



University at Buffalo
The State University of New York

**Complete Search and Analytics Solution
based on dissecting twitter data**

**Introduction to Information Retrieval
CSE 535**

**Project - 4
IR System - SHODH**

SUBMITTED BY:

ANIKET KULKARNI (50289457)

PRANAV BHAGWAT (50290966)

SOUMITRA ALATE (50289133)

VISHAL GAWADE (50290596)

OVERVIEW:

The data was crawled from Twitter for the given topic – Trump using the twitter search API. The crawled data was then processed using python script to extract the required information such as text, hashtags, user names, language etc. The geographical information such as locality, state and country was also added to the processed tweet JSON files using Google Maps Geoencoding API. The processed tweet collection is then indexed in Solr. We developed a UI which we named as SHODH such that Solr acts as the backend of it. We used spark java a web application framework which interacted with backend solr and frontend UI. We also integrated the analysis of search results to make it more interesting for the user. We also added the advanced search feature to allow the user to filter the tweet results based on the language. A query parser was also developed as part of this project to ensure that our IR system caters to the information need of the user. The homepage of our search engine is as shown below:

FEATURES:

- 1) We crawled tweets in five different languages - English, Spanish, French, Thai and Hindi.
- 2) UI displays 10 results on each page and total 500 results are fetched.
- 3) We implemented analysis of returned tweets corresponding to languages, locations, sentiment analysis, top hashtags, twitter age.
- 4) The results can be filtered based on the language according to user requirements.
- 5) User can query in a language and expect to obtain results in other languages.

Preprocessing Data:

We implemented the preprocessing of data similar to the implementation we used for project 1 by using a python script. We gathered information for the fields such as user information fields- user-name, screen-name, profile-image, location and necessary fields such as date, text, emoticons, hashtags, etc. The data is processed in order to index the data effectively corresponding to the required fields. The preprocessing of data also included the addition of location fields – locality, state, country based on the geographical coordinates information present in the raw tweet files.

Microsoft Translator API:

We used the Microsoft Azure Translator API which is a cloud based machine translation service. We used the API to detect the language of the given query and translate it to other languages such that the returned results are relevant to the user.

Spark Java

Spark is a free and open-source software web application framework and domain-specific language written in Java. It is an alternative to other Java web application frameworks such as JAX-RS, Play framework and Spring MVC.

Maven

Maven is a build automation tool used primarily for Java projects. Maven addresses two aspects of building software:

- 1) Describes how software is built

2) Describes its dependencies.

Query Parser and query expansion

We used the BM25 model along with the dismax parser. We translated the query into 5 languages and boosted the results of tweet whose results matches with the query language so that the most relevant tweets were shown to the user.

```
defType=dismax&q=(ट्रम्प+जीत+रहा+है)+OR+(ત્રમ્પ+જીત+રહા+હૈ)+OR+(trump+is+winning)+OR+(Trump+gagne)+OR+(Trump+está+ganando)&qf=text_en^3+text_hi+text_fr+text_th+text_es+hashtags&facet=true&facet.field=hashtags&rows=500&facet.limit=500
```

Faceted Search:

Faceting is the arrangement of search results into categories based on indexed terms. This technique of using faceted fields can be exploited in order to perform the analysis of the returned search results in terms of language count, location of the user. Various charts have been created for the visual representation of the analyzed data. The concept of faceted searching is also used to narrow down the search results by setting parameters such as language in order to get much more relevant results for the user. In our project we have implemented faceted search on trending hashtags. We return a list of the top ten hashtags returned for the given query term/ phrase.

