De volta à semana 6 | **Lições** | Anterior | Próximo

Now that you are familiar with the dataset and exported it into a CSV file, let's start integrating this dataset with another small dataset and analyze it in Spark.

As the Sports Analyst, you are very interested in reporting on the countries with the most popularity in Twitter. So a good way to approach this problem would be to find which countries were mentioned the most in the tweets in your dataset and to analyze what words are being used the most in these tweets.

In addition to the CSV file you just exported from MongoDB, we give you a small dataset with the codes and names of some countries. To see this additional dataset, open the following file:

- Downloads/big-data-3/final-project/country-list.csv

To get you started, we have prepared a Jupyter notebook template, and started a SparkSQL context for you. Please open the notebook in:

- Downloads/big-data-3/final-project/SoccerTweetAnalysis.ipynb.

You will use this notebook to answer the questions below. So let’s get started.

Question 1: As a Sports Analyst, you are interested in how many different countries are mentioned in the tweets. Use the Spark to calculate this number. Note that regardless of how many times a single country is mentioned, this country only contributes 1 to the total.

Question 2: Next, compute the total number of times any country is mentioned. This is different from the previous question since in this calculation, if a country is mentioned three times, then it contributes 3 to the total.

Question 3: Your next task is to determine the most popular countries. You can do this by finding the three countries mentioned the most.

Question 4: After exploring the dataset, you are now interested in how many times specific countries are mentioned. For example, how many times was France mentioned?

Question 5: Which country has the most mentions: Kenya, Wales, or Netherlands?

Question 6: Finally, what is the average number of times a country is mentioned?

Don't forget to save the results of your analysis as you will have a quiz testing the correctness of these results following this exercise.

✓ Concluído

👍 👎 ⚑