

PROJECT 4: Complete Search and Analytics Solution based on dissecting twitter data

Authors :

(Team Squad)

Dhanashree Solanke - ddsolank

Sai Ram Prasanth Tatiraju - sairampr

Uday Bhaskar Reddy Bijjula - udaybhas

Sandeep Kothapally - skothapa

Abstract

The project aim is to gain experience of building an end-to-end IR solution which provides wide-range of knowledge about a particular topic, including relevant tweets and analysis results. We started with collecting twitter data based on social issues(social unrest, politics, environment & crime) across five different cities (NYC, Mexico city, Paris, New Delhi & Bangkok). The language of the tweets also ranges in these 5 city specific languages. Then we created the BM25 Solr schema and indexed the corpus of tweets. We improved the IR System Performance using query parsing and various other query boosting techniques. Post that, we created a Web stack which would display the query results along with analytics. Finally We performed Time series analysis, Comparison across the cities, Sentiment analysis and Faceted search to build a solution that provides insight related to social conversations on important societal issues

Dataset:



We created a corpus of over 3 lakh tweets using streaming twitter API (tweepy library in python). We collected twitter data based on social issues (social unrest, politics, environment & crime) across five different cities, namely


- NYC
- Mexico city
- Paris
- New Delhi
- Bangkok

The language of the tweets also ranges in these 5 city specific languages (English, Spanish, French, Hindi and Thai). Then we indexed the twitter data using Solr BM25 model.

Statistics

Last Modified: a day ago
Num Docs: 312034
Max Doc: 347493
Heap Memory Usage: -1
Deleted Docs: 35459
Version: 847
Segment Count: 39

Optimized: 
Current: 

 optimize now

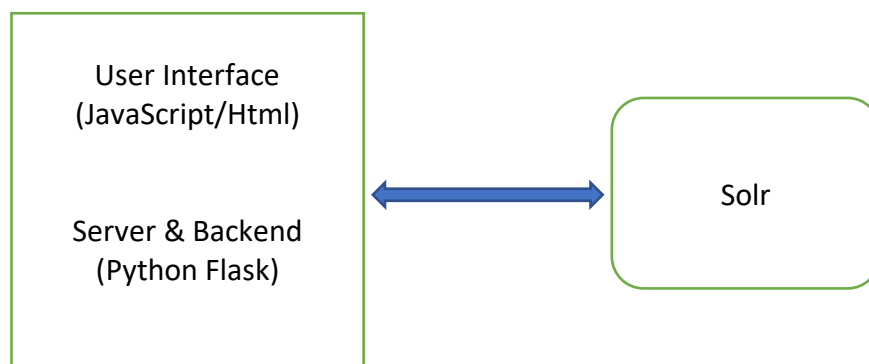
IR Model:

We Implemented a BM25 model schema on Solr-6.6.5 (The default model implemented by Solr) and indexed the 3 lakh corpus of tweets. Then to boost the performance and precision of query results, we did query processing. Broadly, we did query boosting using the dismax qf parameter, by introducing a list of fields and assigned a boost factor to increase that particular field's importance in the query. Also we included a lot of common synonyms across all the five languages in the synonyms.txt file to further improve the results performance.

Analytics and UI:

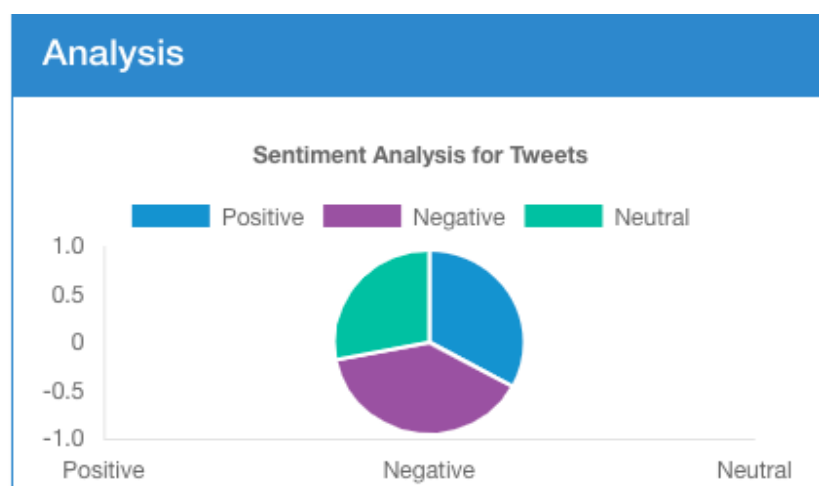
Technology Stack

- We used Python Flask Framework for backend and server.
- JavaScript and html to render the UI.
- Used Google cloud to host the project.



