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.

Task 1

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 comedies in the database.
- It uses the summary() command to display basic statistics about the “users_rating” field.
- It uses the groupby() and count() commands to display all distinct values in the “rating” field and their number of appearances

Task 2

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” and “year” of the movie with the largest “users_rating” that its title starts with the first letter of your last name.

- It returns the average “users_rating” of the movies that their title starts with the *second* letter of your last name.

- It returns the “title” and “year” of the movie with the most votes, when only movies with title starting with the *third* letter of your last name are considered.

Task 3

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 “user_rating” of a movie, using as input, its “metascore”, “runtime”, “genre” (the first one), and “language” (again 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 (80%-20%), (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.