Long Term Trend Analysis

Introduction

- Long term trend analysis is a time series of a set of words
- These words may or may not be related
- Instead of looking at the words one at a time we are trying to define groups as topics

Framework

Pre-Processing

- Removed Special characters (such as &,*,@

Clustering

Post Clustering

- We need to identify the words that most represent the cluster

- We calculate average TFIDF for the whole cluster -

To eliminate noise we remove all the words that have a TF-IDF value less than average

- Viewing the tweets as documents and making them as clusters

- Threshold value chosen for cosine similarity-0.05 -

Finding frequency for each term in documents, idf for each document and taking these values for testing them in cosine similarity for all the tweets in data set.

News Results

- Calculate the number of new articles for representative words
- Google news is considered as a source to retrieve the articles
- Only articles during the time preiod of the tweets are considered

Data

- Tweets collected for specific drug names

Time Period : February 21 - April 30

Total Number of tweets: 33,000

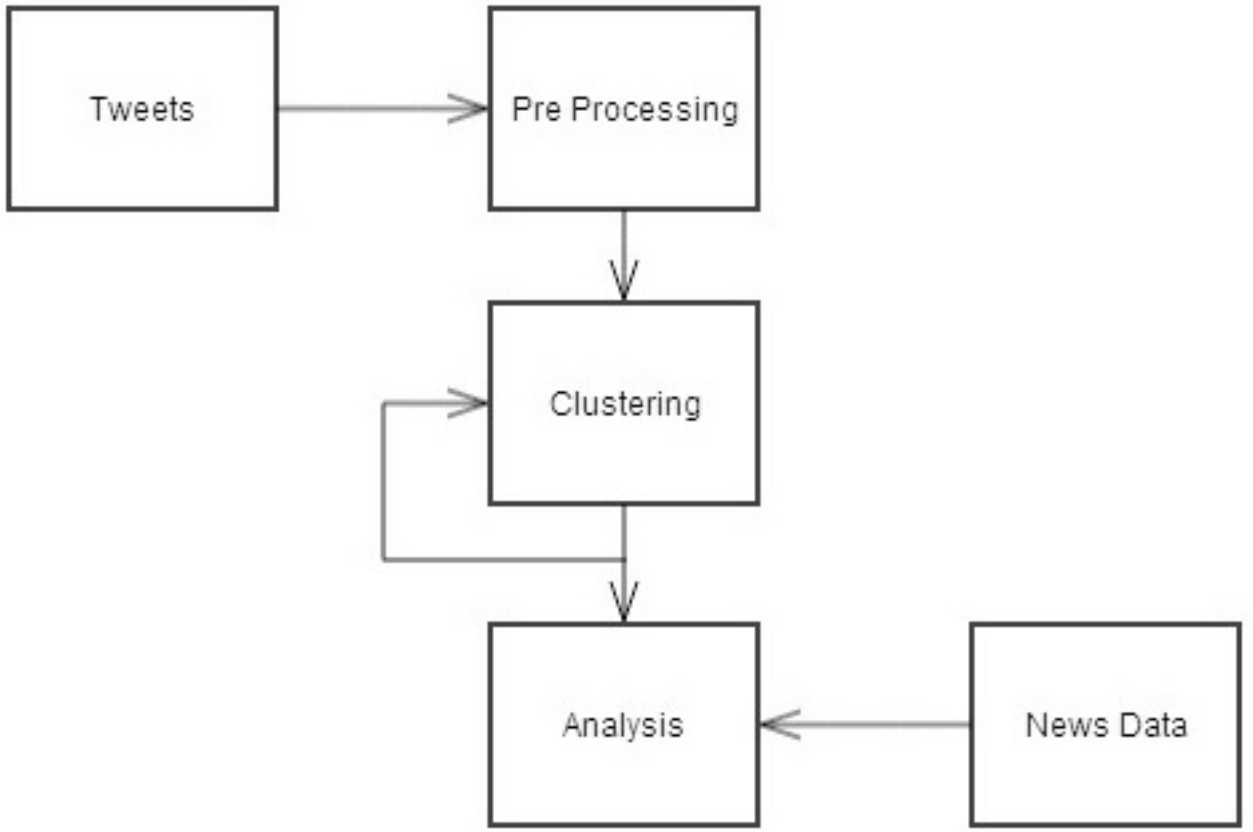
What is topic-clustering?

- Topic clustering-Most representative terms in

different tweets of one cluster are comparing with the occurance of same terms in different clusters. Based on this,we are identifying a topic of terms for both clusters.

- Identifying the co-relation of terms in a cluster versus different clusters,topics and documents.

WORKING MODEL



Analysis of data set

Time series graph for the clusters in both data sets.

Types of data set

- Global data set-

Past 2 months medical(drug) related data from news articles.

Analysis of data set

- Manual analysis-
- Analysis by implemented model-

Manual analysis:

- Manual analysis is done by just noting the occurancy of the representative terms for all clusters by referring the pages that Google website or browser has pulled out over the clusters creation period.

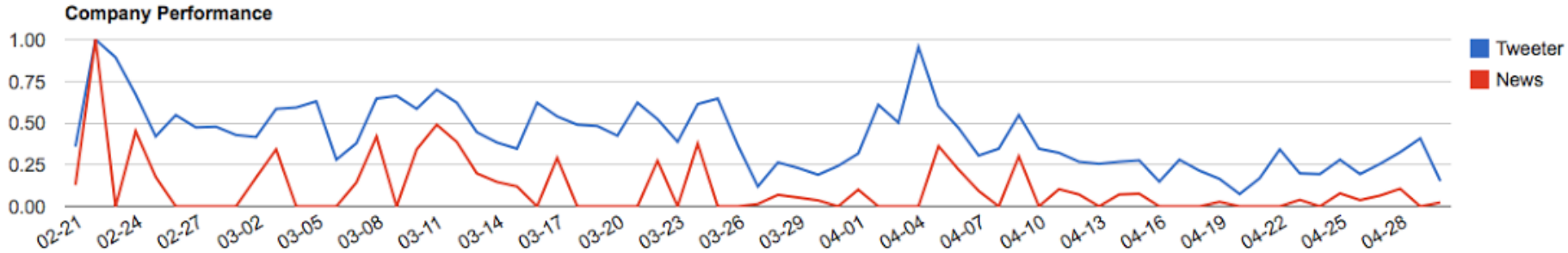
- Drawing term-series graphs for different clusters in both data sets
- Identifying the peak points in all the graphs for both the data sets.
- At the points where peak points are observed,we are looking for the frequent match of user decisions for that particular topic over a period of time.

Analysis by TF-IDF model

- In this, analysis can be done by feeding the data first to the model and trained it by using TF-IDF model and passing the trained data for testing. This testing of data sets is done by criterian values we set in the cosine similarity technique.
- Using K-means for identifying the topic clustering of representative term in the tweets and generating time-series graphs for the clusters in data set.

Graphs screen shot

Cluster1



Words in Cluster

cymbalta depression help hurts

Words in Cluster

Or take cymbalta

Or take cymbalta

@HauteMessFiasco swollen hands, sleepy all day, not enuff sleep at night, achy bones, headaches. feels similar to coming OFF cymbalta to me.

@HauteMessFiasco swollen hands, sleepy all day, not enuff sleep at night, achy bones, headaches. feels similar to coming OFF cymbalta to me.

depression hurts, but cymbalta helps....I think that's what that commercial says

depression hurts, but cymbalta helps....I think that's what that commercial says