Feature Extraction on Twitter Streaming data using Spark RDD

Analyze real-time tweets using Twitter's Streaming API and spark streaming. The anticipated results is to be able to tell what percentage of tweets are classified into a particular feature. The features we looked at were opinions, tentative, vulgarity, positive/negative/neutral.

Getting Started

These instructions will get you a copy of the project up and running on your local machine for development and testing purposes. See deployment for notes on how to deploy the project on a live system.

Prerequisites

1. Apache Spark - Please refer this link to install the spark streaming module to your OS.

   Mac :

   https://medium.freecodecamp.org/installing-scala-and-apache-spark-on-mac-os-837ae57d283f

   Windows :

   http://www.ics.uci.edu/~shantas/Install_Spark_on_Windows10.pdf

2. Unix File system

3. Python 2.7 or above

Installing

1. You need to get twitter credentials in order to access the twitter data. Please follow below link to get auth parameters https://themepacific.com/how-to-generate-api-key-consumer-token-access-key-for-twitter-oauth/994/
2. Update BigProj.py file with your credentials.
3. You need to copy the script to your home path in terminal.

Deployment

Please run below commands in chronological order :
python <path>/BigProj.py | nc -lk 9999
Spark-submit <path>/streaming.py localhost 9999
Go to terminal, run copied script to get results
Run <path>/dataGraph.py to visualize the results.

Built With

- Python
- Unix Shell Scripting