Big Data Analytics

Exercise Sheet 7

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

Information Systems and Machine Learning Lab University of Hildesheim

June 09th, 2016

Submission until June 15th, 2016, to mohsan.jameel@ismll.de

Dataset:

Data set is taken from https://archive.ics.uci.edu/ml/datasets/Wine+Quality . The goal of this dataset is to model wine quality based on the chemical quantities. Download the modified version (that you can directly use as an input for MapReduce application) from:

http://math.ut.ee/~jakovits/upload/winequality-red_mr.csv

Some statistics about the dataset

- Column attributes are: "fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dioxide";"total sulfur dioxide";"density";"pH";"sulphates";"alcohol";"quality"
- Column values for a specific entry are separated by ";" character.
- The categories of "quality" (the last column) has value range : 3,4,5,6,7,8

Exercise: Advanced Analytics with Map Reduce

Q1: **(10 marks)** Your program should answer query *"What is the average value of* "fixed acidity" *for each category of* "quality"*?"*

a. Map should extract the specific column value for "fixed acidity" and output it as <quality, column value>
b. Reduce should then aggregate the list of values it gets as an input, calculate their average and output <quality, average>

Use skeleton code given at http://math.ut.ee/~jakovits/upload/MapReduceSkeletonSecond.java

Q2: **(5 marks)** How can you modify your code logic to answer question like median, maximum and minimum value of a given "column" for each category of "quality". (You are not required to write code for this part. List out the APIs or function/method/class in MapReduce framework that will help you achieve this)

Deliverables:

MapReduce application source code (only part of the code you added to skeleton code)

Output of running the application in Hadoop