Assignment 3

Tools & Techniques for Large-Scale Data Analytics (CT5105)
NUI Galway, Academic year 2017/2018, Semester 1

- ***Submission deadline (strict): Sunday, 22nd October, 23:59***
- *Please put your code into a single .zip archive with name "YourName_Assignment3.zip", submit via Blackboard*
- *Include all source code files (that is, files with name ending .java) required to compile and run your code..*
- *Unless specified otherwise in the question, use only Java 8 together with Apache Spark for this assignment.*
- *Please note that all submissions will be checked for plagiarism.*
- *Use comments to explain your source code. Missing or insufficient comments can lead to mark deductions.*

Question 1 [max. marks: 40]

Add a static method `countTemperature(t)` to your Java class `WeatherStation` from the previous assignment. It should return the number of times temperature `t` has been approximately measured so far by any of the weather stations in *`stations`* ("approximately" here meaning `t +/- 1`). (So this method provides just a part of the functionality of `countTemperatures(…)` from the last assignment.)

Use Apache Spark to implement this method, by making use of RDDs as far as possible, with operations parameterized with lambda expressions, and parallel computing where appropriate. No need to use the MapReduce pattern for this question (although certain map and reduction/aggregation-style operations will perhaps be useful here). Make your code as efficient and concise as possible by freely using the means Spark provides with RDDs/JavaRDDs (but without using DataFrames or Datasets). Hint: only very few lines of code are needed!

Also provide a `main()`-method which invokes `countTemperature(…)` with some test data and prints the results.

Question 2 (might require some knowledge from the next lecture)

a) [max. marks: 40] A typical large-scale data analytics task is *sentiment analysis* using classification, i.e., the computational determination of the attitudes of people towards a certain topic or item, achieved by supervised machine learning from large data sets with given sentiment annotations.

   Download the following data archive:
   https://archive.ics.uci.edu/ml/machine-learning-databases/00331/sentiment%20labelled%20sentences.zip
   Data set yelp_labelled.txt in this archive contains a number of single-sentence restaurant reviews, each labeled with "0" (negative sentiment) or "1" (positive sentiment).

   Create a program which builds a Linear Support Vector Machine (SVM) model using any 60% of the labeled sentences in yelp_labelled.txt as training data, using Spark MLlib with RDDs.

   Afterwards, use the learned model to predict and print the labels (sentiments) of a few test restaurant reviews taken from the remaining 40% of the data file.

b) [max. marks: 20] One way of estimating the accuracy of a classifier is by computing the *Area Under the ROC Curve* (AUROC, see https://spark.apache.org/docs/latest/mllib-evaluation-metrics.html for details). Add code to your program for a) which prints the achieved AUROC. Use the full remaining 40% of the data (i.e., the full data set without the training data) as test data for computing the AUROC.