

# CSE 435/535: INFORMATION RETRIEVAL

Fall 2020

University at Buffalo

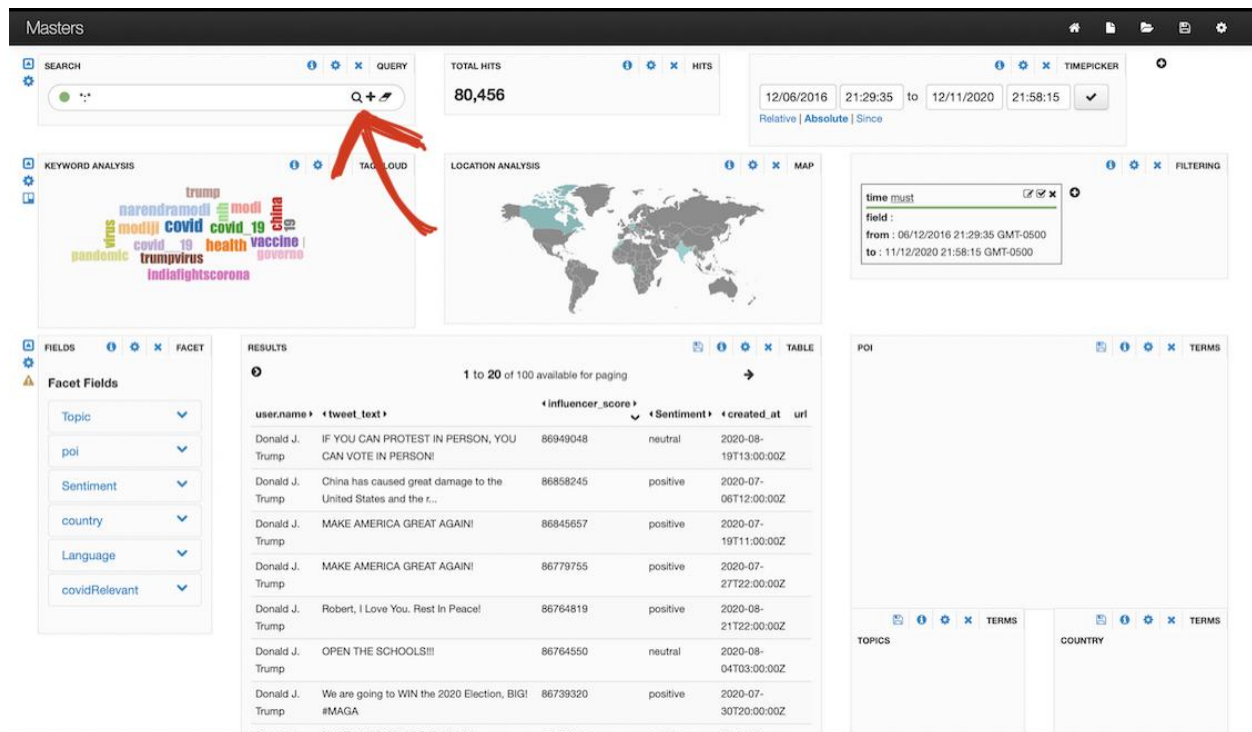
## **PROJECT 4: DISSECTING TWITTER DATA TO ANALYZE GOVERNMENT AND PUBLIC ATTITUDE TOWARDS COVID GOVERNANCE**

Deadline: 11<sup>th</sup> Dec, 11:59 PM EST

By :- Sudhir Yarram(50305566)

Manasa Sai Challa(50356441)

**\*Incase you dont see any panels filled up upon opening the dashboard, click on the search button on top left corner\***



## Overview

In Project 1, we have collected twitter data that fulfilled our requirement from the twitter API and then tokenized and indexed it using Solr. We specify the requirements of the data we are looking for in a python script and then use that to crawl twitter through the twitter API. The data returned from our python script is returned in the form of a JSON file. In Project 2, a given dataset containing the tweet is preprocessed and then the resulting is tokenized. We create a posting list for the tokens generated. We then apply Boolean Query processing, and then Document-at-a-time AND query. After processing the data like such, our dataset is effective enough to create a multilingual IR system. In Project 3, we implement the vector space model and VM25 model, we use these to improve the search results based on understanding of the model, implementation, and evaluation. In Project 4, we unify all the individual tasks performed in the previous projects and unify them to create our very own multilingual IR system.

## Requirement 1 – Social Network Analysis

We calculated the influencer score for each user in the dataset by adding the following metrics together: Retweet count, likes, and follower count. After calculating the score we added an object 'influencer\_score' to each user field in the json file. We can now order search results by descending or ascending count of influencer score.

