Assignment 6
Individual Assignment

MET CS 755 - Cloud Computing
Data Stream Processing using Live Tweeter Stream

1 Description

In this assignment you will implement a stream processing system that processes the live twitter data.

2 Data Set

To get access to the Twitter Live stream data, you need to go to http://apps.twitter.com, create a twitter account and create an app, then your consumer key and secret will be generated for you after that.
consumer_key="YOUR CONSUMER KEY"
consumer_secret="YOUR SECRET KEY"

Then you will be redirected to your app's page. Then create an access token under the the **"Your access token"** section
access_token="YOUR ACCESS TOKEN"
access_token_secret="YOUR ACCESS TOKEN SECRET"

We provided a template implementation on blackboard for setting up tweeter stream in python, and sending data over TCP socket to Spark. You need to get the template implementation and set your keys. Our implementation uses a python library named "tweepy", you need to install it "pip install tweepy".

3 Assignment Tasks

Please select one of the following 2 tasks (You need to do one task only):

3.1 Task 1 : Important users that are tweeting about "Data" (20 points)

In this task you need to implement a spark streaming task that finds the **top-20** important twitter users that are tweeting about **"data"** in the past **5 minutes** (Filter the twitter stream by using the keyword **"data"** capitalization is not important - map all words to lower case and then filter).

You program should deliver results every **20 seconds**.

How important a user is should be calculated based on number of followers of the user (you can find the number of follower from **'[user']['followers_count']** in JSON object).

Let your application run for about 20 minutes and collect results.

You can use our template implementation provided on blackboard - resources.

3.2 Task 2 - YOUR ASSIGMENT IDEA (20 Points)

Describe a small mini project for mining interesting information from live tweeter stream.

Study the JSON object of tweet and try to understand the fields. Here you can find the twitter JSON specification and a JSON example file

- `https://dev.twitter.com/overview/api/tweets`
- `https://gist.github.com/hrp/900964`

1. Describe the research problem and argument why it might be important
2. Implement your solution using twitter data set and Spark Streaming

4 Important Considerations

4.1 Machines to Use

You can run these tasks on your laptop because the time window is small and the size of the data is relatively small.

4.2 Academic Misconduct Regarding Programming

In a programming class like our class, there is sometimes a very fine line between ”cheating” and acceptable and beneficial interaction between peers. Thus, it is very important that you fully understand what is and what is not allowed in terms of collaboration with your classmates. We want to be 100% precise, so that there can be no confusion.

The rule on collaboration and communication with your classmates is very simple: you cannot transmit or receive code from or to anyone in the class in any way—visually (by showing someone your code), electronically (by emailing, posting, or otherwise sending someone your code), verbally (by reading code to someone) or in any other way we have not yet imagined. Any other collaboration is acceptable.

The rule on collaboration and communication with people who are not your classmates (or your TAs or instructor) is also very simple: it is not allowed in any way, period. This disallows (for example) posting any questions of any nature to programming forums such as **StackOverflow**. As far as going to the web and using Google, we will apply the **”two line rule”**. Go to any web page you like and do any search that you like. But you cannot take more than two lines of code from an external resource and actually include it in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found on the web does not render the ”two line rule” inapplicable. It is still a violation to obtain more than two lines of code from an external resource and turn it in, whatever you do to those two lines after you first obtain them.

Furthermore, you should cite your sources. Add a comment to your code that includes the URL(s) that you consulted when constructing your solution. This turns out to be very helpful when you’re looking at something you wrote a while ago and you need to remind yourself what you were thinking.

4.3 Turnin

Create a single document that has results for all three tasks.

For each task, copy and paste the result that your last Spark job.

Also for each task, for each Spark job you ran, include a screen shot of the Spark History.

Please zip up all of your code and your document (use .gz or .zip only, please!), or else attach each piece of code as well as your document to your submission individually.