EE/CSCI 451
Fall 2020
Programming Homework 6
Assigned: October 24, 2020
Due: November 3, 2020 AOE, submit via blackboard
Total Points: 100

General Instructions

- You may discuss the algorithms. However, the programs have to be written individually.
- Submit **the source code, all generated outputs and a simple report** via **Blackboard**. The report should simply record the five returned centroids of kmeans algorithm. Put all the files in a zip file with file name `<firstname>_<uscid>_phw<programming homework number>.zip` (do not create an additional folder inside the zip file). For example, `alice_123456_phw1.zip` should contain only the source codes and the report.

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.

2 Installation

We will download the pre-built binaries of Apache Spark for Hadoop 2.7 or later. **The steps for installation on Windows OS are as follows:**

1. Install python (https://www.python.org/downloads/). Make sure to check the option "Add python to PATH" while installing. (We have tested with python 3.6 and 3.8 but any similar version should work)

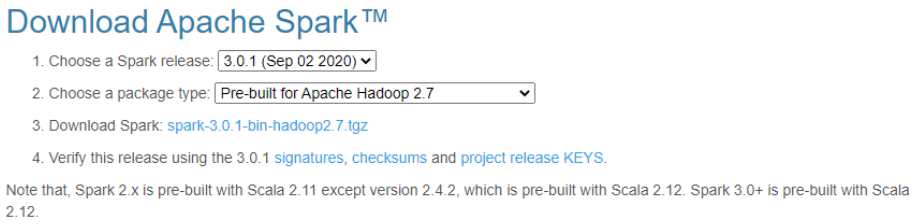

Figure 1: Spark Download link

2. Install Java 8 from (https://www.java.com/en/download/)

3. Download Apache Spark 3.0.1 from the following website

   (http://spark.apache.org/downloads.html). Make sure that the option package type is set to: "Pre-built for Hadoop 2.7". See Figure 1.

4. You can extract the files from the downloaded tarball in any folder of your choice using the 7Zip tool.

5. Open the extracted files and change directory to the root directory, namely, YOURPATH-3.0.1-bin-hadoop2.7

6. 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 (It should print a line "Pi is roughly xxx").

7. Note: You may need to install winutils and create another system environment variable to get rid of the error message "ERROR Shell: Failed to locate the winutils binary in the hadoop binary path". For more information on installing winutils, follow this easy-to-read guide: `https://deelesh.github.io/pyspark-windows.html#disqus_thread`

The steps for installation on MAC OS are as follows:

1. Install Python

2. Install java 8:
   brew tap adoptopenjdk/openjdk
   brew cask install adoptopenjdk/openjdk/adoptopenjdk8

3. Install Apache Spark: brew install apache-spark

4. Open bash profile to add Spark path: vim ∼/.bash_profile
   insert the following lines:
   export SPARK_HOME=/usr/local/Cellar/apache-spark/3.0.1/libexec
   export PYTHONPATH=$SPARK_HOME/python/lib/py4j-0.10.9-src.zip:$PYTHONPATH

5. exit vim by esc, :wq and source ∼/.bash_profile

6. Same as step.6 for Windows: If you successfully see the output line for pi.py, you are all set.

3 K-means Clustering [100 Points]

Based on the discussion slides, complete the Map (mapToCluster) and Reduce (updatemeans) functions of 'kmeans.py' [50 points]. Run the program and submit the output file produced.[50 points]

The command lines for running the program on Windows/MAC OS can be found in the commandline.txt provided to you.

References

[1] "MapReduce,"
`http://static.googleusercontent.com/media/research.google.com/en/us/`

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