Project 4

https://informationretrievalfall2020.herokuapp.com/

(DISSECTING TWITTER DATA TO ANALYZE GOVERNMENT AND PUBLIC ATTITUDE TOWARDS COVID GOVERNANCE)

1. Introduction

The goal of this project is to analyze our Twitter data and find the public attitude to the COVID-19. Our tweets are from 3 counties: the USA, India, and Italy. The range of the language also includes English, Hindi, and Italian. In our project, we will build a web application that provides insights about COVID governance and other statistical information about it by analyzing our tweets data. There are several parts that we need to be concerned about: First, we should define several rules to calculate the influencer score and use it to improve the tweets ranking. Second, we need to choose models or methods to achieve the analysis of different topics. The sentiment analysis will be used on the retrieved results and the statistical results will show to the user. The web application also provides the user with faceted search options.

2. Dataset

Data of this project was collected from Twitter using the Twitter API. They are in JSON files which are created in Project 1:

- The language of the tweets also ranges in these country specific languages (English, Hindi and Italian)

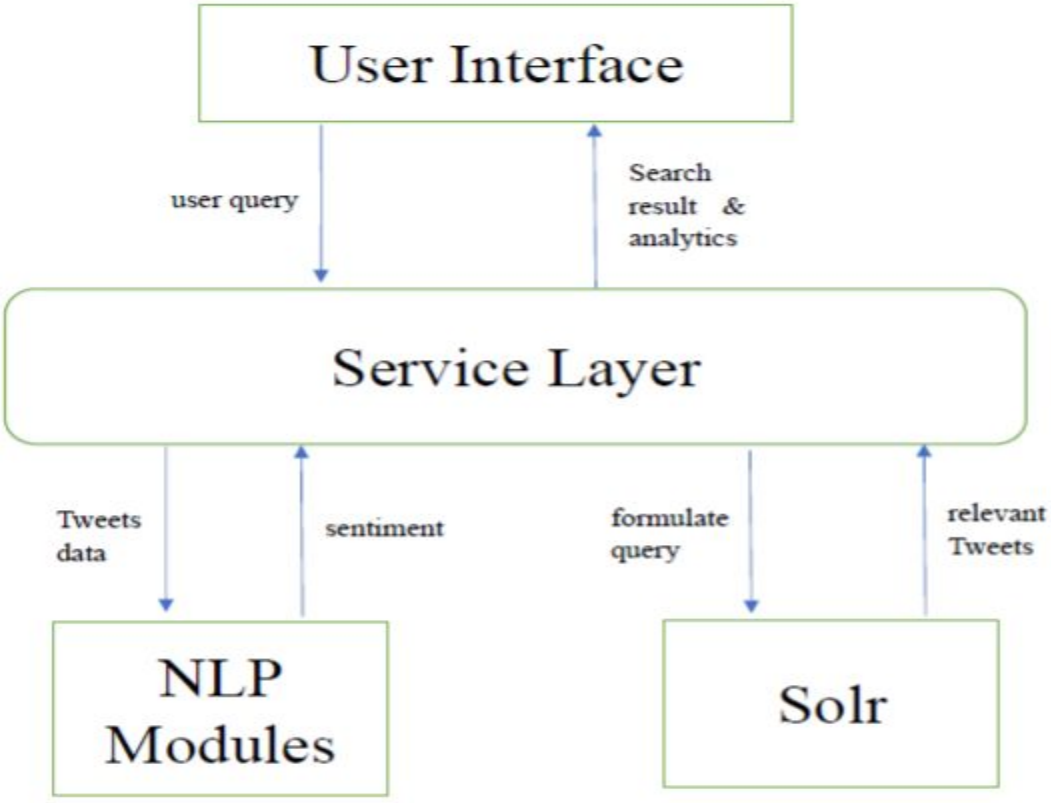
- Tweets posted in 5 consecutive days focused on reactions of general public to government's policies on COVID

3. Architecture Diagram

The application is set up in this architecture.

The User Interface is the user component that captures the user query and sends it on to the Service Layer. It also receives the response from the service layer. The search results, as well as graphical views of the analysis data, will be shown.

The Service Layer receives the user query and queries the Solr instance after adding filters. When receiving the data from Solr, it applies the sentiment analysis, retrieves news and youtube videos on the retrieved content and receives the score.



4. Requirements:

Requirement 1: - Influencer score is calculated using this formula.

If the User is verified, Score = (Number of Likes + Number of retweets)**1.5

Else If the text contains coronavirus or government related policy as a topic, Score = (Number of Likes + Number of retweets)**1.2

Else the Score = (Number of Likes + Number of Retweets) for normal users and the verified must be untrue.