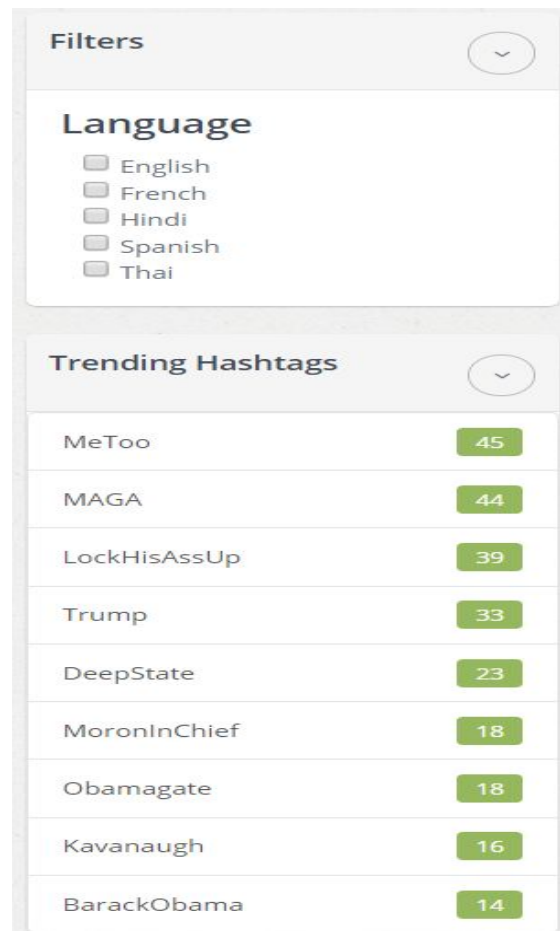


Figure 1: Language filter and top ten hashtags filter for the given query term/ phrase.

SHODH UI:

We developed a UI for our application by using HTML, CSS3 and Javascript. Bootstrap was used to enhance the user experience and make the user interface more aesthetically appealing. We used an ATLANT theme for implementing the UI. The results page displays 10 results at a time in various pages similar to the design observed in Google Search Engine. In individual tweets, various information has been displayed such as user name , screen name, number of retweets and favourite count using the data from processed tweet files. The results page also displays analysis with respect to the particular query term/ phrase. The analysis page shows analysis over the complete corpus.

