INTRODUCTION TO INFORMATION RETRIEVAL CSE 535

FALL-2019

PROJECT-4 REPORT

SUBMITTED BY:

BHARGAVA REDDY KUNCHA (50317932) (bkuncha)
VAMSHIDHAR REDDY KOMMIDI (50310287) (vamshidh)
ROHITH KUMAR GADALAY (50314899) (rohithku)

ANALYISING THE IMPACT OF POLITICAL RHETORIC IN TRADITIONAL AND SOCIAL MEDIA

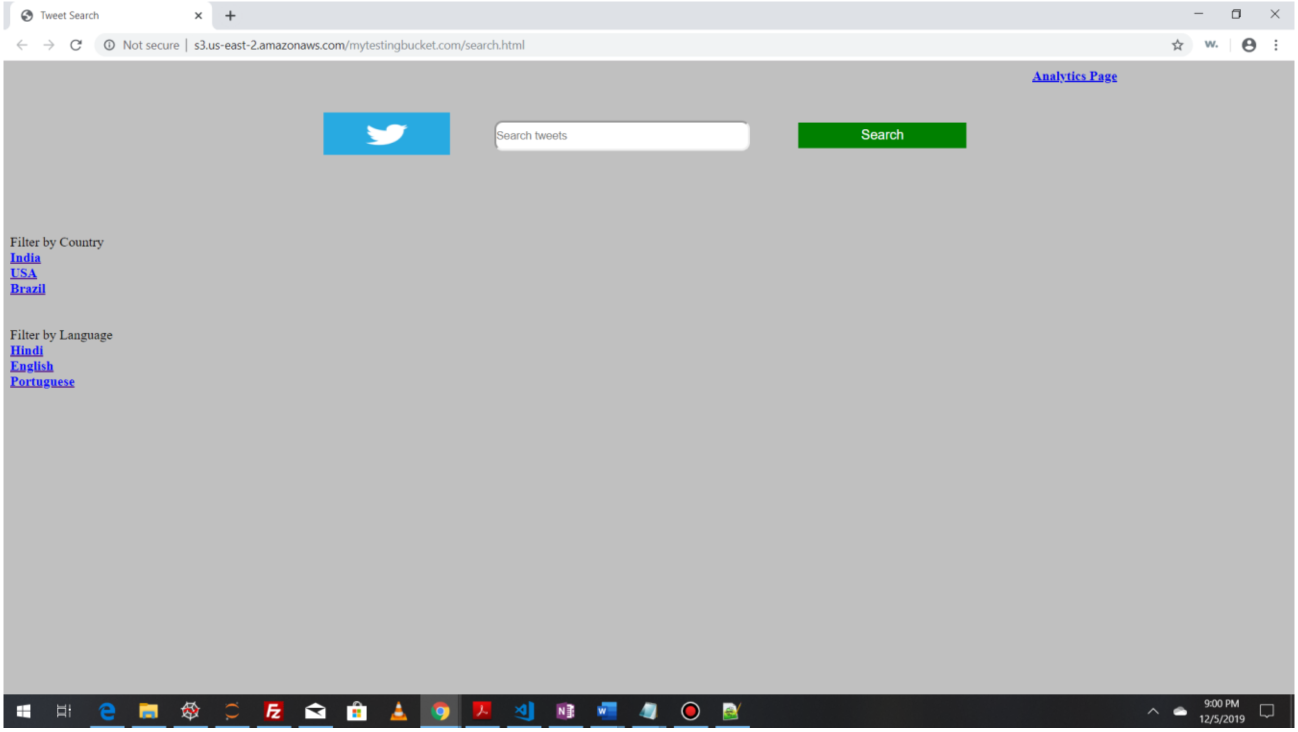
OVERVIEW:

The data was crawled from Twitter 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 processed tweet collection is then indexed in Solr. We developed a UI in which Solr acts as the backend of it and HTML and JavaScript is used for front end. We also included the filter search by language and search by country in-order to make a better interactive experience to the user.

