**COSC2671 Social Media and Network Analysis**

**Assignment 1:Analysing and Tracking the Sentiments and Topics on Social Media**

Student No: S3714805 Student Name: Vishwa Gandhi

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Table of Contents

**Introduction**

Entity selection

**Data Collection**

Approach of data collection

Data Statistics

**Pre-processing and Data Cleaning**

**Analysis Approach**

Sentiment Analysis

Topic modeling

**Analysis & Insights**

Sentiment Analysis

Topic modeling

**Conclusion**

**References**

**Introduction**

For this assignment I have decided to analysis social profile of India's current Prime Minister 'Narendramodi'. He is among world's most powerful leaders and his work is continuously influencing people all across the world. Being a Prime Minister of world's largest democracy, his actions affects large group of people. It fascinates me to learn about highly followed and influencing personality. Throughout this report I will use data of Tweets mentioning '#Narendramodi' to answer following statements:

1. What is people's overall sentiment towards Indian political leader and a Prime Minister Narendra Modi.
2. What are major topics being discussed on social media related to Narendra Modi.

**Data Collection**

**Approach of data collection**

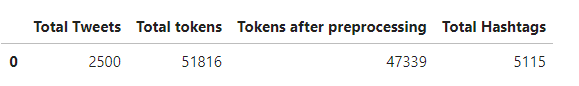
Two widely used approaches for Twitter data collection is REST API and Streaming API. Following are the key difference between these two techniques.

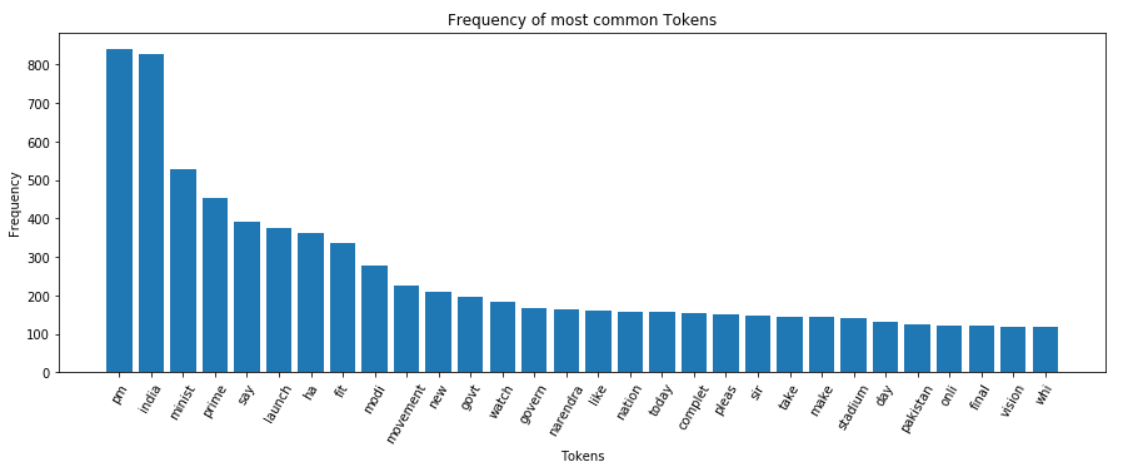
* Implementation of request-response mechanism mainly differentiates REST and Streaming API.
* REST API fetches past data for search query which can be a mention, hash tag or twitter handle by requesting server. It is request and response based approach.
* Whereas, Stream API is a continuous live data fetching mechanism. It retrieves tweets in real time to client server, demands continuous net connection.

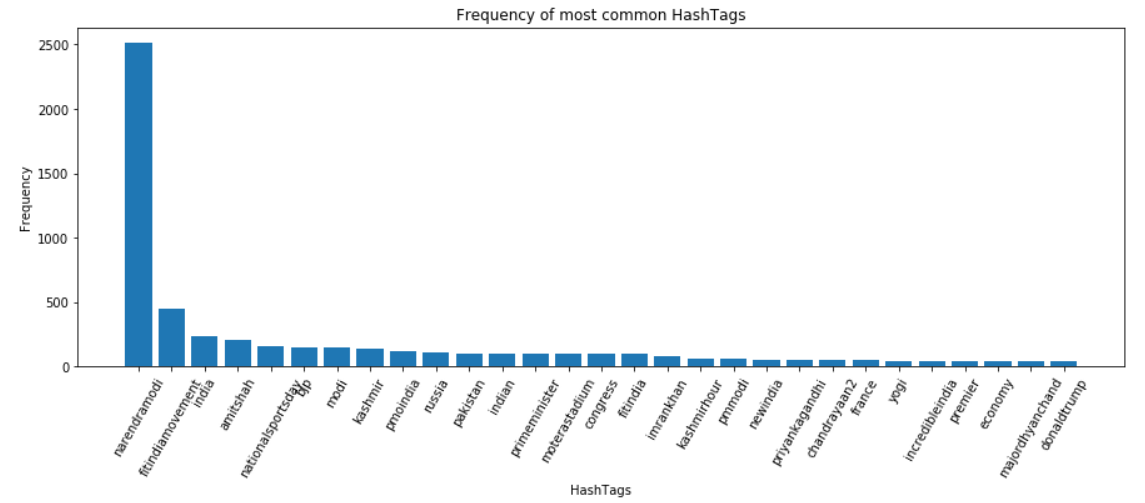
I have used REST API to fetch data from Twitter. According to the scope of my analysis, live data fetching is not required at the moment. Information facilitated by REST API is sufficient enough to gather meaningful insights. Hence, I have used REST API to fetch data from Twitter. I have collected 2500 tweets with '#narendramodi' in text using tweepy package and stored it in a JSON file. This JSON file will be used to access data.