Using the above formula, the covid tweets and government policy tweets are boosted and if it is from POI then they are boosted more.

Requirement 2: Content/Topic Analysis

Topic Analysis is done by considering each and every tweet present in the full_text field, gensim and LDA models are used to analyze the topics. We have used pyLDAvis to display the analysis in a static HTML page where the user can analyze most frequent topics that are discussed. The algorithm of topic analysis is as follows.

1) Text Cleaning (full_text) is done by using nltk, removed all the stopwords from all the different languages and tokenized the data.

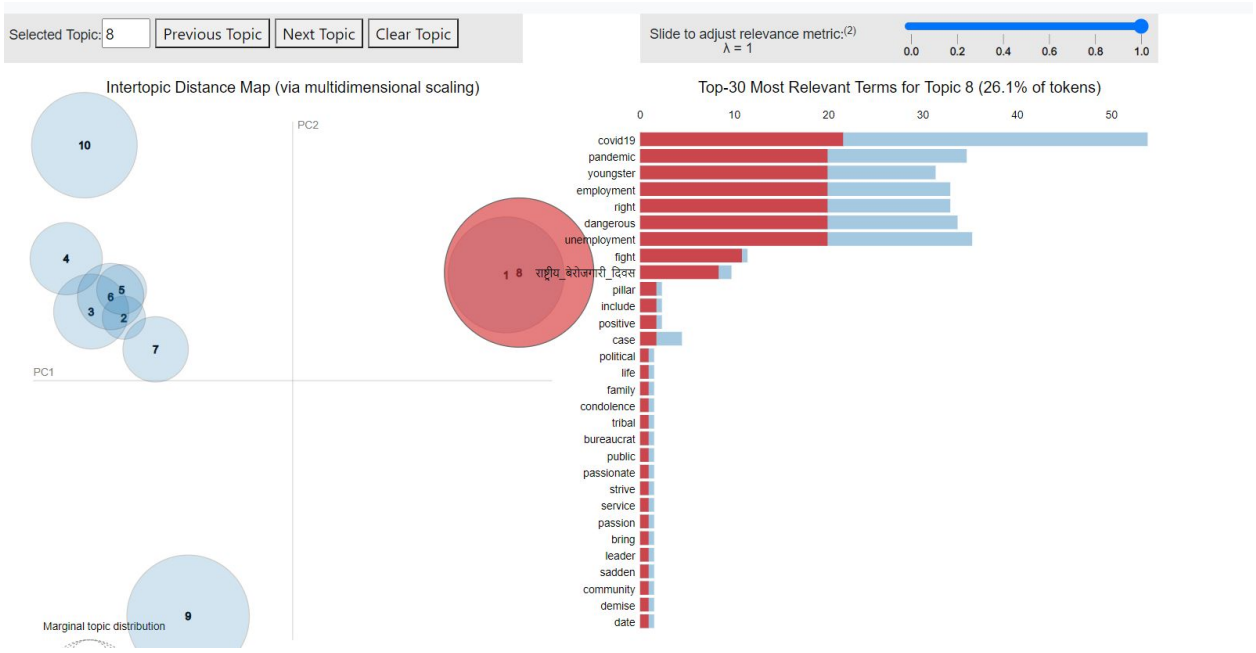
2) Prepare text for LDA, then add to a list.

3) Create a dictionary from the data, then convert to a bag-of-words corpus and save the dictionary and corpus for future use.

4) We pick the number of topics ahead of time even if we're not sure what the topics are.

5) Each document is represented as a distribution over topics.

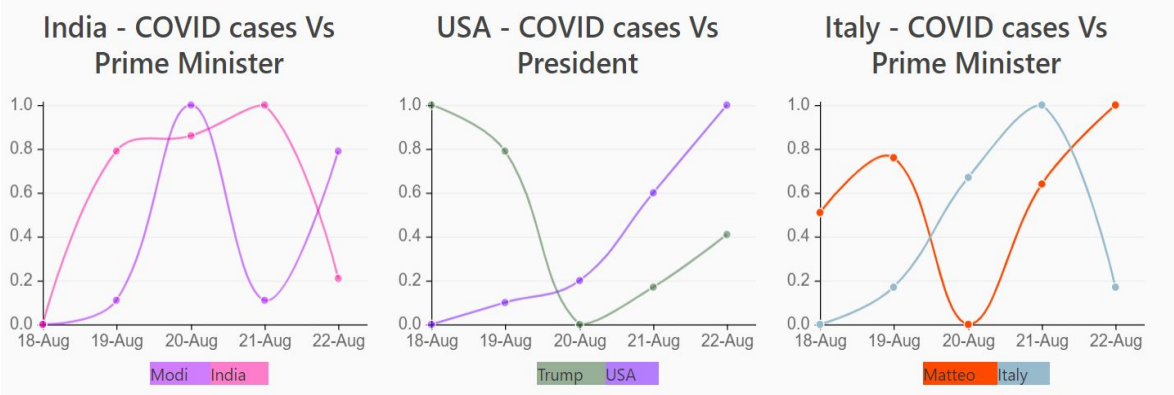
6) Each topic is represented as a distribution over words.



