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

CIIC 8995/5995 5016 – Big Data Analytics
Spring 2017

Project 3: Sentiment Analysis for Tweets
Due Date: Jun 23, 2017, 11:59 PM

Objectives

1. Use MLLib, Spark Streaming, SparkSQL, and Kafka to analyze trends contained in a collection of tweets.
2. Become familiar with Machine Learning concepts

Overview

You will design, implement and test a series of programs that will analyze the sentiments of 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 Kafka
3. Read the tweets from Kakfa with Spark Streaming
4. Use Spark, Spark Streaming , Hive, and SparkSQL to collect and store the tweets.
5. Implement sentiment analysis for groups of tweets that contain the following keywords:
6. Count the number of occurrences for these keyword, in intervals of 1 hours, on each day,
   a. MAGA
   b. Dictator
   c. Impeach
   d. Drain
   e. Swamp
   f. Comey

   You Must accumulate tweets for at least 3 days

Your solution will consist of a collection of Python programs, and SQL queries that perform tasks 1-6.

Visualization

Provide a means to visualize the analysis of sentiments, using the D3.js library. You are free to use the charts that you think best fits the visualization. For example, you can use pie charts to show % of positive or negative tweets for each group of tweets (groups are defined by keyword).

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 – June 23, 2017**.