BIG Data Management-1

Assignment 1

Due Date: 26th May 2018

Assignment type: Individual

Submission documents:
(A word/PDF document with results/screenshots of results, along with the code files to be uploaded on LMS)

Assignment 1 contains four questions and will ask you to get familiar with aspects of Apache Spark. While first three questions require you to get familiar with Spark programming, the last question will ask you to understand an existing code and explain it in simple terms.

Q1. Consider the two data files (users.csv, transactions.csv). Users file has the following fields:

a) UserID
b) EmailID
c) NativeLanguage
d) Location

Transactions file has the following fields:

a) Transaction_ID
b) Product_ID
c) UserID
d) Price
e) Product_Description

By making use of Spark Core (i.e. without using Spark SQL) find out:

a) Count of unique locations where each product is sold.
b) Find out products bought by each user.
c) Total spending done by each user on each product.

Remember, you have to make use of Spark Core for this question. You cannot make use of Spark SQL for this.

Q2. For this question, please make use of the attached JSON file (tweets.json). Make use of Spark SQL library to answer the following questions:

a) Save the dataset as a DataFrame, and print the schema.
b) Get all of the tweets made by a user (any user would work. We should be able to replace user names to get tweets by that particular user).
c) Find count of all tweets by each user user.
d) Get a list of all of the people who are mentioned in tweets.
e) Count the number of time each person is mentioned in the entire dataset of tweets.

f) Give top 50 users who are mentioned the most.
g) Get a list of all hashtags mentioned in the dataset.
h) Find how many times each hashtag is mentioned in the dataset.
i) Get a list of all of the people who are located in a particular city (e.g. Paris)
j) Get country wise distribution of users, and find out which country ranks highest in terms of number of tweets, and number of users.
k) Find out number of tweets where a user is from France and mentions Paris in their tweets.

Q3. For this question, you would need to use the concepts learnt in Graph analytics session, and use datasets trip.csv and station.csv. The two files contain bike sharing data provided by SF Bay Area Portal. Trip.csv file contains following fields:

a) tripId
b) Duration
c) StartDate
d) EndDate
e) StartStation
f) StartTerminal
g) EndDate
h) EndStation
i) EndTerminal
j) BikeID
k) SubscriberType
l) ZipCode

Station.csv file contains following fields:

a) stationId
b) Name
c) Lat (Latitude)
d) Long (Longitude)
e) Dockcount
f) Landmark
g) Installation

Using the two files, please perform the following:

a) Import the data and create a graph using GraphFrames (Hint: Your graph will have nodes and edges. Nodes here would be individual stations so id field would be name field in station.csv file. Edges would have src and dst so it would Start Station and End Station fields in trip.csv file respectively. You can make use of other fields as properties of nodes and edges).
b) Find out number of incoming connections and outgoing connections for each node and print the top 10 nodes.
c) Find out which are the most common direct routes that people take and print top 10.
d) From the analysis in b, see which are the stations where people most frequently start their trips but do not come back. (Hint: You might have to

think of incoming connections as a ratio of outgoing connections). Print top 10 such stations.

e) Find all such patterns where any station a is connected to station b, b is connected to c, but c is not directly connected to a.

f) Run a PageRank algorithm to figure out which is the most important station in the entire graph.

Q4. Consider the Movie Similarities code and problem that was discussed during the class (Session 5). Please provide a brief write-up on the problem, steps needed to arrive at the solution (recommendation system), and how exactly those steps are implemented in the code. While you are doing so, please also mention what each line of code does (It is not sufficient to mention what each block of code does, you would have to provide explanation for each line).

All the Best!
Sadaf Zabeen & Monisha.k