**CSE4/535 - INFORMATION RETRIEVAL FALL 2021 Project REPORT**

**DISSECTING TWITTER DATA TO ANALYZE GOVERNMENT AND PUBLIC ATTITUDE TOWARDS COVID AND VACCINES**

**SUBMITTED BY:**

**Name and UB number**

**INSTRUCTOR:**

ROHINI SRIHARI

**INTRODUCTION:**

Search engines essentially act as filters for the wealth of information available on the internet. They allow users to quickly and easily find information that is of genuine interest or value, without the need to wade through numerous irrelevant web pages.Search engines provide users with search results that lead to relevant information on high-quality websites. The operative word here is "relevant". To attain and retain market share in online searches, search engines need to make sure they deliver results that are relevant to their users’ searches.

**IMPLEMENTATION:**

As a part of this project we have implemented a Search Engine and performed content analysis.

Here are some of the tasks we have performed

• Created a webpage to perform search operations on our indexed data related to covid and vaccines.

• There is search page on our home similar to google search where users can search for any term related to covid.

• Once the user clicks enter, he will be able to see different information on both the sides:

Left side of our web page renders faceted search functionality which displays the tweets and and the top replies. We have also displayed the overall public response for that particular tweet along with some top replies .

Right side of our web page displays the graphs obtained from the results of sentiment analysis of the tweets as well as their replies.

**Technologies used:**

The Web stack that we have used to implement this search engine is Python with Django Framework. Solr is used for indexing whereas Bootstrap, CSS is used for styling and several other libraries such as matplotlib, plotly etc to aid the development.

**How we collected the data:**

During Project 1, we have collected several tweets from multiple POIs from three different countries - India, USA and Mexico. We have also collected replies for these tweets. We have used the tweets collected during Phase 1 to power the data for this project.

This Data is present in the data folder of the project.

**Indexing:**

We have indexed the data from the data folder into the solr. ( The code for this is present in indexer.py / indexer-aws.py which are used for indexing data in local / remote respectively) We are using the BM25 to rank the documents for a given search term.

**Code:**

The main API that powers the backend is the search API. This API also takes care of the filtering. Whenever the user enters a search term and clicks on enter the API will be hit and the following process takes place:

-> Perform Query cleaning by removing special characters, whitespace.

-> Performs query optimization to improve the search results. ( Using the techniques learned in Phase 3 while we evaluated several IR models and tried to improve the search results)

-> Prepares a solr url by using the search term and the filters provided by the user.

-> Hits the Solr with the URL and get the top results for the search term.

-> For each tweet, collects replies, if any, and modify the result object to incorporate the replies.

-> Performs sentiment analysis on the replies to understand the public response to a particular tweet. Each tweet will be assigned one value (positive, negative, neutral) which is calculated by considering all the replies.

-> Modify the result and send the result object which contains tweets and related information.

-> Result Object will contain a list of tweet objects.A sample tweet object:

    {

        "verified": **true**,

        "country": "USA",

        "id": "1437614790789287941",

        "replied\_to\_user\_id": 0,

        "tweet\_text": "cameras carry covid. https://t.co/ozmxfsXijK",

        "tweet\_lang": "en",

        "tweet\_date": "2021-09-14T03:00:00Z",

        "poi\_name": "tedcruz",

        "poi\_id": 23022687,

        "mentions": [

            "NaN"

        ],

        "tweet\_urls": [

            "https://t.co/ozmxfsXijK"

        ],

        "text\_en": "cameras carry covid. https://t.co/ozmxfsXijK",

        "hashtags": [

            "NaN"

        ],

        "tweet\_emoticons": [

            "NaN"

        ],

        "text\_es": "NaN",

        "\_version\_": 1717727613409034244,

        "sentiment": "neutral",

        "replies": [

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438501629024604164",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz @ChadConnelly Ban cameras! ;)",

                "tweet\_lang": "es",

                "tweet\_date": "2021-09-16T14:00:00Z",

                "reply\_text": "@tedcruz @ChadConnelly Ban cameras! ;)",

                "mentions": [

                    "tedcruz",

                    "ChadConnelly"

                ],

                "text\_en": "NaN",

                "tweet\_emoticons": [

                    ";)"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "@tedcruz @ChadConnelly Ban cameras! ;)",

                "\_version\_": 1718185150590746625

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438334099610116100",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz Can we fire all them along with you?",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-16T03:00:00Z",

                "reply\_text": "@tedcruz Can we fire all them along with you?",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz Can we fire all them along with you?",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185150827724804

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438311361835061256",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz Independent bank in Texas is garnishing the employees $100.00 a month if they do not provide proof of vaccination.",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-16T01:00:00Z",

