EE/CSCI 451
Spring 2020
Programming Homework 5

Assigned: March 24, 2020
Due: April 6, 2020, before 11:59 pm
Total Points: 50

1 Login to HPC

- The host is: hpc-login**3**.usc.edu
- Username and password are the same as your email account
- **Do not** run your program in the login node.
- After login, use the 'srun' command to run your program on a remote node. For example:

```
srun -n1 ./<your executable>
```

1.1 Spark Examples

Spark is installed in /home/rcf-proj/xq2/spark. To run a Spark python program, for example, the 'pi.py', follow the steps:

1. Login to HPC
2. Enter your project directory:

```
cd /home/rcf-proj/xq2/<your usc id>
```

3. Run:

```
srun -n1 ../spark/bin/spark-submit
    ../spark/examples/src/main/python/pi.py
```

4. Expected output: There is a lot of console output of spark framework. If the program runs correctly, you can find a line:

```
  Pi is roughly 3.142740
```

2 Introduction

The objective of this assignment is to gain experience with programming using the MapReduce programming model [1] in Apache Spark Cluster programming framework [2]. Apache Spark supports SCALA, python and java as programming languages. This assignment uses python as the programming language. If you use any other language, please provide detailed instructions for running the program in your submission.

3 K-means Clustering [20 points]

Based on the discussion slides, complete the Map (mapToCluster) and Reduce (updatemeans) functions of 'kmeans.py' [15 points]. Run the program and submit the output file produced. [5 points].

```
srun -n1 ../spark/bin/spark-submit kmeans.py data.txt means.txt
```

4 Triangle Counting [30 points]

Based on the discussion slides, write a program which uses map reduce function in Apache spark to count the number of triangles in a graph. The input graph and the description of its format is provided in the file named: p2p-Gnutella06.txt.

A python helper program named readgraph.py is provided which reads the input file and populates the nodes and edges to help you get started (you can run it using: python readgraph.py).

Your program should produce an output file which contains the number of triangles to which each vertex belongs to. [25 points]. Run the program and submit the output file produced. [5 points].

5 Submission Instructions

- Code: 'kmeans.py' and output file. (20 pts)
- Code: 'trianglecounting.py' and output file. (30 pts)
- Report: Write clearly how to compile and run your code. Screenshot of the execution on HPC.

You may discuss the algorithms. However, the programs have to be written individually. Submit the code via Blackboard. Make sure your program is runnable.

References

[1] "MapReduce,"
`http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf`

[2] "Apache Spark,"
`https://spark.apache.org/`