## Requirement 2 – Content/Topic Analysis

In order to perform topic analysis on our tweet text, we used LDA with Gensim. The data is first converted in a dictionary then to a bag of words corpus, we saved this for future use. Then we choose how many topics we want as a result from the dataset. It gives out different topics in each tweet, we further integrated this into our tweet data and added it to our facet field. This can be further singled out for a specific country as well.

## Requirement 3 – Insights/Analytics

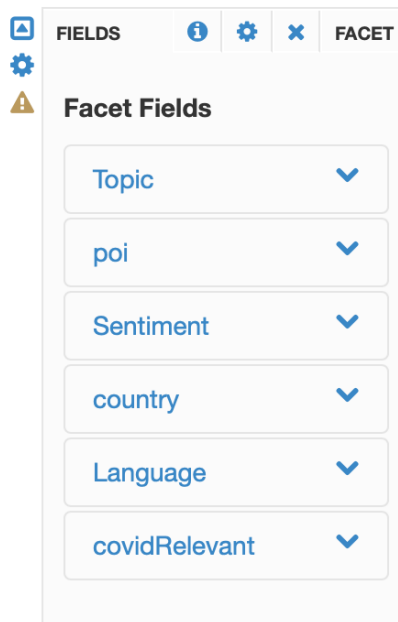
Insights are used to represent data in a more visualized manner. The things implemented in requirement 1 and 2 are visualised in different ways such as piecharts, tagcloud, heatmaps, maps etc. Additional implementation of sentiment analysis, location analysis and keyword analysis, time analysis were also carried out. For sentiment analysis, the text was first cleaned using regex function, followed by calculating sentiment value using the TextBlob library. If  $\text{analysis.sentiment.polarity} > 0$  then the sentiment is positive, if  $\text{analysis.sentiment.polarity} == 0$  then the sentiment is neutral and if  $\text{analysis.sentiment.polarity} < 0$ , it means the sentiment is negative. A sentiment field is then added to each user field in the json file. Location, keyword, time of the tweets were already present in the json dump, so we used charts and other visualization techniques to correlate these values with other fields and presented them.

## Requirement 4 – Faceted Search

We used banana dashboard to visualize the data sourced. The left side of the page has the faceted search functionality, which include fields like topic, poi, sentiment, country, language etc.

There's search bar on the top of the page, to the right and bottom of the page you can see several graphs, pie charts and other images depicting different relationships between fields. In the center of the screen you can see the search results ordered by influencer score by default.

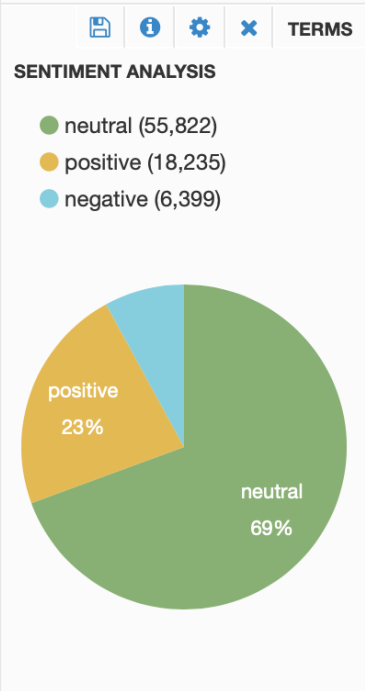
### Facet Fields search



The image shows a sidebar panel titled "Facet Fields" with a warning icon. The panel has a header bar with "FIELDS", an information icon, a settings gear, a close "X" button, and a "FACET" button. Below the header, the "Facet Fields" section contains a list of six fields, each with a dropdown arrow: "Topic", "poi", "Sentiment", "country", "Language", and "covidRelevant".

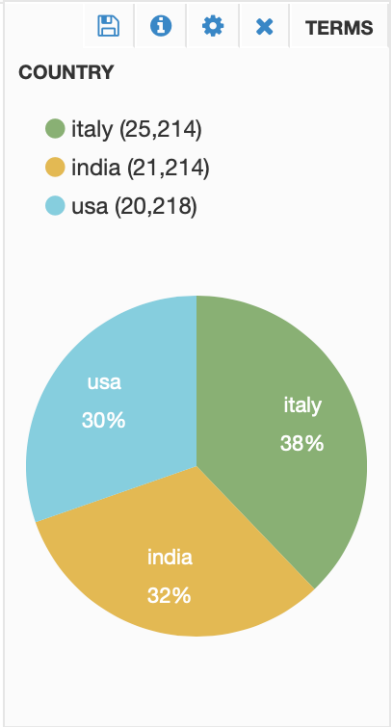
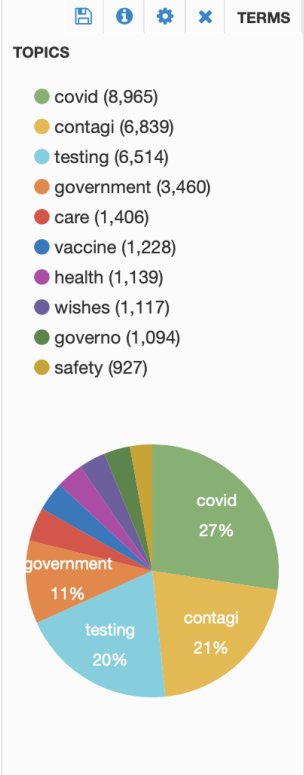
FIELD	SETTINGS	CLOSE	FACET
<b>Facet Fields</b>			
Topic	▼		
poi	▼		
Sentiment	▼		
country	▼		
Language	▼		
covidRelevant	▼		

