Spark Assignment

MSc in Business Analytics

Big Data Systems

Deadline: Tue April 2nd 2024, 23:59

Background

You have been hired by a new video streaming company that wants to use data science techniques to optimize their sales. It has been assigned to you to analyse a dataset of movies using Apache Spark (and PySpark, in particular) to reveal useful insights. You can find the dataset in the e-class page (the file "movie.json.zip").

Part A - SparkSQL

Task A1 [20 points]

Your first task is to explore the dataset. You need to use SparkSQL with Dataframes in a Jupyter notebook that delivers the following:

- It uses the json() function to load the dataset.
- It counts and displays the number of movies in the database.
- It counts and displays the number of PG-13 rated movies in the database.
- It uses the summary() command to display basic statistics about the "votes" field.
- It uses the groupby() and count() commands to display all distinct values in the "genre" field and their number of appearances

Your deliverable should a ready-to-run Jupyter notebook (named "id-tA1.ipynb"), containing your details (name, id) and explanations for each step of the code.

Task A2 [40 points]

For this task you continue to work with SparkSQL. This time, you need to provide a Jupyter notebook (again using PySpark and Dataframes) that delivers the following:

- It returns the "title", "year", and "users_rating" of the movie with the lowest "users_rating" that its title starts with the first letter of your first name and it has at least 200,000 votes.
- It returns the average "users_rating" and "metascore" of the movies that their title starts with the *second* letter of your first name and for which there are at least 200,000 votes.
- It returns the "title", "year", and "votes" of the movie with the most votes, when only movies with title starting with the *third* letter of your first name are considered.
- It returns the maxium and minimum users_rating score for all movies into which Bruce Willis is involved as an actor.

Your deliverable should be a ready-to-run Jupyter notebook (named "id-tA2.ipynb"), containing your details (name, id) and explanations for each step of the code.

Part B - MLlib

Task B1 [40 points]

As a final task, your supervisor assigned to you to investigate if it is possible to train a linear regression model (using LinearRegression() function) that could predict the "metascore" of a movie, using as input, its "user_rating", "runtime", "language", and "genre" (the first one). Similarly to the previous tasks you should use Python and Dataframes, this time with MLlib.

You should pay attention to transform the string-based input features ("runtime", "genre", "language") using the proper representation format, and you should explain your choices. Your code should (a) prepare the feature vectors, (b) prepare the training and testing datasets (75%-25%), (c) train the model, and (d) evaluate the accuracy of the model (based on the Rsquared metric) and display the corresponding metric on the screen.
Your deliverable should be a ready-to-run Jupyter notebook (named "id-tB1.ipynb"), containing your details (name, id) and explanations for each step of the code.

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 (using PySpark) 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!