Big Data Analytics

Exercise Sheet 6

Prof. Dr.Dr. Lars Schmidt-Thieme, Mohsan Jameel

Information Systems and Machine Learning Lab University of Hildesheim

June 02nd, 2016

Submission until June 8th, 2016, to mohsan.jameel@ismll.de

Exercise 1: Design problem (KMeans) (7 points)

Using MapReduce Framework give a solution for KMeans Clustering algorithm (see tutorial lec 2). In this task you are required to define key-value pair, Mapper and Reducer. How can you learn iteratively using MapReduce framework. You have to explain your algorithm with help of following 1D example. (Dry run your algorithm with following example)

Data (20, 23, 19, 29, 33, 29, 43, 35, 18, 25, 27)

Find 4 centroids i.e. k = 4

http://courses.cs.washington.edu/courses/cse599c1/13wi/slides/mapreduce-kmeans.pdf

Exercise 2: Implementation (KMeans) (5 points)

In this task you are required to implement KMeans using Hadoop MapReduce. Your problem should work for any number of dimensions i.e. data point are not just 1D each data point can have N features.

Use help from https://github.com/himank/K-Means/blob/master/src/KMeans.java

Exercise 3: Using Cluster (3 points)

Follow the instruction given in the tutorial 5, log in to the cluster and run HelloWorld program. When you submit job provide screenshot of output of qsub, qstat and output of your program.

Instruction:

Login to cluster (windows user use putty)

$ username@cluster.ismll.de

Copy script and java program from /tmp to your home director

$ cp -r /tmp/test_program ~/

Now you can submit your job with qsub, check status with qstat (provide output of these command)