**Data Statistics**

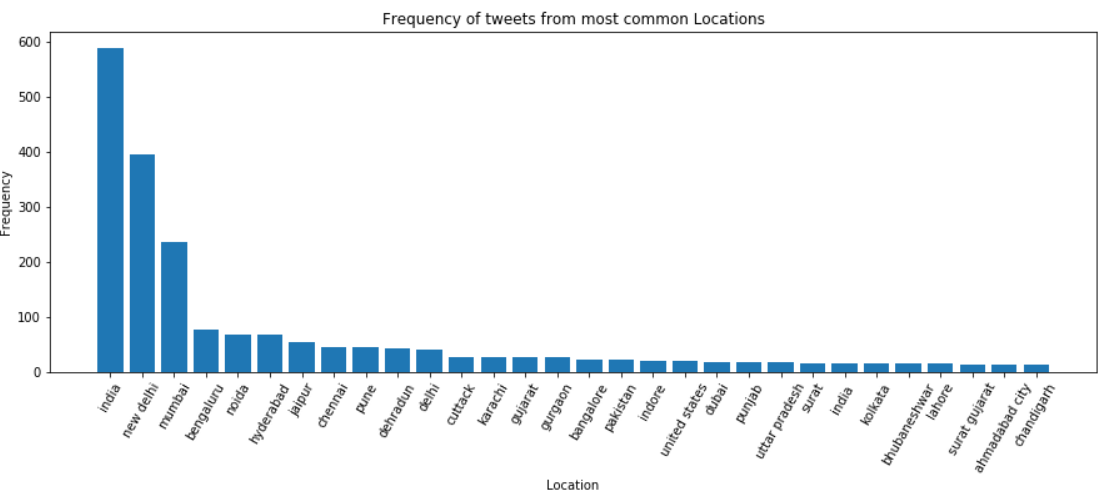
* From total 2500 tweets retrieved, 51816 tokens are generated from tweet text. This also contains redundant tokens. Moreover, 5115 hash tags are also present in data.



* Below given bar plot provides frequency count information of most common 30 tokens used in our collected data. Most of these terms are related to government operations. It can be said that major discussion happens about nation and government related topics. Few key words like fit, launch, vision, stadium and movements can be mapped to ongoing movements related to fitness awareness, recent launch activity happened at NASA India. 
* Below given bar plot gives information regarding most frequently used hash tags. As hash tag has become new effective way of learning an event, it is very useful to study trending hash tags. These hash tags are about few leading politicians in India, fitness movements, construction of a motera stadium and India's activities related to India's domestic & international affairs.

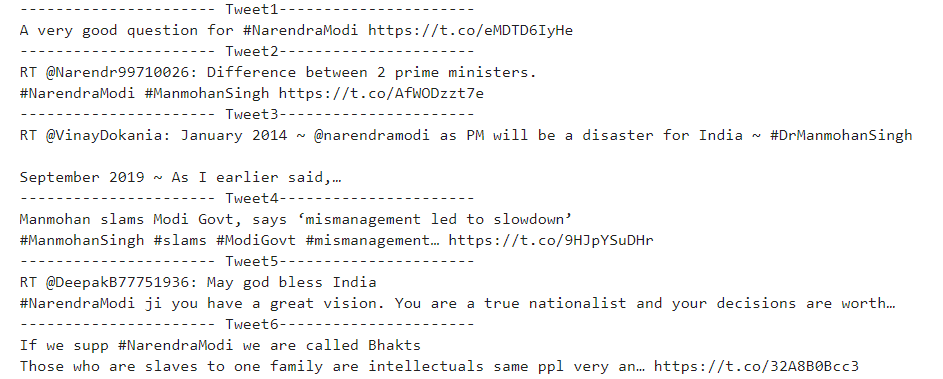


* Apart from this data, geographical affinity of user can also be analyzed as shown in below given most common region based tweet frequency bar plot. It can be clearly seen that most of the tweets are from users in India. Also, some tweets are by users from United States and Dubai.