                "reply\_text": "@tedcruz Independent bank in Texas is garnishing the employees $100.00 a month if they do not provide proof of vaccination.",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz Independent bank in Texas is garnishing the employees $100.00 a month if they do not provide proof of vaccination.",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185150848696323

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438248992115085313",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz You’re one of “these people”.",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T21:00:00Z",

                "reply\_text": "@tedcruz You’re one of “these people”.",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz You’re one of “these people”.",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185150919999489

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438184818374848513",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz #cruz…..will never get it.  Hardened heart,

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T17:00:00Z",

                "reply\_text": "@tedcruz #cruz…..will never get it.  Hardened heart,

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz #cruz…..will never get it.  Hardened heart,

                "tweet\_emoticons": [

                    ",

                    ",

                    "❣️"

                ],

                "hashtags": [

                    "cruz"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151063654401

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438177881637621765",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz Wow",

                "tweet\_lang": "und",

                "tweet\_date": "2021-09-15T16:00:00Z",

                "reply\_text": "@tedcruz Wow",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "NaN",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151084625923

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438170252995616771",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz I might be more sympethetic to a message if they lived it....  Just like AOC MET Gala dress...start with her @100%",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T16:00:00Z",

                "reply\_text": "@tedcruz I might be more sympethetic to a message if they lived it....  Just like AOC MET Gala dress...start with her @100%",

                "mentions": [

                    "tedcruz",

                    "100"

                ],

                "text\_en": "@tedcruz I might be more sympethetic to a message if they lived it....  Just like AOC MET Gala dress...start with her @100%",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151121326081

            },

            {

                "verified": **true**,

                "country": "USA",

                "id": "1438164672537141253",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz They're outside, moron.",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T16:00:00Z",

                "reply\_text": "@tedcruz They're outside, moron.",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz They're outside, moron.",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151142297602

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438163348957241346",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz @mschlapp You should post some pictures or videos of maskless Larry Elder posing for photos - you know, for contrast.",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T16:00:00Z",

                "reply\_text": "@tedcruz @mschlapp You should post some pictures or videos of maskless Larry Elder posing for photos - you know, for contrast.",

                "mentions": [

                    "tedcruz",

                    "mschlapp"

                ],

                "text\_en": "@tedcruz @mschlapp You should post some pictures or videos of maskless Larry Elder posing for photos - you know, for contrast.",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "NaN"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151143346179

            },

            {

                "verified": **false**,

                "country": "USA",

                "id": "1438152577237995522",

                "replied\_to\_tweet\_id": 1437614790789287941,

                "replied\_to\_user\_id": 23022687,

                "tweet\_text": "@tedcruz Hey Ralphy, correction, no one trusts you.\n#weaselTed\n#CancunCruz",

                "tweet\_lang": "en",

                "tweet\_date": "2021-09-15T15:00:00Z",

                "reply\_text": "@tedcruz Hey Ralphy, correction, no one trusts you.\n#weaselTed\n#CancunCruz",

                "mentions": [

                    "tedcruz"

                ],

                "text\_en": "@tedcruz Hey Ralphy, correction, no one trusts you.\n#weaselTed\n#CancunCruz",

                "tweet\_emoticons": [

                    "NaN"

                ],

                "hashtags": [

                    "weaselTed",

                    "CancunCruz"

                ],

                "tweet\_urls": [

                    "NaN"

                ],

                "text\_es": "NaN",

                "\_version\_": 1718185151150686209

            }

        ],

        "positive\_replies": 2,

        "negative\_replies": 1,

        "neutral\_replies": 7

    }

**METHODOLOGY:**

Evaluation Methodology:

We are using the BM25 to rank the documents for a given search term. BM25 (BM is an abbreviation of best matching) is a [ranking function](https://en.wikipedia.org/wiki/Ranking_function" \o "Ranking function) used by [search engines](https://en.wikipedia.org/wiki/Search_engine" \o "Search engine) to estimate the [relevance](https://en.wikipedia.org/wiki/Relevance_(information_retrieval)" \o "Relevance (information retrieval)) of documents to a given search query. It is based on the [probabilistic retrieval framework](https://en.wikipedia.org/wiki/Probabilistic_relevance_model" \o "Probabilistic relevance model).

**SAMPLE SCREENSHOTS:**

**WORK BREAKDOWN BY TEAMMATES:**

**Sri Sai Omkar Akilla**:

Performed Data Collection and indexing

Backend:

Developed Search API (returning tweets along with the replies for a search term).

Calculated Sentiment analysis of each tweet.

Found number of COVID and non- COVID tweets by POIs.

**CONCLUSION:**