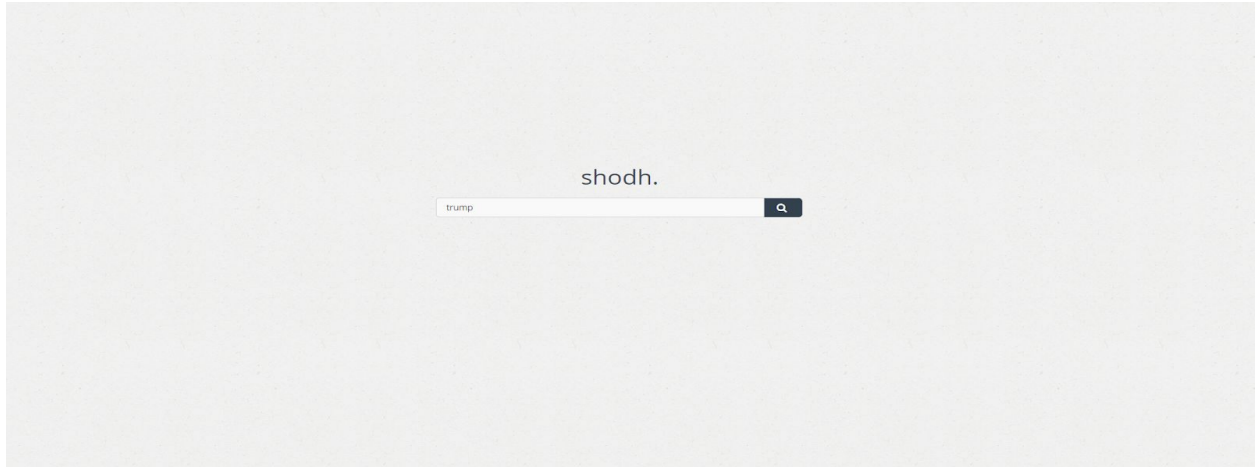


Figure 2: Shodh homepage











trump			
	prouddeplorable.mbaus @0809Mmy Trump! Trump! Trump!	0	0
	Susan Kaspar @StdPoodleMom @RealMuckmaker 1. Trump 2. Trump 3. Trump 4. Trump 5. Trump	0	0
	michele sinopoli @MicheleSinopoli Trump supporters are believed in Trump are stupidly!	0	0
	People V. Trump @Not45Th #StopKavanaugh @SenatorCollins @IIsamurkowski @SenatorTimScott @SenJoeManchin @SenCapito #WhatsAtStake TRUMP... https://t.co/Y2qhKaVetN	0	0
	Mark @markarmy2 @HillaryClinton For Trump	0	0
	Letter Lout @Letterlout @Smittyme @BishopJacob @RealSaavedra @JackPosobiec @DonaldJTrumpJr @AC360 @realDonaldTrump @FoxNews @OANN Trump is... https://t.co/Ykz4V7VaRO	0	0
	Tsk Tsk, Tut Tut, Piffle & Pshaw @Myshiloh @Bluepeople1 @Juliebythecoast @ElaineEguthrie1 @eugenegu @LindseyGrahamSC Trump.	0	0
	Big Dog @kenk22 Trump.... https://t.co/GBbsZMXief	0	0
	Jon Holden Galluccio @Gallucciojon #TFA trump	0	0
	Unamüdo @Unamudo @Tu_diabla @Alcalinus_ Trump? https://t.co/9Sa4BsRKfV	0	0
<div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>50</div> </div>			