**Pre-processing and Data Cleaning**

Tweets text data is generally in informal writing which can have typos, abbreviations, multiple languages, special characters and other random words. To gain insights from such data, it needs to be cleaned to remove such terms. Given figure is displaying first 6 tweet text data. It can be noticed that many characters and words like https,.. , ~, A, RT are present which do not contribute to our analysis. These unwanted tokens are removed to smoother further processing.



In order to preprocess data to prepare it for advance processing, I have used following steps:

**Analysis Approach**

**Sentiment Analysis**

In this section, I have applied unsupervised **'vader sentiment analysis'** which generates sentiment polarity score for each tweet. Twitter data is in unsupervised and unstructured manner. It is tricky to analyze sentiment of such data as a sentence may contain multiple sentiments, slangs and sarcasm. Vader has following benefits which made me use it for this assignment:

* It is a lexicon based analysis approach which uses social media lexicon set. Hence, it is more relevant to our social media analysis.
* Vader considers both polarities of the sentence to take polarity shift into account. Hence, it generates efficient results in case of a sentence having multiple sentiments.
* It generates compound scores between -1 to 1 for entire tweet which depicts bigger picture and gives better sentiment mapping.

**Topic Modeling**

I have implemented Latent Dirichlet Allocation (LDA) technique to discover topic from twitter data.

* It is suitable for unsupervised data. LDA is a probabilistic model to discover topics from data given.
* It is easy and faster to train LDA models and its performance can be improved by tuning number of topics to be discovered.
* It iterates over entire collection of data to check topic assignment for every word. These features allow improving the result generated by the function.
* It is also possible to derive proportion of word in a document. Moreover, spread of each sentence by word count can be inferred.

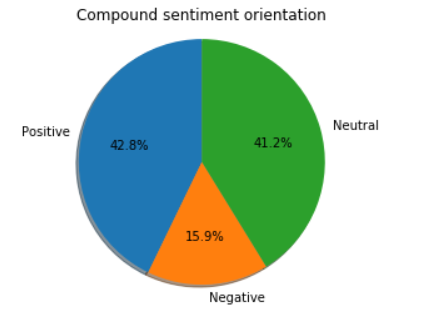
Hence, considering the nature of problem in this assignment, LDA seems a good choice to discover trending topics being discussed.

**Analysis & Insights**

**Sentiment Analysis**

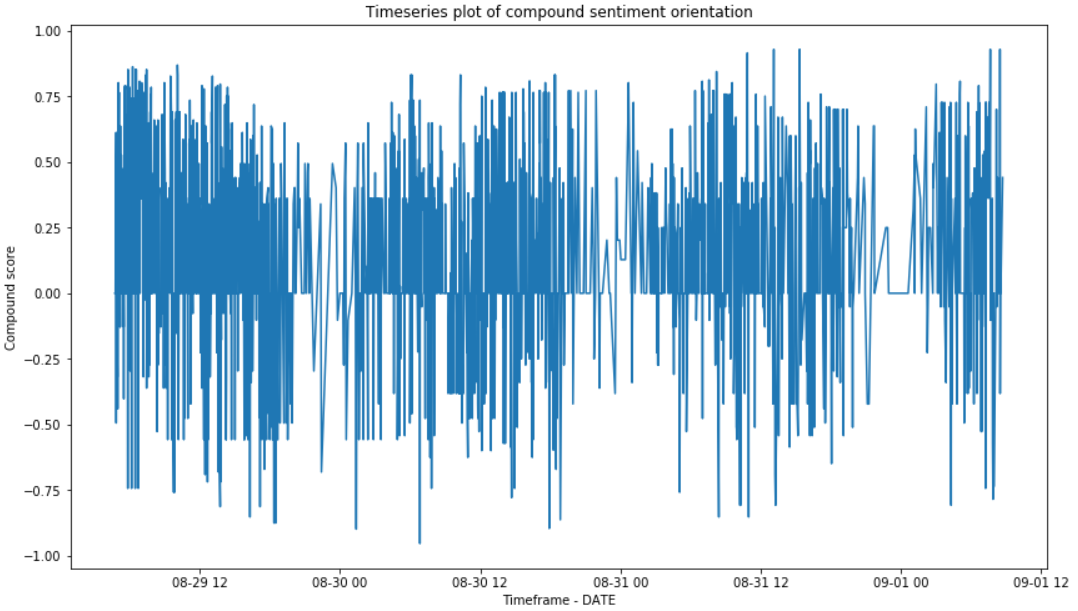
Applying Vader algorithm on preprocessed text data has generated polarity scores for each tweet. It generates 4 types of scores for each query: {positive, negative, neutral, compound}. To draw conclusion about overall sentiment of entire data collection, compound score is a best measure. But to learn sentiments at granular level, rest of the score types give detail information. As of now, I have used compound score to find the solution.