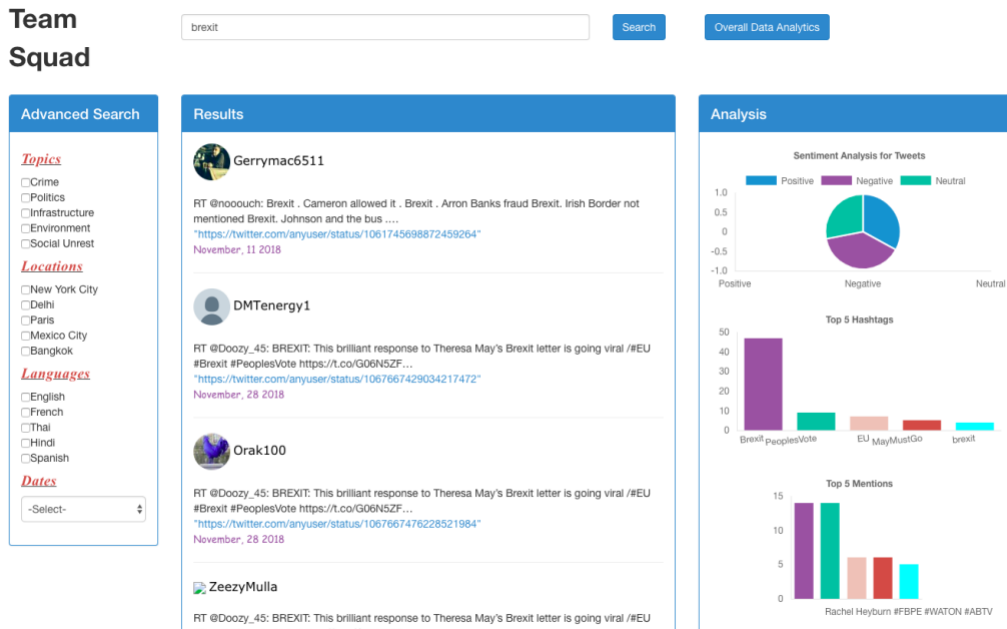
Sentiment Analysis:

Sentiment Analysis (opinion mining) is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. For this project we used textBlob python library for processing textual data, and performing sentiment analysis. To render the pie chart we used Chart.js library.

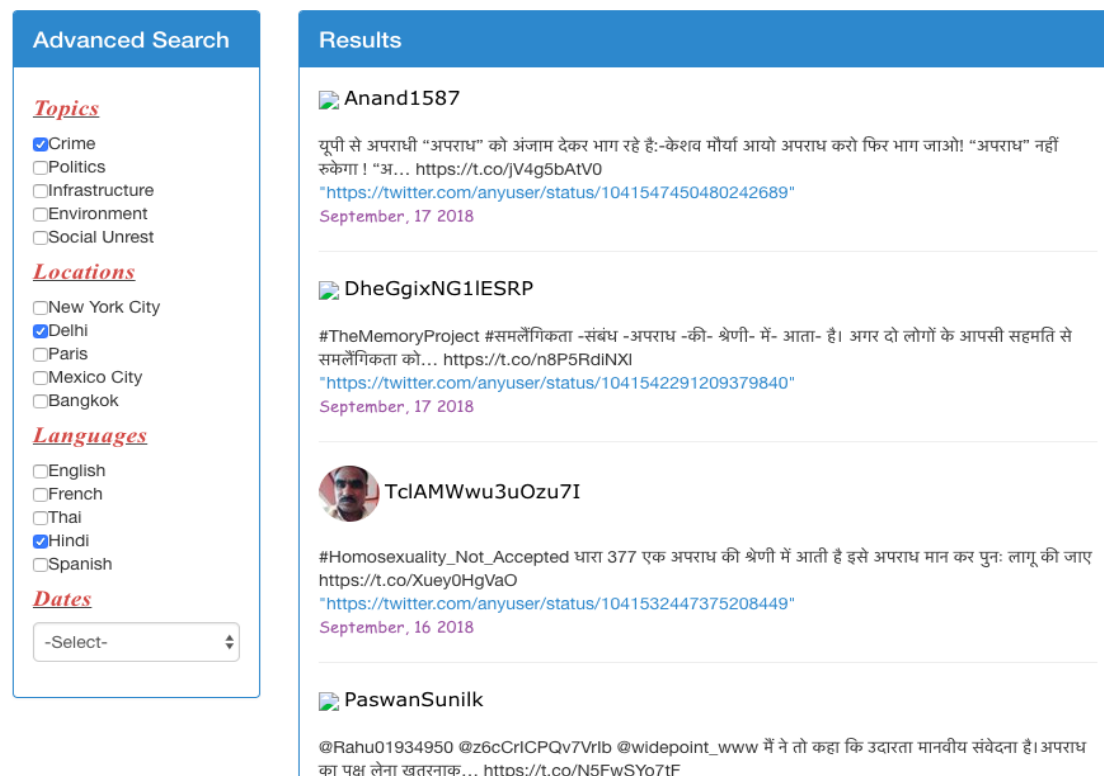


Search Results Page:

Following pics depict the search results page in the application with left pane having advanced search options, middle pane displaying the tweet results, right pane showing the sentiment analysis, top 5 hashtags, top 5 mentions for the search results.



