CS 6350- ASSIGNMENT 5

Please read the instructions below before starting the assignment.

- There are 2 parts in this assignment. They can be done on Databricks or the UTD cluster. For both cases, you have to provide your code, README file indicating the dependencies of your code, and the output. For Databricks, you also have to submit a public URL of your notebook. Create separate folders for each part.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6350/CS6350_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page. Only one submission per team is required.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions on Piazza, and not through email to the instructor or TA.

ASSIGNMENT 5

Part I: GraphX

In this assignment, you will use Spark GraphX to analyze the high energy physics collaboration network data that is available at:
https://snap.stanford.edu/data/ca-HepTh.html

The data consists of a large list of collaborators, i.e. scientists that have collaborated in the past. You will use this data to construct a GraphX graph and run some queries and algorithms on the graph.

Below are the steps that you will perform. You should use Scala or Python under Spark to accomplish all of these tasks. You are free to use either UTD cluster or Databricks. You have to submit your code, not just a link to the public URL.

Step I:

Load the data into RDD using Spark. Define a parser so that you can identify and extract relevant fields.
Note that the dataset contains two-way relationships for each collaborator. That is, if X and Y have collaborated, there will be two lines in the file as:

X Y
Y X

Step II:

Define edge and vertex structure and create property graphs.

Step III:

Run the following queries using the GraphX API:

a. Find the nodes with the highest outdegree and find the count of the number of outgoing edges

b. Find the nodes with the highest indegree and find the count of the number of incoming edges

c. Calculate PageRank for each of the nodes and output the top 5 nodes with the larges PageRank values. You are free to define the threshold parameter.

d. Run the connected components algorithm on it and find the nodeids of the connected components.

e. Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, you can randomly select the top 5 vertices.

Part II: Spark Streaming

For this assignment, you will use Spark Streaming along with Twitter API to extract continuous tweets about restaurants in the Dallas area. You will then analyze each of them for positive and negative sentiments and report the net sentiment index for a particular time period e.g. from 12 noon to 3 pm on Friday March 31, 2017 the average restaurant sentiment for Dallas was +2.0

Below are the steps of this assignment:

1. For a given time period, extract tweets using Twitter handle for the following search criteria:

"Dallas" and "Restaurants"

An example of how to do this using Databricks can be seen at:
https://docs.databricks.com/_static/notebooks/2016-election-tweets.html

2. Analyze the sentiment of these tweets. This can be done in many different ways, such as using Stanford's coreNLP library. An example of how to do this using Databricks notebook is available here:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/1233855/3233914658600709/588180/latest.html

You can also analyze the sentiment of the tweets using any of the other standard methods, such as the Tweet Sentiment API https://www.tweetsentimentapi.com.

If none of the above two options work, a third option is for you to use a dataset of positive and negative words, as you did in a previous assignment. For each tweet, you can count the number of positive and negative words and classify it to the majority class. That is, if the positive words exceed the negative words, it will be positive tweet and vice versa.

3. Report the average of sentiment for all tweets. The range should be between 1 – 5, where 1 indicates totally negative reviews and 5 indicates totally positive reviews.

4. Repeat this process for at least 10 intervals using Spark streaming during different times and report your results.