J. Chon, A. Debruyn, R. Raymond, K. Orton, K. Wong
Faculty Advisors: Dr. Feng Wang, Dr. Haiyang Wang
School of Mathematical and Natural Sciences

Twitter Data Collection, Analysis & Modeling Framework

Purpose of Project

Social media provide new channels for people to share and exchange information and generate immense data every day. Applying big data analysis techniques on social media data can discover and predict a wide array of phenomena in real life. For example, previous works have found a high Pearson correlation factor between flu data collected from Twitter and CDC ILI (Influenza Like Illness) data. However, due to the fact that the social media data are usually noise, unstructured, and large scale, collecting, processing, and modelling such data is an intimidating task. The goal of this project is to build an efficient and flexible Twitter data collection, analysis, and modeling framework. This framework provides the necessary modules to apply big data analysis to social media data to reveal patterns and events in the social media.

Data Collected

The collected raw Tweet data covers several dimensions and categories as described below:

Category	Amount
Total Size Of Data	150 GB
Total # Tweets	121,556,931
# Source Tweets	40,518,977
# Unique Users	19,083,164
Data Collection Start	October 11, 2013
Data Collection End	March 17, 2015

- Any Tweet that contains the keyword “flu”. This is one of the most common illnesses and that the CDC tracks very closely
- Tweets containing the keyword “Ebola”. Identifying the spreading pattern for an actual outbreak of an unexpected disease, increases the accuracy & speed of the predictive model in identifying new outbreaks
- Tweets related to the Malaysian Airlines flight disappearance. Tracking an event with high media coverage & duration, gives insight into the unique communication fluctuations due to such events

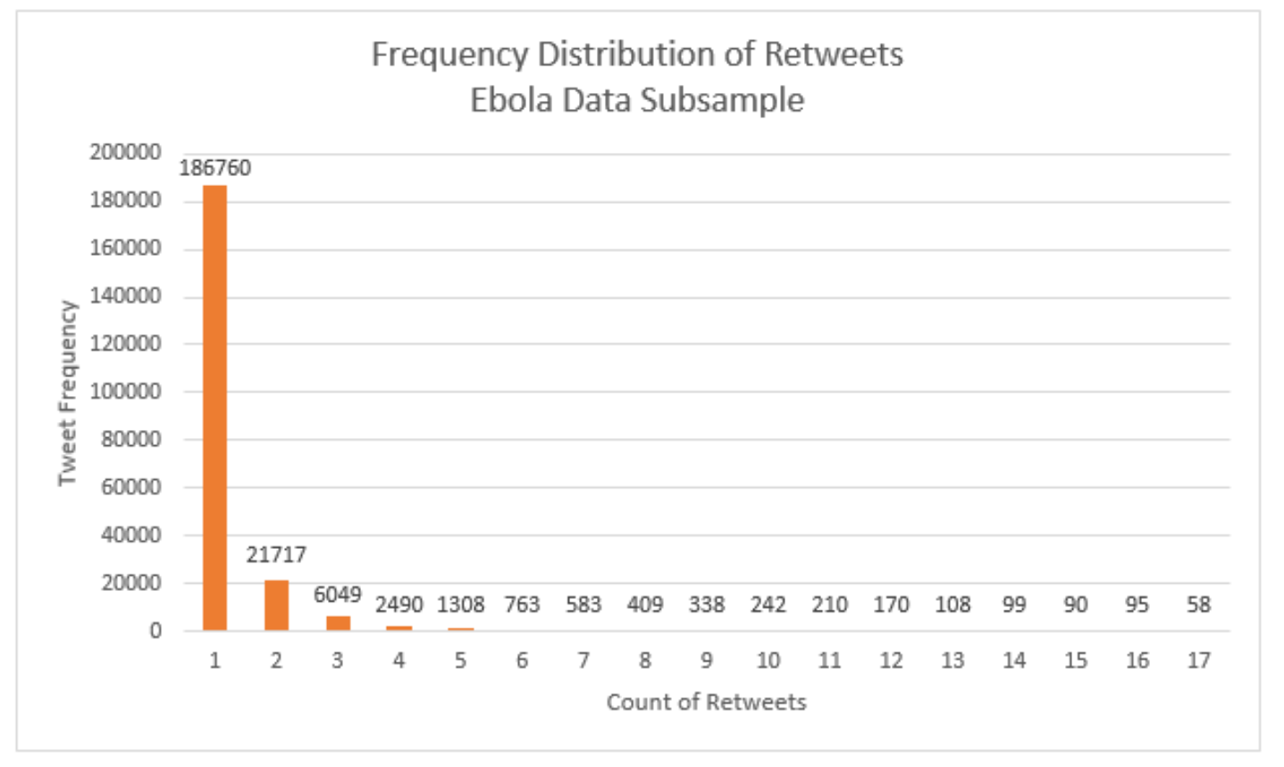
Framework

As illustrated in Figure 1, the raw data consists of Tweets posted by users of Twitter, the users’ profiles, and the followers of the collected users. Tweet Stream Collection module handles establishing and maintaining the connection with Twitter to retrieve Tweets based on chosen keywords and/or Tweets with a particular user as a source. For every unique user encountered the user’s profile is extracted and the user’s followers retrieved from Twitter and stored in a MongoDB database. The first stage of processing the raw data involves extraction of certain features, e.g., user distance from the source, geo-location of a tweet, user interaction. The second stage generates a matrix by aggregating the results of the modules from the first stage. Generated matrices from stage two are used as input into cluster module and prediction model e.g. Linear Diffusive Model.

Twitter
Stream API
Router
Twitter
Rest API
Twitter Stream
Data Collection
(24hrs x 7days/wk)
User Profile
Extraction Engine
Tweet files (e.g.
“flu/tweets.
2014-Jun-08”
User Profile
MongoDB
Data Processing Engine
Geo-Location
Extraction Module
Tweet Analysis
Module
User Distance
Analysis Module
Tweet Filtering
Module
User Activity
Analysis Module
Tweet Statistics
User – Tweet
Interaction matrix
Cluster module
User Clusters
Geo-tagged flu tweets
CDC ILI (Influenza-like
Illness data)
PDE Modeling
(Linear Diffusive
Model)
Model Parameters

Figure 1. Twitter Data Collection, Processing, and Modeling Framework

Ebola Statistics



For the subset of Twitter data relating to the Ebola outbreak (2014), on the first day of collection (October 5th 2014) 222,839 tweets were collected. We had a total of 24 tweets with more than 1000 retweets with the highest retweet count being 3,457 for an image that was trending depicting that the TV show The Simpsons had predicted the Ebola outbreak back in 1997.

Conclusion

The power that comes from applying big data techniques to social media is unprecedented. We build a Twitter data collection, processing, and modeling framework to demonstrate the feasibility of automating these processes. Our framework is flexible in the sense that newly developed clustering algorithms and mathematical models can be easily plugged in and verified. Our framework provides us the capability to discover and predict epidemiological outbreaks in their early stage. By improving the prediction accuracy, we may be able to track outbreaks in nearly real-time. The system built so far has proven to be a very powerful tool for analyzing big data.

Future Work

- Improve geocoding hit ratio
- Redesign algorithms to handle real-time data analysis
- Improve the parameter estimations of the model

Acknowledgements

This project is supported by NSF and NCUIRE.