* To learn about overall distribution of the compound score, consider following pie chart :



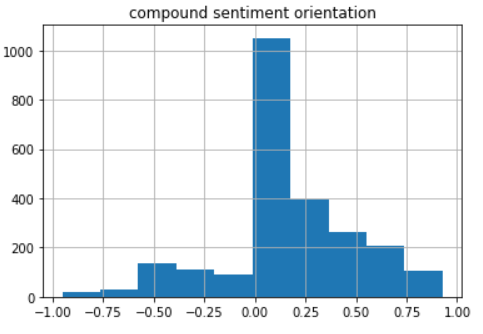
As it can be seen, there is very low proportion of negative sentiment present in tweet data. Proportion of positive and negative is nearly same, having around 2% more positive sentiment. It can be concluded from the pie chart that there is positive sentiment in general public towards Indian politician and Prime Minister 'Narendra Modi'.

* Consider below given time series graph to analyze any time based sentiment demonstrated due to occurrence of particular event.



There is no sudden spike in time based compound scores. Hence, there is no event happened in the duration for which data has been collected which had led to significant sentiment change. It is evident from the graph that generally for entire duration, intensity of positive sentiment is higher than negative score. Therefore, we can say that there is possible sentiment among public towards Narendra Modi.

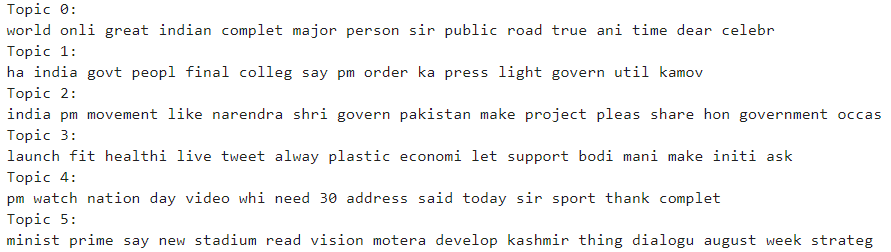
* Below given graph gives insight that overall sentiment is positive. Degree of positivity is not very strong and it is more towards neutral sentiment. But by considerable shift towards +1 makes it overall positive.



**Topic Modeling**

I have implemented LDA technique to discover 6 topics from the tweet data. LDA parameter 'number of topics' is tuned by trial & error approach. In case of 6 topics, it gives considerably tightly closed word distribution in a topic to get the meaningful domain out of it and also it is big enough not suppress any topic in finding.

* Following topics were generated by LDA with 6 topics:



* To better understand word distribution in topic, consider given below word clouds.



* These topics generated gives following information:
* Topic 1: It has words like government, Pakistan, projects, security, submarin, independ. These discusses about India's relation with world, Foreign affairs, naval security.
* Topic 2: More occurring meaningful words are light, placement, college, govern, people. Analyzing these words, it can be said that it is mainly discussing about employments, job creation and related governing decision.
* Topic 3: This topic has more words related to India, pm and movement. Apart from that, it has lot of noise.
* Topic 4: This topic is strongly indicating towards health, economical, scientifically growth, environment concerns. It can be generalized as better living standards, climate and economy. It is due to fitness promoted by Narendra Modi, launch of space rocket.
* Topic 5: it is about PM's video address on independence day which was in last few days
* Topic 6: This is about motera stadium event happened

**Conclusion**

Following conclusions can be drawn from the social media analysis performed on twitter data for 'narendramodi'.

* Sentiment Analysis: As per the results of vader sentiment analysis, overall sentiment orientation is positive towards Mr.Narendra Modi.
* Topic modeling: I found 6 distinct topics from LDA and some of the interesting ones are fitness movement, motera stadium, employment, government discussions.

**References**

[1] O. Genc, "The basics of NLP and real time sentiment analysis with open source tools", *medium.com*, 2019. [Online]. Available: https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe. [Accessed: 26- Aug- 2019].

[2] M. Sarkar, "Using LDA Topic Models as a Classification Model Input. [Online]. Available: https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html [Accessed: 27- Aug- 2019].

[3] A Ahemed, "Complete Guide to Topic Modeling. [Online]. Availablehttps://nlpforhackers.io/topic-modeling/ [Accessed: 27- Aug- 2019].

[4] J. Chan, " Social Media and Network Analytics, *Class Notes (Internal)* [Online] [Private]. [Accessed: 21- Aug- 2019].