Figure 3: Shod results page table for the given query term/ phrase

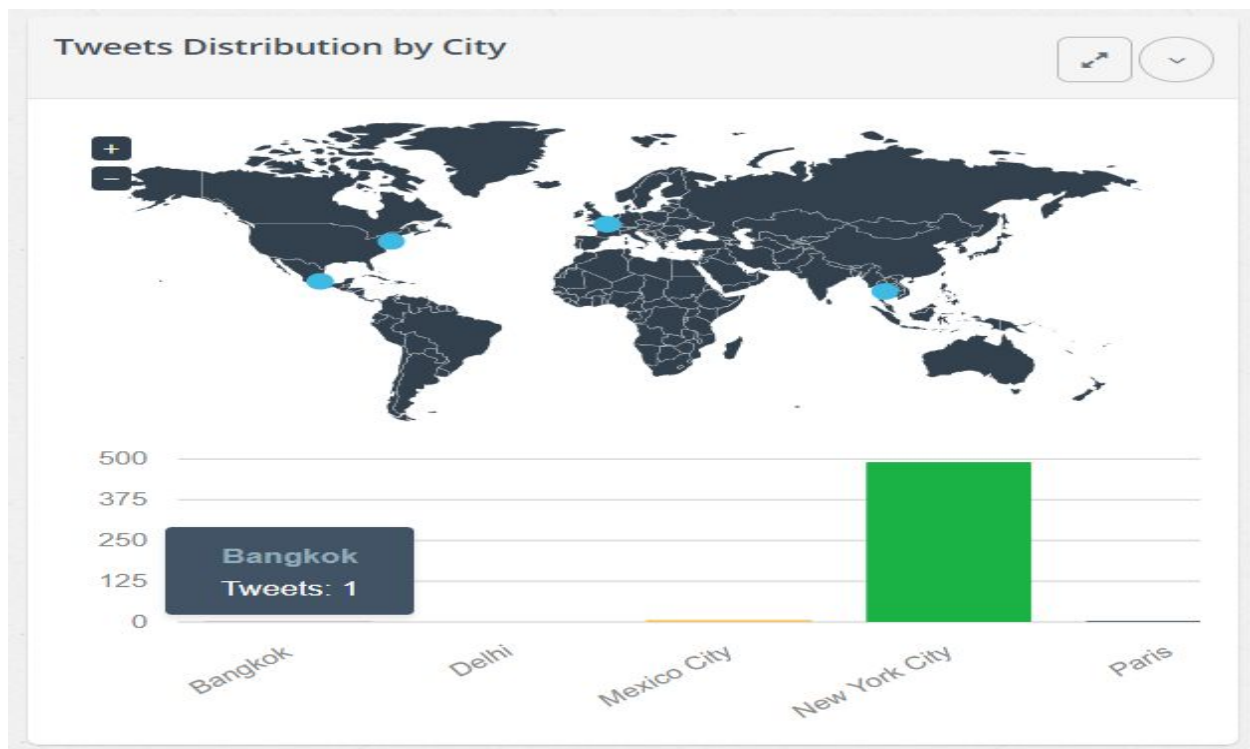


Figure 4: Tweet distribution with respect to the target cities for the given query term/ phrase

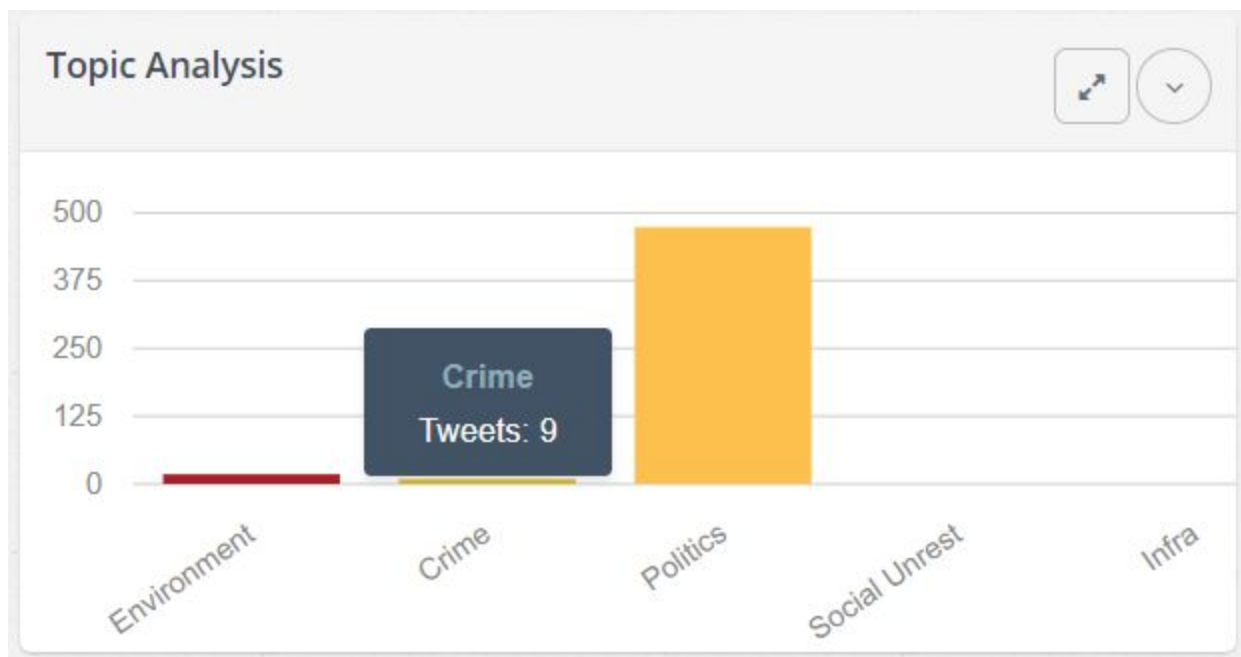


Figure 5: Topic distribution with respect to the target topics for the given query term/ phrase