The url of the search engine is:

http://s3.us-east-2.amazonaws.com/mytestingbucket.com/search.html

The homepage of our search engine is as shown below:



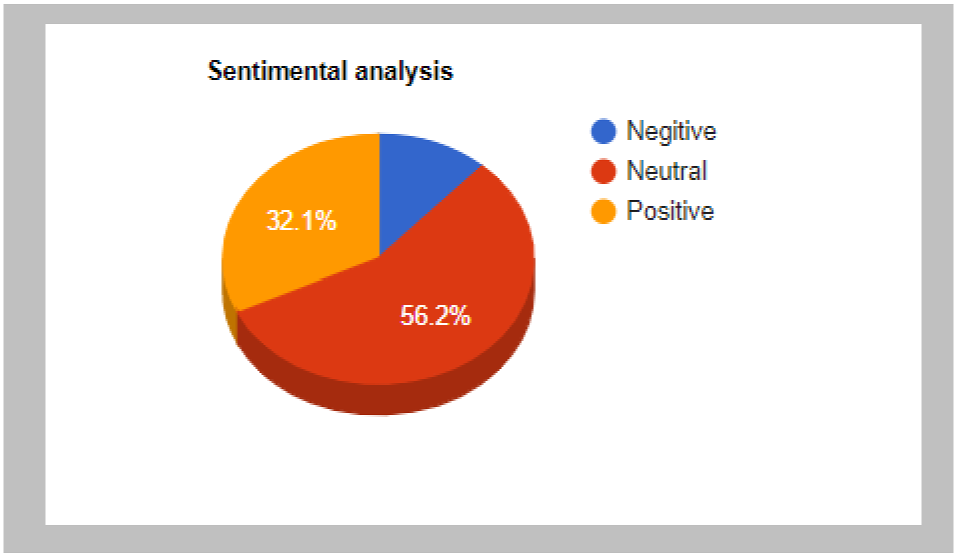
IMPLEMENTATION DETAILS:

Preprocessing Data:

We utilized the twitter data retrieved of project 1 and implemented the preprocessing of it by using a python script. 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 fields –topic obtained from topic analysis step and sentiment obtained from sentiment analysis step.

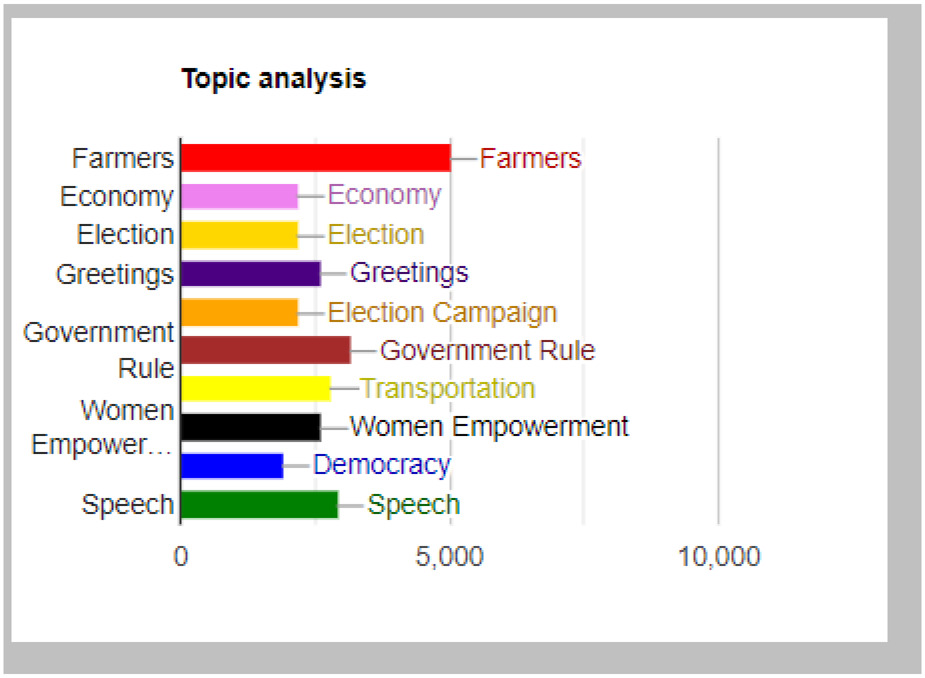
Sentiment Analysis:

We have used VADER to perform sentimental analysis on the data retrieved. We used SentimentIntensityAnalyzer package and filtered the text of the retrieved data based on the language. The polarity scores are returned based on the tweet data and the sentiment values are classified as positive, negative and neutral respectively. The pictorial representation is done by pie-chart.



Topic Analysis:

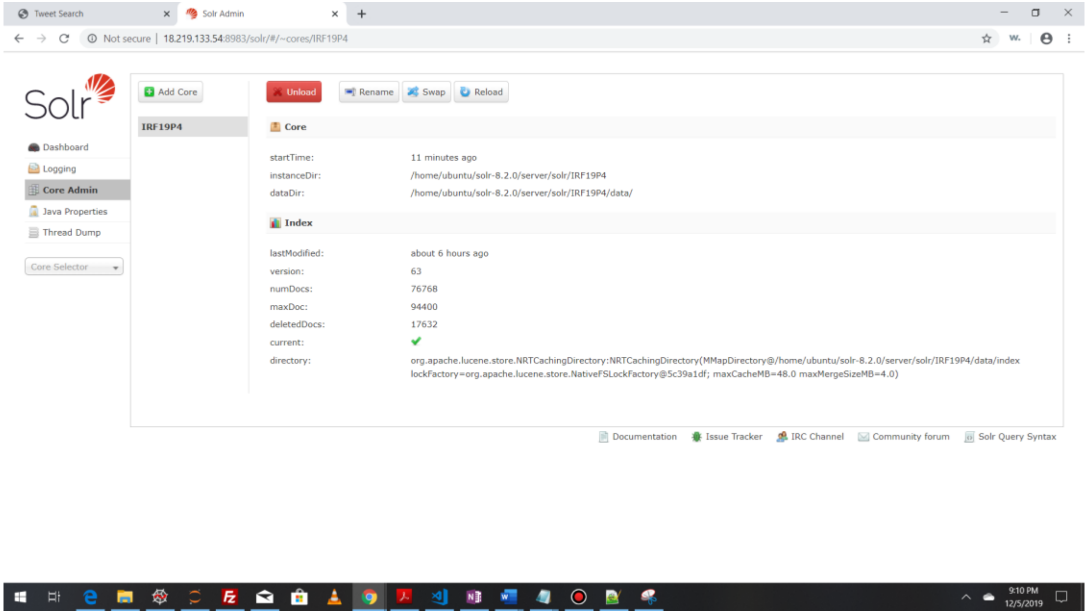
We have used unsupervised machine learning algorithm, LDA (Latent Dirichlet Allocation) to explore the topics on the retrieved data. We have used open source library spacy for text processing to load the language specific text. LDA'S approach to topic modeling is that it considers each document as a collection of topics in a certain proportion. And each topic as a collection of keywords, again in a certain proportion. Based on the scoring values of each word, we have identified the top 10 topics. We have performed topic modeling on the POI's tweet categorized by the country to show the impact of POI's tweets to show the impact on the society. Ten different topics from each country are chosen based on the POI's tweets. A sample of topic analysis for country India is represented in below figure.