Correlation analysis:

Correlation is analyzed between Covid cases in each country, This is achieved by picking up the most important person of the country (President/Prime Minister) and data is normalized since tweets are not keeping up to the increase in the number of covid cases in each country. The data is maintained in a sqlite database and displayed at the point of retrieval.

Interesting facts are obtained from the 5 day period (18-Aug-23-Aug), In India and USA, prime minister and president had made tweets related to covid, as per the number of increase in cases. But in Italy the reverse has happened, when the number of cases are low the tweets related to covid are high and when the cases are high, tweets related to covid are low.



Requirement 3: Insights/Analytics.

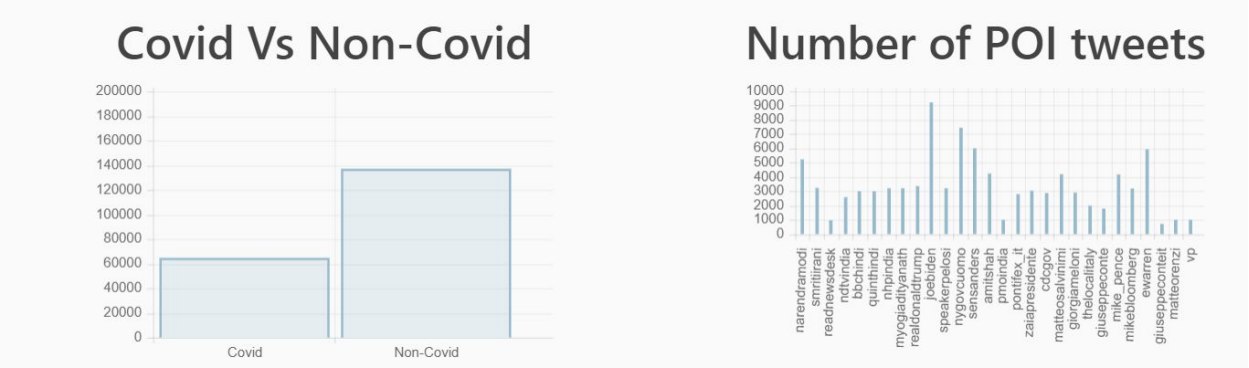
1) Sentiment Analysis: Sentiment analysis is done using vaderSentimentAnalysis package and the score in the website is obtained while they are loading, different languages are translated to English and then sentiment score is calculated. If it is 0 – neutral, > 0 – Positive, < 0 - Negative.

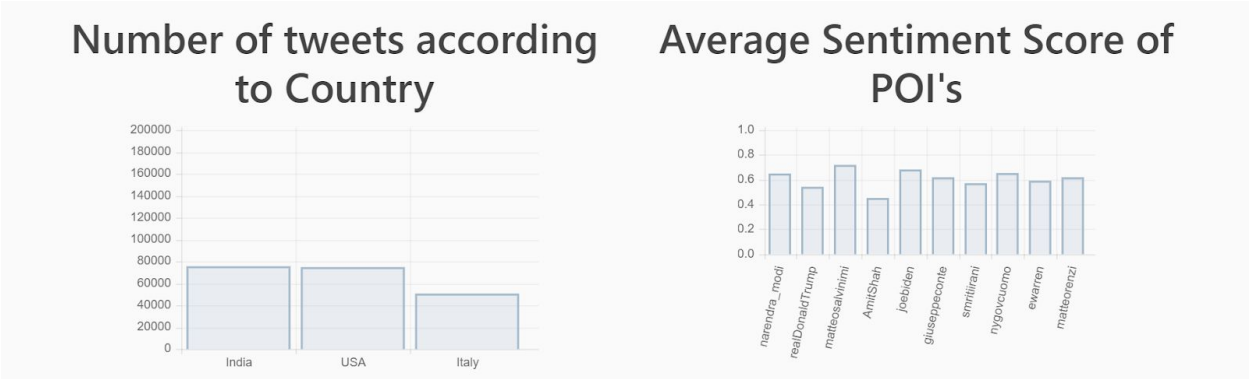
2) News Articles: News articles are fetched using newsapi, it has a 30 day limit and a rate limit of 100 requests per day. The search query is directly passed into the news api and translated across all languages(en,it,hi).