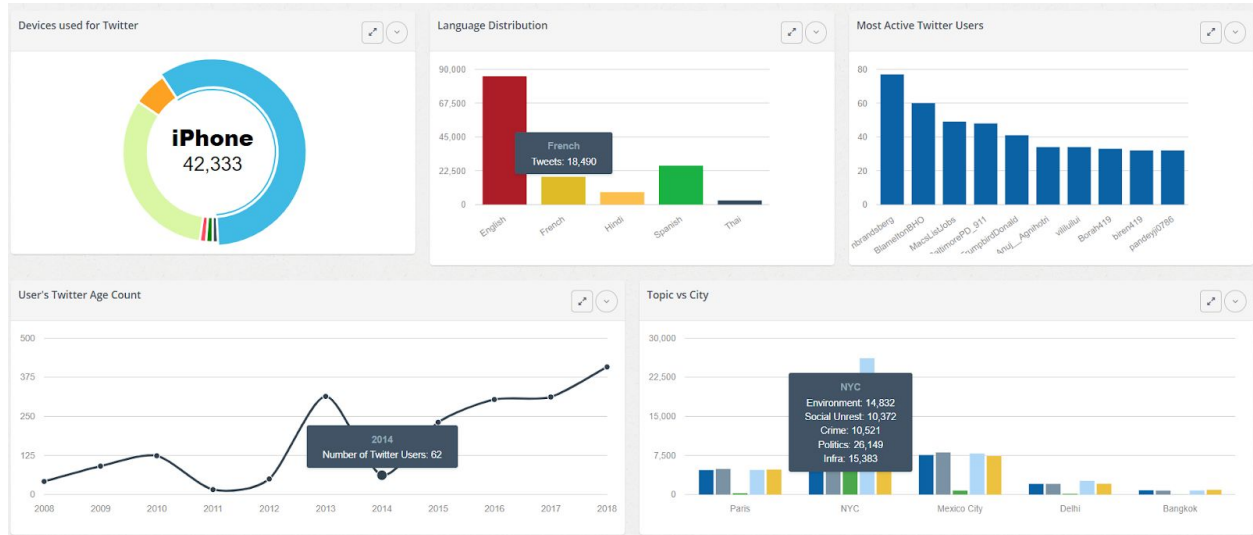


Figure 6: Analysis over the complete corpus

Sentiment Analysis:

We implemented the simplest form of Sentiment analysis on the tweet results set using a word list called AFINN-111. The file AFINN-111 is the latest version which consists of 2477 words and phrases. The approach we used compares the terms in tweet_text with the list of words in AFINN-111 file which contains pairs of word and precomputed sentiment score associated with the word (ranging from -5 to +5). The sentiment of the tweet is determined by adding up the sentiment components of the individual words in the tweet. For the results returned for the user-given query, the number of positive, neutral and negative tweets present in the search results is displayed. AFINN-111 file was available only in English language. This file was parsed and translated into the other languages in order to create sentiment based word-lists for other languages. By doing this, we can now derive the sentiment analysis for tweet results of all languages.