Search Engine:

We developed a UI for our application by using HTML, JavaScript for front end and backend. We used Google charts tool for making the pie chart and bar graph for the visualization of analytical data displayed on the webpage. The results page displays 10 results at a time and pagination is achieved. We have included the features search by language and search by country to make the UI more interactive. We have included an analytics page, when clicked will navigate to a new page which shows the visualizations implemented. We have used dismax query parser and boosted the query results in the Solr to achieve more relevant tweets based on the query. When a keyword is queried, the results are displayed accordingly and the person who has tweeted the tweet is also displayed.

The screenshot obtained from Solr after indexing is:



Core

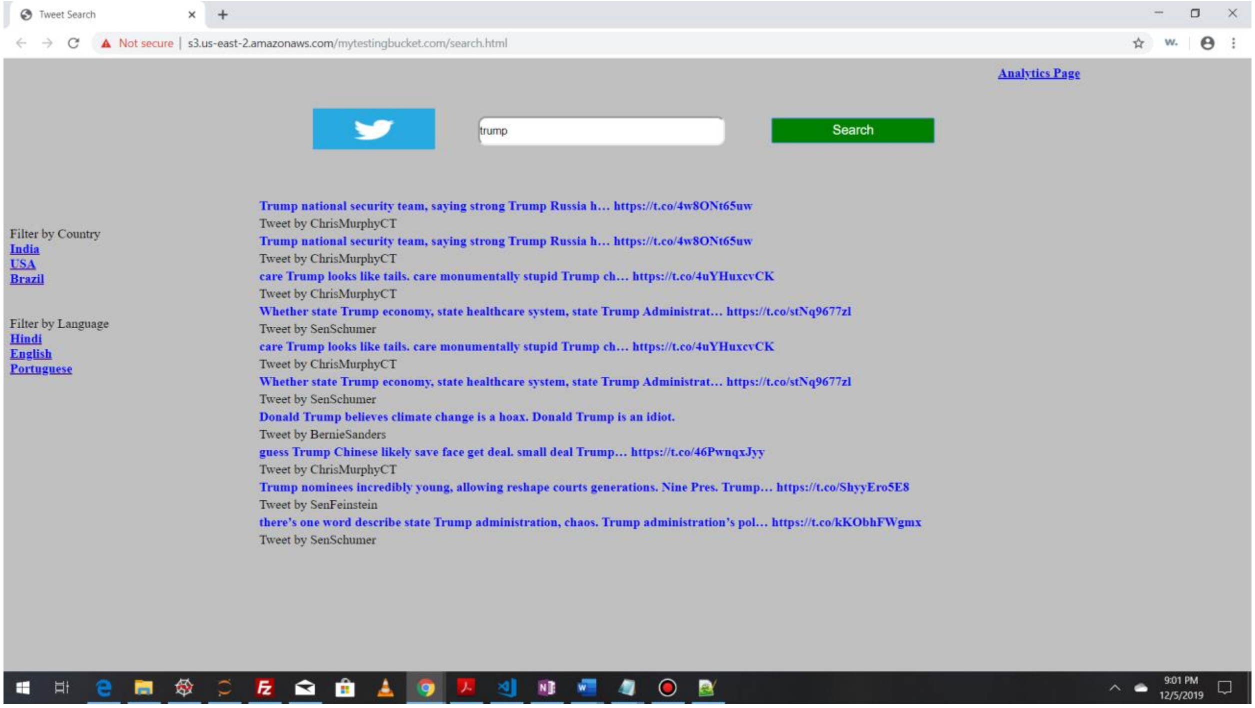
startTime: 11 minutes ago
instanceDir: /home/ubuntu/solr-8.2.0/server/solr/IRF19P4
dataDir: /home/ubuntu/solr-8.2.0/server/solr/IRF19P4/data/

Index

lastModified: about 6 hours ago
version: 63
numDocs: 76768
maxDoc: 94400
deletedDocs: 17632
current: ✓
directory: org.apache.lucene.store.NRTCachingDirectory:NRTCachingDirectory(MMapDirectory@/home/ubuntu/solr-8.2.0/server/solr/IRF19P4/data/index lockFactory=org.apache.lucene.store.NativeFSLockFactory@5c39a1df; maxCacheMB=48.0 maxMergeSizeMB=4.0)

