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

Big Data Analytics
Spring 2017

Project 3: Sentiment Analysis for Tweets
Due Date: December 4, 2018, 11:59 PM

Objectives

1. Use Keras, TensorFlow, and SparkSQL, 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 HDFS (steps 1 and 2 were done un project 2).
3. Read the tweets from HDFS with Spark
4. Use Spark, Hive, and SparkSQL to collect and store the tweets in a manner easy for Keras to ingest the data.
5. Classify the tweets that contains the following keywords:
   a. Flu
   b. Zika
   c. Diarrhea
   d. Ebola
   e. Headache
   f. Measles

   The classification will put the tweet into one of this classes:
   a. 0 – Does not talk about a medical condition
   b. 1 – Does talk about a medical condition

c. 2 – ambiguous

You need to use at least 2 different classifier models.

You will be given a labelled data set to train your classifier. You will then use for this classifier to classify the data you collected in project 2.

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 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 – Dec 4, 2018**.