Assignment 02:

Data Analysis using Spark, MongoDB

Due Date: 13 Dec 2020

In this assignment, there are a total of **two** questions related to Apache Spark and Mongo. **Only ONE question needs to be answered correctly**. You can answer either of them or both. Submit your work in a single document with the **answers** to the questions and the **code**. Preferably it can be the **pdf** of a documented **jupyter notebook.**

Question 01: Analyzing Tweets About Countries in Spark

Download the tweets data from below mentioned link:
https://nustedupk0.sharepoint.com/sites/BigDataAnalyticsMSCS-2k19MSDS-2k19/Class%20Materials/Assignment02/users.json

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 JSON file containing the **tweets** data, we give you a small dataset with the codes and names of some countries. To see this additional dataset, open the following file:

https://nustedupk0.sharepoint.com/sites/BigDataAnalyticsMSCS-2k19MSDS-2k19/Class%20Materials/Assignment02/country-list.csv

To get you started, we have prepared a Jupyter notebook:

https://nustedupk0.sharepoint.com/sites/BigDataAnalyticsMSCS-2k19MSDS-2k19/Class%20Materials/Assignment02/SoccerTweetAnalysis.ipynb

For CoLab users, you can upload this notebook on CoLab and run it there. Just be sure to setup the SQLcontext according to how it is setup in the tutorials given in the class.

You need to use Spark and answer the questions below (provide the code as well):

Question 1.1: 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 1.2: Next, computes 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 1.3: Your next task is to determine the most popular countries. You can do this by finding the three countries mentioned the most.

Question 1.4: 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 1.5: Which country has the most mentions: Kenya, Wales, or Netherlands?

Question 2.6: Finally, what is the average number of times a country is mentioned?

Question 02: Expressing Analytical Questions as MongoDB Queries [OPTIONAL]

Use the Cloudera VM to answer this question. Follow the b/m steps to setup mongoDB and the dataset.

- Download datasets and neccesary scripts to install the tools on the VM: https://github.com/mjawadak/Big_Data_Analytics/
- Install MongoDB `Big_Data_Analytics/datasets/big-data-3/mongodb/setup.sh`

Start MongoDB server and MongoDB shell. Open a terminal window by clicking on the square black box on the top left of the screen.

```
cd Downloads/big-data-3/mongodb
```

```
./mongodb/bin/mongod --dbpath db
```

Open a new terminal shell window, change to the *mongodb* directory, and start the shell: `cd Downloads/big-data-3/mongodb`
`./mongodb/bin/mongo`

If you cannot use the VM. Download the json file for the collection from https://nustedupk0.sharepoint.com/sites/BigDataAnalyticsMSCS-2k19MSDS-2k19/Class%20Materials/Assignment02/users.json

You will need to run MongoDB on your machine and upload the JSON into the database using **mongoimport** tool

On the cloudera VM, use the database called **sample** and collection called **users** to answer the following questions:

Imagine you are the Sports Analyst for a big magazine. The goal of this assignment is to demonstrate your data-driven reporting skills and express the following natural language questions as MongoDB queries on soccer-related tweets in English. Provide the MongoDB statement as well for each question.

- **Question 2.1:** How many tweets have location not null?
- **Question 2.2:** How many people have more followers than friends?

- **Question 2.3:** Return text of tweets which have the string "http://" ?
- **Question 2.4:** Return all the tweets which contain text "England" but not "UEFA" ?
- **Question 2.5:** Get all the tweets from the location "Ireland" and contains the string "UEFA"?