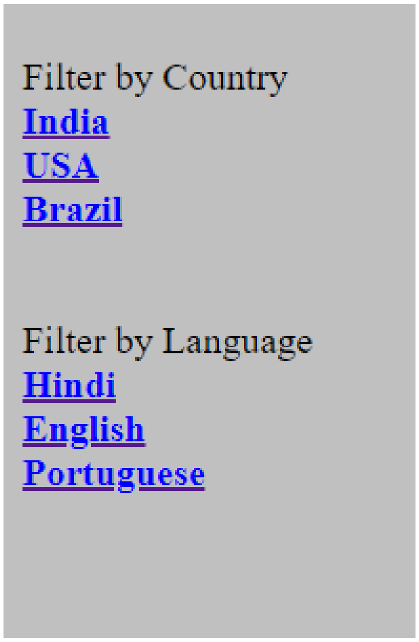
RESULTS:

When we search for trump keyword, the tweets are retrieved and the person who has done the original tweet is also represented. The resulting user interface is



Trump national security team, saying strong Trump Russia h… https://t.co/4w8ONt65uw
Tweet by ChrisMurphyCT
Trump national security team, saying strong Trump Russia h… https://t.co/4w8ONt65uw
Tweet by ChrisMurphyCT
care Trump looks like tails. care monumentally stupid Trump ch… https://t.co/4uYHuxcvCK
Tweet by ChrisMurphyCT
Whether state Trump economy, state healthcare system, state Trump Administrat… https://t.co/stNq9677zl
Tweet by SenSchumer
care Trump looks like tails. care monumentally stupid Trump ch… https://t.co/4uYHuxcvCK
Tweet by ChrisMurphyCT
Whether state Trump economy, state healthcare system, state Trump Administrat… https://t.co/stNq9677zl
Tweet by SenSchumer
Donald Trump believes climate change is a hoax. Donald Trump is an idiot.
Tweet by BernieSanders
guess Trump Chinese likely save face get deal. small deal Trump… https://t.co/46PwnqxJyy
Tweet by ChrisMurphyCT
Trump nominees incredibly young, allowing reshape courts generations. Nine Pres. Trump… https://t.co/ShyyEro5E8
Tweet by SenFeinstein
there's one word describe state Trump administration, chaos. Trump administration's pol… https://t.co/kKObhFWgmx
Tweet by SenSchumer

