Spark Assignment

MSc in Business Analytics
Big Data Systems
Deadline: Tue April 6th 2021, 23:59

Background

You have been hired as a data scientist by a startup company that gathers and analyses transportation data, i.e., data for flights, marine traffic, train schedules, etc., to provide relevant services to its clients. Your company has recently gathered a new dataset that contains information about international flights in 2017. It has been assigned to you to analyse this dataset using Apache Spark to reveal insights about these data.

You can find the dataset here:
http://andrea.imis.athena-innovation.gr/aueb-master/flights.csv.zip

Task 1 [20 points]

Your first task is to calculate the average flight delays in the dataset (i.e., two numbers: average arrival & average departure delay). Your supervisor made it clear that you should choose SparkSQL with Python and DataFrames, so that your code should be compatible with other software products of your company. Your deliverables for this task are the following:

- A Python file (named “task1.py”) containing the code to produce the desired result.
- A report (named “task1.pdf”) explaining the basic intuition of your code.
- A screenshot (named “task1.png”) of the produced output (e.g., showing the result in the console).

Task 2 [35 points]

For this task you continue to work with SparkSQL. The objective is to create reports on the average and median departure delays of (a) all the airports, and (b) all the airways in the dataset. You should give four reports, two for the airports (average/median delays) and two for the airways (average/median delays). Each report is a CSV file containing one line for each airport/airway and the lines of each file should be ordered (in descending order) based on the corresponding criterion (average/median delay). No header files are required for these files. An extra instruction you have from your supervisor is that you should take care of some data outliers: you should not consider in your analysis any airports/airways that have extremely low number of flights; the criterion is that any airport/airway belonging in the lowest 1% percentile, regarding the number of flights, should be omitted. Your deliverables for this task are the following:

- A Python file (named “task2.py”) containing the code to produce the reports.
- A report (named “task2.pdf”) explaining the basic intuition of your code.
- The four report files (named “task2-ap-avg.csv”, “task2-ap-med.csv”, “task2-aw-avg.csv”, and “task2-aw-med.csv”) having the determined file structure. Please restrict the number of lines in the files (keep only the first 100 lines of each file) to keep the file size small.

Task 3 [45 points]

As a final task, your supervisor assigned to you to investigate if it is possible to train a linear regression model that could predict the departure delay a flight may have by using, as input, its origin (column “ORIGIN”), its airways (column “CARRIER”), and its departure time (column “DEP_TIME”). Again you should use Python and DataFrames, this time with MLlib. You should pay attention to transform the string-based input features using the proper representation format, and you should explain your choices. Special attention should be given to the “DEP_TIME” column: your supervisor told you that you should only consider the corresponding hour of the day (which, of course, should be transformed in a one-hot representation). For your training and testing workflow you should first remove outliers (see Task 2) and then split your dataset into two parts, one that will be used for training reasons and it will contain 70% of the entries, and a second one containing the remaining entries and which will be used for the assessment of the model. No need to implement a more sophisticated evaluation process (e.g., based on k-fold) is required in this phase. Your code should (a) prepare the feature vectors, (b) prepare the training and testing datasets, (c) train the model, (d) print on the screen the first 10 predictions, i.e., pairs of feature vectors (in compact format) and predicted outputs, on the screen, and (e) evaluate the accuracy of the model and display the corresponding metric on the screen. Your deliverables are the following:

- A Python file (named “task3.py”) containing the code of the preprocessing, training, and evaluation phase of your machine learning workflow.
- A report (named “task3.pdf”) explaining the basic intuition of your code and your design decisions and assumptions.
- A screenshot (named “task3.png”) showing the output of your machine learning workflow (e.g., showing the results in the console).

Bonus question [+5 points]

Write down something you liked for this assignment and something you disliked. You get all the points if you write something relevant to the question, regardless of which your opinion is. If your comments are irrelevant or do not provide any feedback you get no points. Your deliverable should be a report (named “bonus.pdf”) containing your feedback.

Submission instructions & honor code

Your code files should be fully replicable and readable (documentation comments are required and appreciated). Your code should work with Spark v.3 and should be ready to be executed (e.g., containing all the required import statements). Code failing to execute or producing wrong results will be penalised. You understand that this is an individual assignment, and as such you must carry it out alone. You may discuss with your fellow students to better understand the tasks/questions but you should not ask them to share their answers with you or to help you by giving you specific advice.

GOOD LUCK!