt /test/out1 /test/out2

3. Read the output file:

> hdfs dfs -cat /test/out2/part-r-00000

**Output**

16539,40423 64

3610,3634 61

31482,31545 60

41851,41903 58

2689,29425 55

47034,47047 55

32867,32869 54

7983,8003 52

4069,16215 52

4425,4447 52

**Running Code for Question 3**

Input Files:

1. soc-LiveJournal1Adj.txt

2. userdata.txt

Jar File: MapReduce.jar

Class: Part3

1. Delete the output directories if they already exist:

> hdfs dfs -rm -r /test/out1

> hdfs dfs -rm -r /test/out2

2. Run the jar file:

> hadoop jar <Location-of-MapReduce.jar-on-local-PC> Part2 /test/userdata.txt /test/soc-LiveJournal1Adj.txt /test/out1 /test/out2 <userid\_1> <userid\_2>

3. Read the output file:

> hdfs dfs -cat /test/out2/part-r-00000

**Output**

Sample output:

0,1 [Juan Figueroa: Columbia,Beth Watson: Lake Worth]

**Running Code for Question 4**

Input Files:

1. soc-LiveJournal1Adj.txt

2. userdata.txt

Jar File: MapReduce.jar

Class: Part4

1. Delete the output directories if they already exist:

> hdfs dfs -rm -r /test/out1

> hdfs dfs -rm -r /test/out2

2. Run the jar file:

> hadoop jar <Location-of-MapReduce.jar-on-local-PC> Part2 /test/userdata.txt /test/soc-LiveJournal1Adj.txt /test/out1 /test/out2

3. Read the output file:

> hdfs dfs -cat /test/out2/part-r-00000

**Output:**

24907 Amanda,1460 Rose Street,Harvey,Illinois,89

12547 James,4868 Lincoln Street,Camden,New Jersey,88

41403 Richard,3799 Dog Hill Lane,Hays,Kansas,88

2450 Kathleen,2741 Timber Ridge Road,Sacramento,California,88

2448 Dan,3057 Elmwood Avenue,Scottsdale,Arizona,88

2618 Leonard,4449 Kildeer Drive,Hampton,Virginia,88

2447 Mary,1833 Boring Lane,San Francisco,California,88

2435 Richard,4287 Locust Street,Albany,Georgia,88

2433 Jeane,4999 Bee Street,Traverse City,Michigan,88

21076 Joseph,1826 Alpha Avenue,Cumby,Texas,88