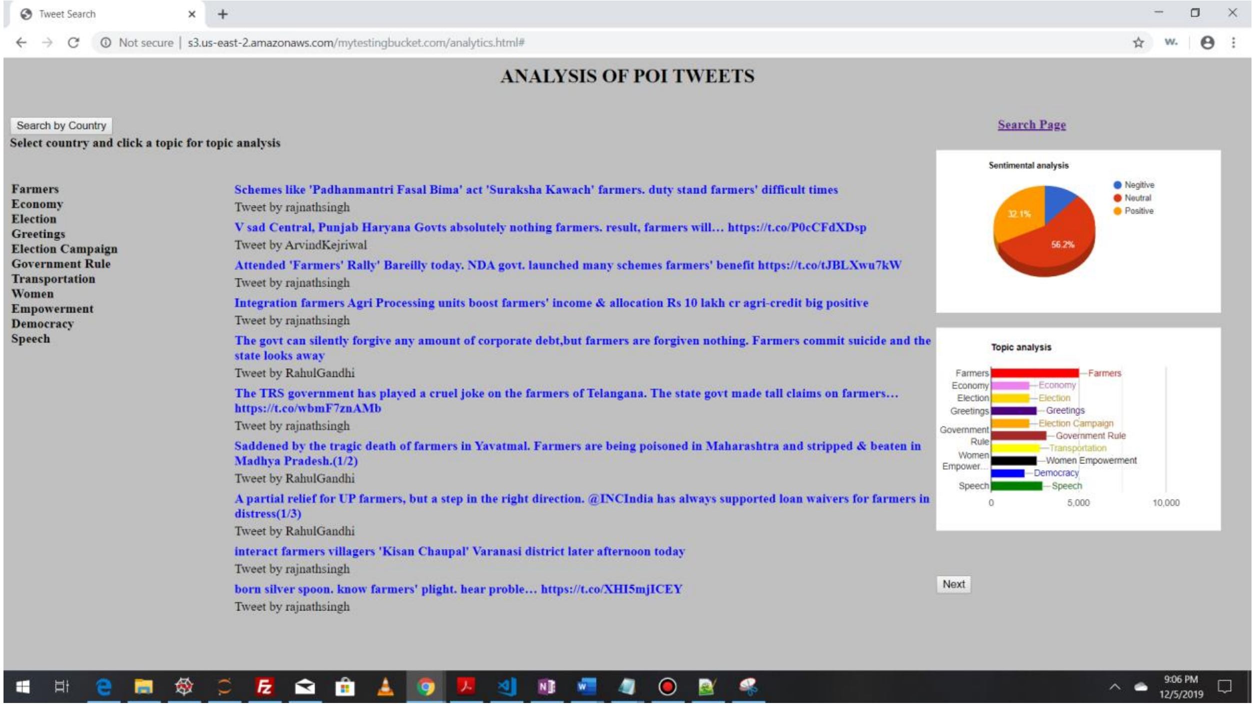
The Filter by Country and Filter by Language section is:



Filter by Country
[India]
[USA]
[Brazil]

Filter by Language
[Hindi]
[English]
[Portuguese]

When the analytics page is clicked and one of search by country is selected, the resulting user interface is:



ANALYSIS OF POI TWEETS

Select country and click a topic for topic analysis

Farmers
Economy
Election
Greetings
Election Campaign
Government Rule
Transportation
Women
Empowerment
Democracy
Speech

Schemes like 'Padhanmantri Fasal Bima' act 'Suraksha Kawach' farmers. duty stand farmers' difficult times
Tweet by rajnathsingh

V sad Central, Punjab Haryana Govts absolutely nothing farmers. result, farmers will… https://t.co/P0cCFdXDsp
Tweet by ArvindKejriwal

Attended 'Farmers' Rally' Bareilly today. NDA govt. launched many schemes farmers' benefit https://t.co/tJBLXwu7kW
Tweet by rajnathsingh

Integration farmers Agri Processing units boost farmers' income & allocation Rs 10 lakh cr agri-credit big positive
Tweet by rajnathsingh

The govt can silently forgive any amount of corporate debt,but farmers are forgiven nothing. Farmers commit suicide and the state looks away
Tweet by RahulGandhi

The TRS government has played a cruel joke on the farmers of Telangana. The state govt made tall claims on farmers… https://t.co/wbmF7znAMb
Tweet by rajnathsingh