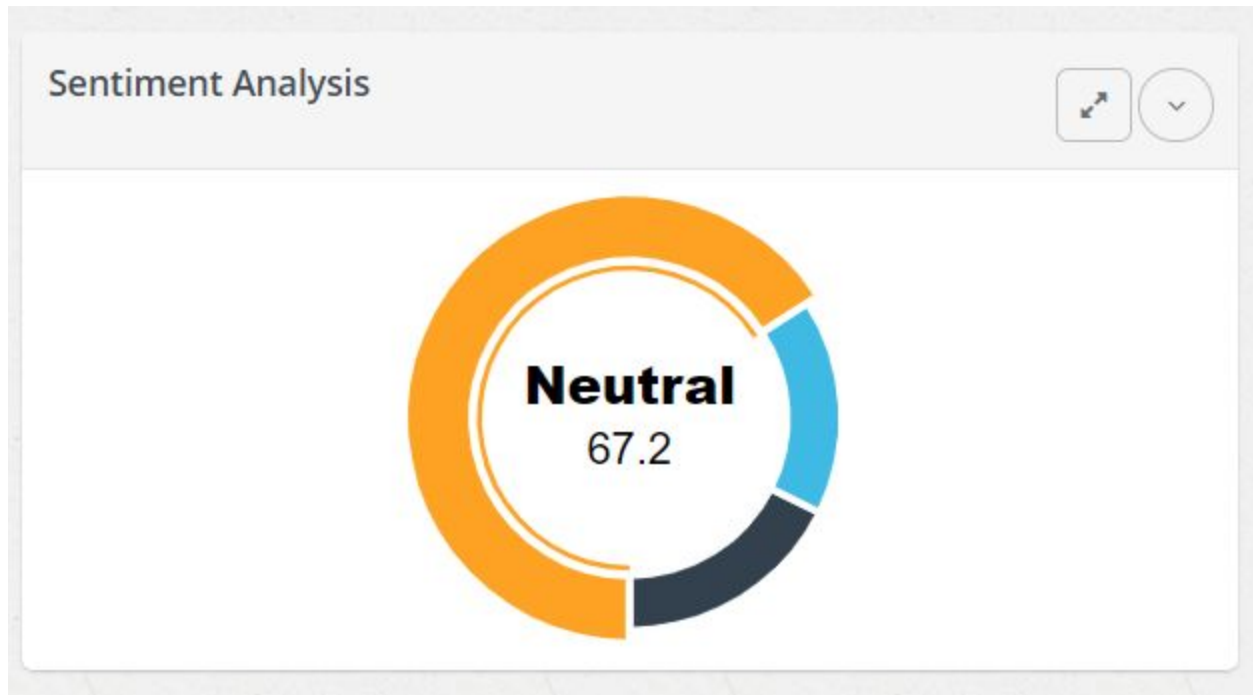


Figure 7: Sentiment analysis for the given query term/ phrase

VIDEO DEMONSTRATION:

We demonstrated the functionality of our IR system in a video:

https://drive.google.com/file/d/15PkLHSezE1tRuOIXTXPXiSACgW2GsB_h/view

In this, we queried for – Trump. It was demonstrated in the video that the language of the query was detected and relevant tweets in other languages were returned. The analysis of the results was highlighted in the demonstration.

TEAM CONTRIBUTIONS:

TEAM MEMBER	UBIT NAME	UBID	TASKS
Aniket Kulkarni	aniketvi	50289457	1.Front End development(UI) 2.UI Data Processing 3.Integrating Backend calls
Pranav Bhagwat	pbhagwat	50290966	1.Tweet Data Analysis 2.Tweet Collection
Soumitra Alate	smalate	50289133	1.Sentiment Analysis 2.Report
Vishal Gawade	vgawade	50290596	1.Query Expansion and Boosting results,Faceted Search 2.Backend(SparkJava,SolrJ) Development 3.Integration of Solr, SparkJava and UI 4.Deployment on AWS

REFERENCES:

- 1) Microsoft Translator API - <https://www.microsoft.com/en-us/translator/translatorapi.aspx> , <https://tika.apache.org/1.8/examples.html>
- 2) Faceted searching - <https://examples.javacodegeeks.com/enterprise-java/apache-solr/solr-facetedsearch-example/>
- 3) Sentiment Analysis - http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010 , <http://www.slideshare.net/faigg/tutotial-of-sentiment-analysis>
- 4) Spark Java - <http://sparkjava.com/documentation#views-and-templates>
- 5) Apache Solr- <https://www.baeldung.com/apache-solrj>
- 6) Maven - <https://maven.apache.org/guides/getting-started/maven-in-five-minutes.html>
- 7) Faceted Search - <https://lucidworks.com/2009/09/02/faceted-search-with-solr/>
- 8) SolrJ - https://lucene.apache.org/solr/6_5_1/solr-solrj/index.html?overview-summary.html