Analyzing Tweets About Countries

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.

You will use the *SoccerTweetAnalysis.ipynb* notebook located at the *big-data-3* directory of your *jupyter-coursera* container to answer the questions below. Let's get started.

Question 1: As a Sports Analyst, you are interested in how many different countries are mentioned in the tweets. Use 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.

Go to next item   ✓ **Completed**

👍 Like   👎 Dislike   ⚑ Report an issue