3) Youtube Videos: Youtube Videos and tweet text are correlated if a user mentions a youtube video in the twitter data, then that youtube video is displayed and also if the videos are displayed according to the topics that they are talking about in full_text and related to search query.

4) Visualization: Covid vs Non-Covid, a bar graph is represented on overall text data if the text is related to covid or non-covid. Number of POI tweets – various numbers of POIs are extracted and a bar graph is represented according to the number of tweets.

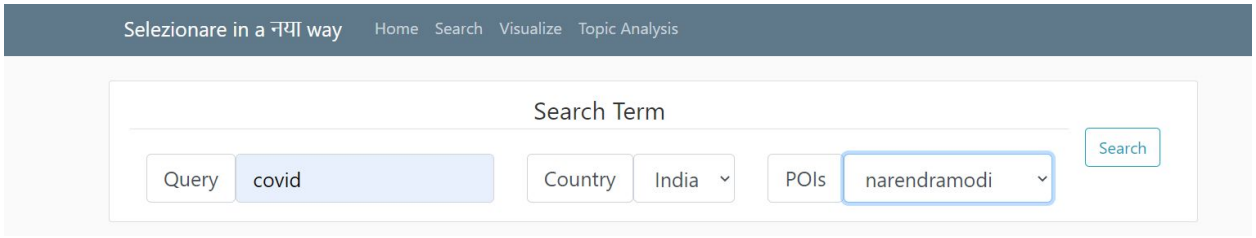
Number of tweets according to country, which is represented in a bar graph.





Requirement 4: Multi faceted search

Multi faceted search is implemented where the user can select a country a poi and search for information or he could select a single country or single poi and search for data. All possible combinations of country, poi name and search term works.



Additional Comments:

The Initial Home page consists of trending posts which are sorted by influencer score and the score that is obtained by BM25 with k1=1.3 and b=0.6 which gave us the best results.

5. A Short Note on Architecture

Front end - HTML, JS, CSS

Back End - Apache SOLR, News API, Youtube API

Database - SQlite

Framework - Flask

6. Future improvements

In order to make the web application more perfect and get more accurate results. some improvements can be brought to the system in the future:

- Add a date-based search. The user can select a specific date or a range of dates to search a particular query.

- We can try to use Pseudo relevance for query optimization. When completing the original query, the user can select the tweet on which he wants (for example the first 100 tweets) to expand the original query, and then the final result will be more accurate. Some results which are missed in the initial query can be finally obtained.

7. Conclusion

Implementing this project, we understood how to build an end-to-end search engine and how it works. We can also index, extract, and analyze the data. We got a chance to implement many concepts such as faceted search, sentiment analysis, and correlation analysis in our project. Different models like Gensim and LDA models were studied and used to achieve our topic analysis part. After analyzing the data. We found that the COVID-19 is still a topical issue around the world, the number of daily tweets posted by the government has a strong connection to the number of increase in cases. The public has been closely watching the trend and news of this virus.

References:

1) https://towardsdatascience.com/topic-modelling-in-python-with-nltk-and-gensim-4ef03213cd21
2) https://www.youtube.com/watch?v=MwZwr5Tvyxo&list=PL-osiE80TeTs4UjLw5MM6OjgkjFeUxCYH
3) https://blog.ruanbekker.com/blog/2017/12/14/graphing-pretty-charts-with-python-flask-and-chartjs/
4) https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
5) https://lucene.apache.org/solr/guide/8_5/
6) https://flask.palletsprojects.com/en/1.1.x/
7) https://sqlite.org/docs.html
8) https://newsapi.org/docs
9) https://pypi.org/project/langdetect/
10) https://py-googletrans.readthedocs.io/en/latest/
11) https://towardsdatascience.com/a-step-by-step-tutorial-for-conducting-sentiment-analysis-a7190a444366

Contributions:

Krishna Teja Mellacheruvu (kmellach, 50336871) – Developed the front end of application, Requirement 1, Requirement 2.

Venkata Narayana Rohit Kintali (vkintali,50336890)– Indexed the data and developed back end for application, Requirement 3 and Requirement 4.

Yanting Li (yli288, 50244287) – Report part and the improvement part for all requirements.