Big Data 2019 17CS313

Assignment 3

Tweet Analysis with Streaming Spark

1. Assignment Objectives and Outcomes

a. The objective of this assignment is for the students to become familiar with the Spark Streaming programming paradigm.

For this assignment, students will stream data that is provided to them via a CSV file and then write Spark Streaming code using PySpark to perform a sequence of simple analyses on the Twitter FIFA World Cup 2018 dataset.

b. At the end of this assignment, the student will be able to write and debug Spark Streaming code.

2. Ethical practices

Please submit original code only. You can discuss your approach with your friends but you must write original code. All solutions must be submitted on the portal. We will perform a plagiarism check on the code.

3. Twitter FIFA World Cup Dataset:

a. The dataset is obtained from <link>

b. The CSV contains a collection of tweets during the 2018 FIFA World Cup

c. Use the provided dataset ONLY for the tasks given below.

Note: The provided dataset is semicolon (;) separated and can have empty values

The dataset does not have column names, please find metadata below for reference

Metadata:

ID, Lang, Date, Source, len, Likes, RTs, Hashtags, UserMentionNames, UserMentionID, Name, Place, Followers, Friends

4. Software/Languages to be used:

a. Python (Version **3.5.2**) with PySpark

b. Spark (Version **2.4.4**)

c. Please make sure to use Hadoop Version **3.2.0** and Java Version 1.8.0 only.

5. Marks:

a. The assignment is for 5 marks.

b. Each Task is for 2 marks.

c. The viva is for 1 mark.

6. Submission Date:

a. October 31 (Thursday), 2019 (Further details will be shared later)

7. Tasks:

a. The assignment is divided into two tasks each of which carries two marks:
   - i. Task 1 - Structured Streaming using HDFS:
     1. Most common hashtag
     2. Most popular twitter user
   - ii. Task 2 - Spark Streaming using Sockets: Top 3 hashtags in each window

b. For Structured Streaming, the data should be streamed from HDFS.

c. For Socket Streaming, the data will be streamed from a Socket script, which will be provided.

d. The output of the tasks are to be written to STDOUT.

e. Submit a one page report based on the given template and answer the questions on the report.

8. Submission Link:

Will be shared with you on Piazza at a later date.

9. Task Specifications:

a. Task 1:

i. Problem Statement Subtask 1:

An etymological study - most common hashtag

Description

Create a directory in HDFS “/stream” and monitor it using readStream

Upload the CSV files to the directory

Stream CSV files to spark as they are discovered in that directory

Determine the most common hashtag of each file received by spark

The number of files given is 1, we may use arbitrary number of files to test your code

Write output to STDOUT

ii. Problem Statement Subtask 2:

A popularity study - the person with the highest followers to friends ratio.

Description

Create a directory in HDFS “/stream” and monitor it using readStream

Upload the CSV files to the directory

Stream CSV files to spark as they are discovered in that directory

Determine the Name of the tweet with the highest followers to friends ratio

Output should be of the form Name,Ratio

The number of files given is 1, we may use arbitrary number of files to test your code

Write output to STDOUT

Note: Please use Structured Streaming and not Spark Streaming

Refer: https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

b. Task 2:

i. Problem Statement:

A sliding window study - 3 most common hashtags of each window

ii. Description:

Stream the data from the Socket using the provided code

Accept the Window Size and Batch Duration as command line arguments

Determine the 3 most common hashtags of each window

Write the output to STDOUT

iii. Output Format:

1. Comma separated values in the form of -
   1st most common hashtag , 2nd most common hashtag, 3rd most common hashtag
2. Each pair must be in a new line with there should be **no spaces** in between the values. (The values below are just for representational purposes)
   Eg. a,b,c
   x,y,z
   ...