Jeff Levesque

Professor Landowski

IST 652

Homework #2

**Data / Source**

The provided json dataset represents historical twitter data between 2013-08-30 through 2018-08-30. More specifically, these tweets are an aggregation of tweets containing the keyword apple. The corresponding dataset has been git versioned:

* <https://github.com/jeff1evesque/ist-652-hw/blob/master/hw2/apple--2018-09-03T13-11-54.json>

However, the following script was implemented to attain the above dataset:

* <https://github.com/jeff1evesque/ist-652/blob/master/utility/twitter_scraper.py>

**Analysis**

Data Preparation:

Preparation for the acquired dataset was not needed. Specifically, the twitterscraper package was used to generate the necessary json file:

* <https://pypi.org/project/twitterscraper/>

Therefore, if a tweet exists with the defined keywork apple, then an entry will exist in the overall json output file.

**Program Description**

Various package dependencies were required. However, these packages were streamlined into an automated build using vagrant:

* <https://github.com/jeff1evesque/ist-652-hw/blob/master/hw2/Vagrantfile#L10-L15>

Therefore, the only necessary dependency, is to git clone the above repository, as well as downloading the following applications:

* Vagrant
* VirtualBox

Once completed, running the command vagrant up, will build then run the entire application, including generating the corresponding traceback of the analysis in the same terminal window. Specifically, descriptive statistics of apple was computed for all tweets, where retweets were zero or more. Then, the same statistics were aggregated and determined for tweets having one or more retweets.

**Result / Output**

Once the vagrant up built the corresponding virtual machine with all the necessary packages, and mongodb running, the custom run.py was automatically executed:

=============================================

Retweets on apple

=============================================

Duration: 2013-08-xx through 2018-08-xx

Retweets >= 0

total retweets: 38.00

mean retweets: 0.19

stdev retweets: 0.96

=============================================

Duration: 2013-08-xx through 2018-08-xx

Retweets = 1

total retweets: 18.00

mean retweets: 1.00

stdev retweets: 0.00

=============================================

Duration: 2013-08-xx through 2018-08-xx

Retweets > 1

total retweets: 20.00

mean retweets: 6.67

stdev retweets: 4.51

=============================================

The results indicate the frequency of retweets greater than 1, is roughly half the frequency when retweets occurred just once. Another interesting result is that the standard deviation of the retweets is zero, for the case when retweets only occur once. Since retweets was relatively sparse for the given timespan, it would be more interesting to cover more historical data, perhaps 10 years. Additionally, more analysis incorporating likes, as well as sentiment analysis, on a variety of dimensions would provide more interesting results on the provided dataset.