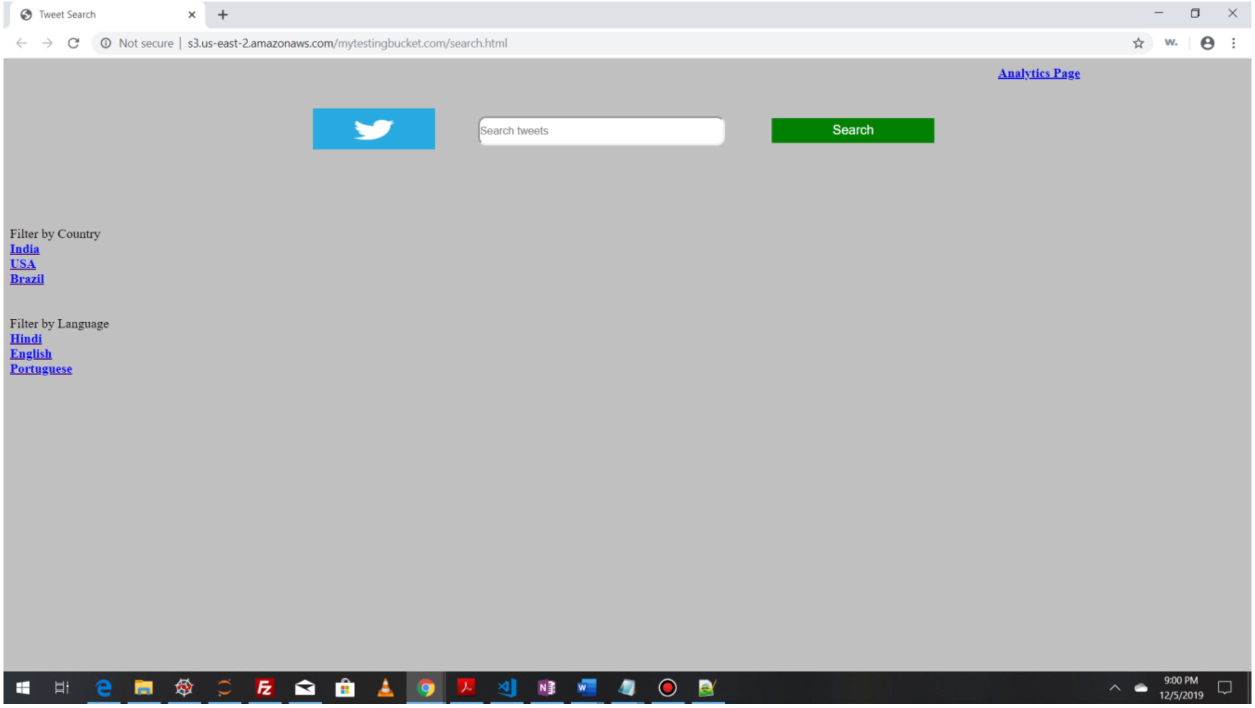
Saddened by the tragic death of farmers in Yavatmal. Farmers are being poisoned in Maharashtra and stripped & beaten in Madhya Pradesh.(1/2)
Tweet by RahulGandhi

A partial relief for UP farmers, but a step in the right direction. @INCIndia has always supported loan waivers for farmers in distress(1/3)
Tweet by RahulGandhi

interact farmers villagers 'Kisan Chaupal' Varanasi district later afternoon today
Tweet by rajnathsingh

born silver spoon. know farmers' plight. hear proble… https://t.co/XHI5mjICEY
Tweet by rajnathsingh

When the search page link in above screenshot is clicked, it navigates to the original page/home page



Tweet Search
Not secure | s3.us-east-2.amazonaws.com/mytestingbucket.com/search.html
Analytics Page
Search tweets
Search

Filter by Country
India
USA
Brazil

Filter by Language
Hindi
English
Portuguese

VIDEO DEMONSTRATION:

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

https://www.youtube.com/watch?v=A4oOCzHSeus&feature=youtu.be

In this video we queried for trump and retrieved the appropriate results from the tweets data.

TEAM CONTRIBUTIONS:

TEAM MEMBER	UBIT NAME	UBID	TASKS
Bhargava Reddy Kuncha	bkuncha	50317932	UI Design, hosting and preprocessing tweets
Vamshidhar Reddy Kommidi	vamshidh	50310287	Sentimental Analysis and Topic Analysis
Rohith Kumar Gadalay	rohithku	50314899	Solr Indexing and writing report

REFERENCES:

1) Google Translator API – https://cloud.google.com/translate/docs/
2) Topic Analysis- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/
3) Faceted searching – https://examples.javacodegeeks.com/enterprise-java/apache-solr/solr-faceted-search-example/
4) Sentiment Analysis - https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/
5)Google Chart tools - https://developers.google.com/chart/interactive/docs/