Assignment 3

Tools & Techniques for Large-Scale Data Analytics (CT5105)
NUI Galway, Academic year 2019/2020, Semester 1

- ***Submission deadline (strict): Wednesday, 30th October, 23:59.*** *Late submission only with a medical or counsellor's certificate.*
- *Put all source code files (and other answers, if any) into a single .zip archive with name "YourName_Assignment3.zip" and submit via Blackboard by the deadline.*
- *Include all source code files (that is, files with name ending .java) required to compile and run your code, including all sources for your classes, interfaces, etc.*
- ***Please include the question number in the name of the respective .java file or package*** *(e.g., "MyClass_Q1.java")*
- *Unless specified otherwise in the question, use only plain Java together with Apache Spark for this assignment.*
- *All submissions will be checked for plagiarism.*
- ***Use Java comments to explain your source code. Missing or insufficient comments lead to mark deductions.***

Question 1 [40 marks]

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`* (that is, counting across all the stations). "Approximately" here means `t` is within interval `[t-1..t+1]`.
(So this method provides just a part of the functionality of `countTemperatures(…)` from the previous assignment.)

Use Apache Spark to implement this method, by making use of RDDs as far as possible, with operations parameterized with lambda expressions, and using parallel computing as appropriate. No need to use the MapReduce pattern for this question (but certain map and reduction/aggregation-style operations will probably 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 a small amount of code is required.

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

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

a) [40 marks] 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 learning from 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 imdb_labelled.txt in this archive contains a number of single-sentence movie 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 imdb_labelled.txt as training data, using Spark MLlib and RDDs (no Dataframes or –sets).

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

b) [20 marks] 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 Q2 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.