Team Squad



Time series Analysis:

Advanced Search

Topics
☒ Crime
☐ Politics
☐ Infrastructure
☐ Environment
☐ Social Unrest

Locations
☐ New York City
☒ Delhi
☐ Paris
☐ Mexico City
☐ Bangkok

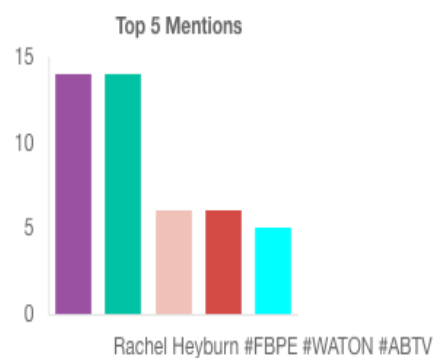
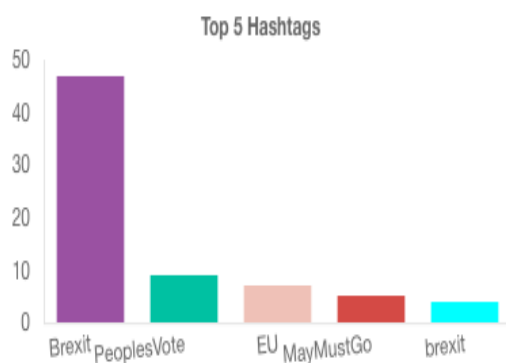
Languages
☐ English
☐ French
☐ Thai
☒ Hindi
☐ Spanish

Dates
✓ -Select-
Last week
Last Month
Last 6 Months

In the advanced search we added date selection filter, to filter out tweets based on the time frame like (last week, last month, past 6 months). All the analysis and graphs were displayed taking time frame into consideration.

Tweet Analysis:

For the user input query ,from the retrieved results we displayed top 5 hashtags and top 5 mentions bar charts from the results retrieved. To render the bar chart we used Chart.js library.

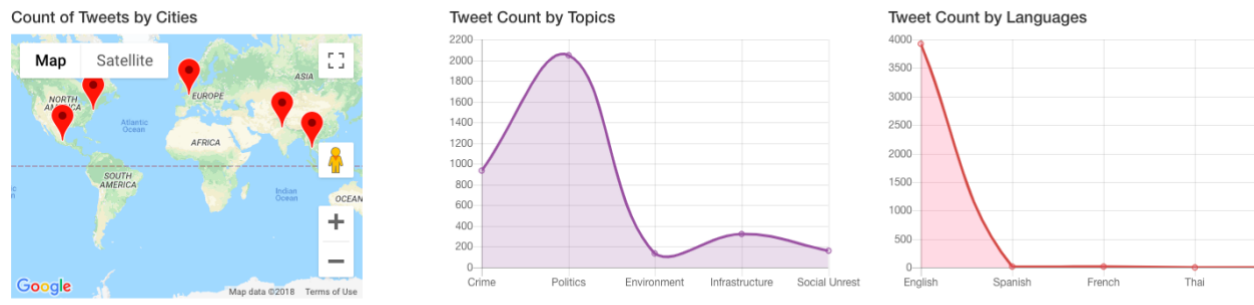


Comparison by category using Faceting:

- City
- Topic
- Language

For the user input query, we displayed the count of tweets by cities, topics and languages using Solr faceting.

Overall Analysis



Member contributions:

Even though the work is distributed among the team , we made sure that each team member is aware of every aspect of the project.

Tweet Collection and Solr setup – Dhanashree , Uday Bhaskar, Sandeep, Prasanth.

User Interface - Dhanashree , Uday Bhaskar

Server Setup & Backend dev - Prasanth , Sandeep

Tweet Analysis - Dhanashree , Uday Bhaskar, Sandeep, Prasanth

Application hosting : Prashanth , Uday Bhaskar

Modifications & Report : Sandeep ,Dhanashree.

Video making : Dhanashree, Prasanth.

Video and App URL:

Video -- <https://youtu.be/rq32AA2D4pY>

App -- <https://irproj4-1544172225843.appspot.com/>

Conclusion:

Thus we built a multi-lingual IR system that provides insight related to social conversations on important societal issues.