EE/CSCI 451
Spring 2017
Programming Homework 5
Assigned: April 11, 2017
Due: April 21st, 2017, before 11:59 pm, submit via blackboard
Total Points: 50

1 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.

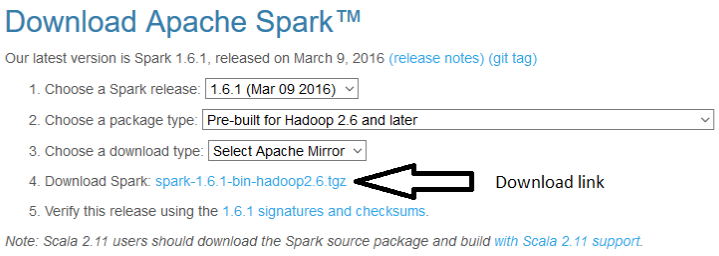

Figure 1: Spark Download link

2 Installation

We will download the pre-built binaries of Apache Spark for Hadoop 2.7 or later. Note that the figures shown here are based on the previous versions of the softwares. The instructions are updated for the latest version. The steps are as follows:

- Install python (https://www.python.org/downloads/). Make sure to check the option "Add python to PATH" while installing. (We have tested with python 2.7.3 but any similar version should work)
- Install Java 7 from http://www.oracle.com/technetwork/java/javase/downloads/jre7-downloads-1880261.html. Add the location of the bin directory of java to PATH variable as shown in Figure 2.
- Download and extract Hadoop 2.7 into a directory from http://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.7.3/hadoop-2.7.3.tar.gz. Add an environment variable named HADOOP_HOME which points to the root directory of the hadoop (The directory containing bin, examples directories). Update the PATH variable with the path to the directory containing the hadoop.cmd executable (the bin directory, e.g. see Figure 4)

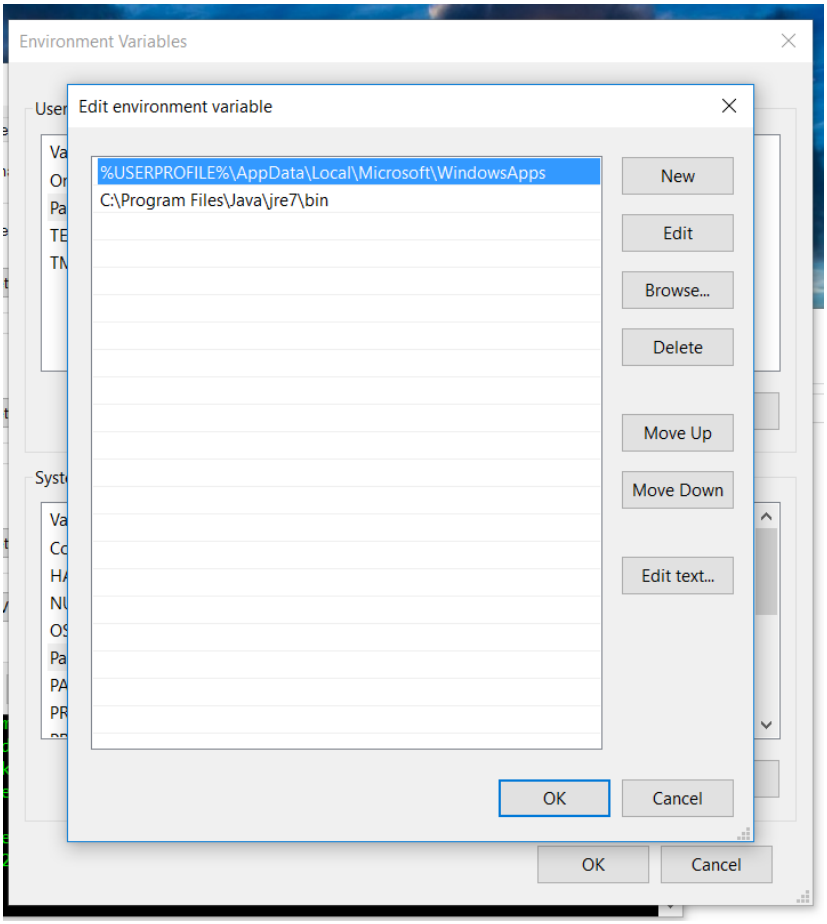

Figure 2:

- Download Apache Spark 2.1.0 from the following website (https://spark.apache.org/downloads.html). Make sure that the option package type is set to: "Pre-built for Hadoop 2.7 and later". See Figure 1.
- Extract the archive. You can use 7zip (http://7-zip.org/download.html).
- Open a commandline and change directory to the root of the extracted archive as shown in Figure 3
- Type: bin\spark-submit examples\src\main\python\pi.py. This command runs a spark application to calculate pi. If the program runs correctly, you are all set (See the figure pi-output.png provided as an attachment).
- You can follow the quickstart guide to try more examples of the spark framework (https://spark.apache.org/docs/latest/quick-start.html).

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].

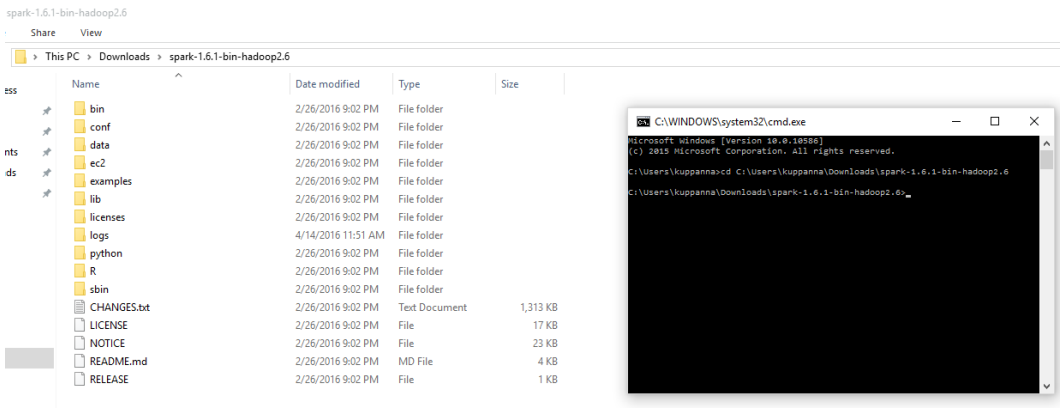

Figure 3: Apache Spark working directory

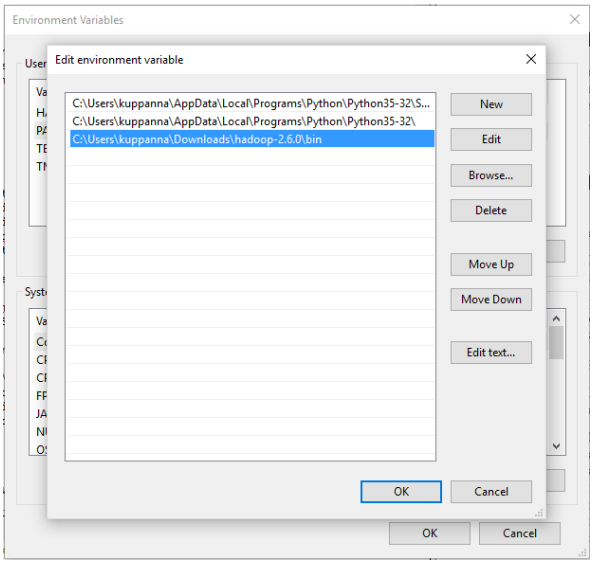

Figure 4: Hadoop path variable

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). The 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

You may discuss. However, the programs have to be written individually. You need to submit your python programs (kmeans.py, trianglecounting.py). You also need to submit the output files generated by both the programs.

References

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

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