Department of Computer Science and Engineering
University of Puerto Rico
Mayagüez Campus

Big Data Analytics
Fall 2018

Project 2: Spark Twitter Analysis
Due Date: November 27, 2018, 11:59 PM

Objectives

1. Use Spark, and SparkSQL analyze trends contained in a collection of tweets.
2. Become familiar with streaming concepts

Overview

You will design, implement and test a series of programs that will analyze live tweets. Your solution will:

1. Capture the tweets from the Tweet Stream API
   https://dev.twitter.com/streaming/overview

   For this purpose, you can use:
   - python twitter (https://pypi.python.org/pypi/twitter)
     - pip install twitter
   - tweetpy (http://www.tweepy.org/)

2. Put the tweets into HDFS
3. Read the tweets from HDFS with Spark
4. Use Spark, Hive, and SparkSQL to implement the following operations:
   a. Capture the top 10 trending hashtags (most viewed) in the last 60 minutes. Refresh hourly
   b. Capture the top 10 trending keywords (most viewed) in the last 60 minutes Refresh hourly. (No stop words here)
   c. Capture the top 10 participants (most tweets posted) in the last 12 hours Refresh every hour

5. Count the number of occurrences for these keyword, in intervals of 1 hours, on each day,
   a. Trump
   b. Flu
   c. Zika
   d. Diarrhea

e. Ebola
f. Headache
g. Measles

You Must accumulate statistic for at least 3 days

Your solution will consist of a collection of Python programs, and SQL queries that perform tasks 1-5.

Visualization

Provide a means to visualize the results of the tasks 1-5, using the D3.js library. You are free to use the charts that you think best fits the visualization.

Deliverables

- **GitHub repo with all the code**

Grading

- **Project will be graded via demonstration of working code, running in cluster mode, forked from GitHub repo.**

PROJECT DUE DATE: **11:59 PM – November 27, 2018**.