Sentiment analysis

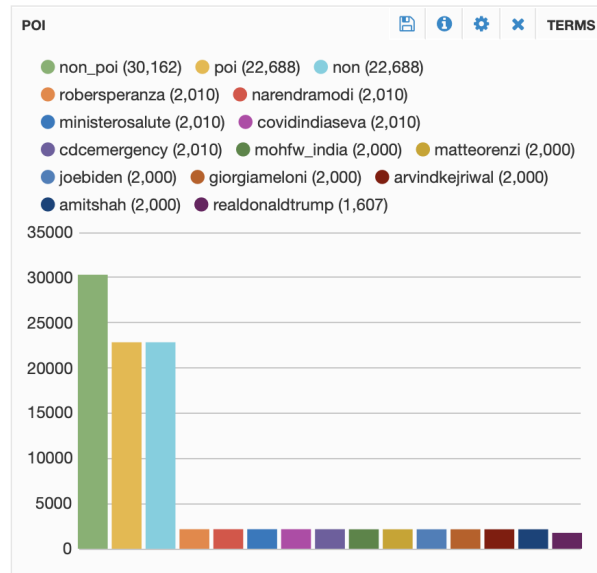


Countries

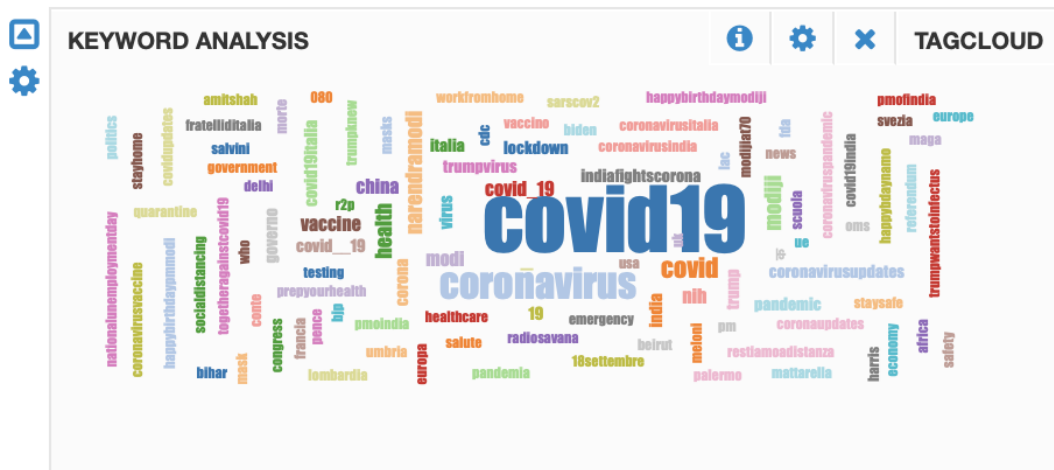
Topic Analysis



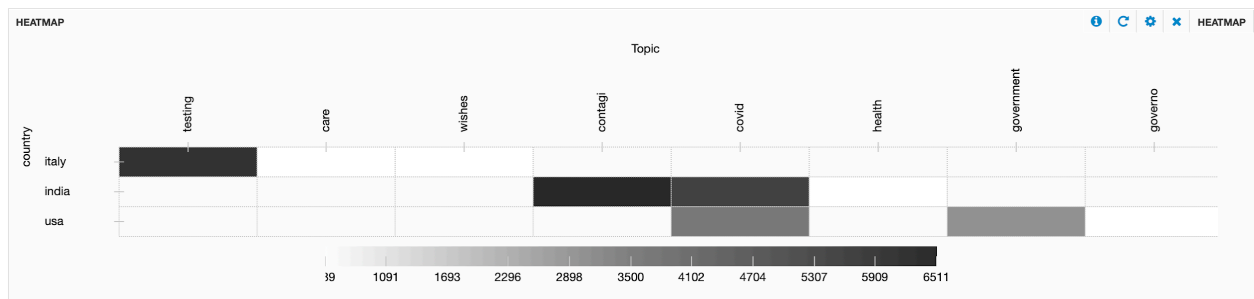
## Persons of Interest



## Topic Analysis



## Heatmap



## Location Analysis

