Github URL: https://github.com/enekotrevi/IRWA-2024.git **TAG:**IRWA-2024-part-4

Part 4: User Interface and Web Analytics

1. User Interface.

In the development of our search engine's user interface, we undertook several crucial steps to ensure a seamless and intuitive user experience

1.1. Data Preparation.

Meticulously read and adapted the load corpus file to read the file about the farmers protest tweets, ensuring a seamless integration into our code. To do this, we read and made sure we understood the code of the skeleton and then merged it with the knowledge we gained from the first lab, where we also had to read this file with a required format. Here, we save each tweet as a document object to be used on the search engine to generate results once a query is given.

1.2. Algorithm Integration.

The previously developed algorithms, those used in Part 2, have been harmoniously integrated and carefully tuned for compatibility

- The TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is a numerical representation technique in natural language processing that we use. It measures the importance of a word in a tweet relative to the collection of tweets. We calculated a weight for each word based on its frequency in the tweet (TF) and its rarity across all tweets (IDF). High TF-IDF scores indicate words that are specific to a document, which helps us in information retrieval and text analysis by highlighting key terms while reducing common words. Following the previous score, we applied cosine similarity to generate the most similar results to the user's query.

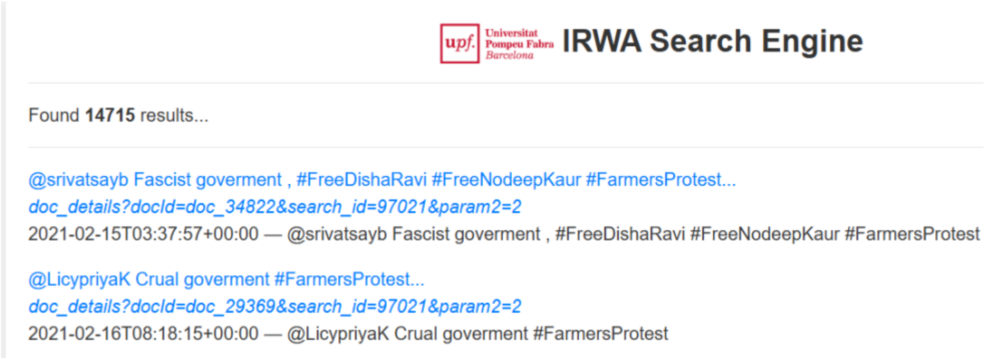
To do this, we took the algorithm we implemented in previous parts of the project and we have adapted it to the structure of the skeleton. As we did in parts 2 and 3, to avoid creating the index at each execution—since it is time-consuming—we use a function to import the serialized index, which is located in the *data* folder.

1.3. Results Presentation.

- Engineered a visually refined presentation of search results. Drawing inspiration from the streamlined design principles observed in the results page of search engines like Google.

- Each result is thoughtfully composed, featuring usernames, a clickable title leading to the complete tweet and the publish date of the tweet.

In the following image you can see both elements.



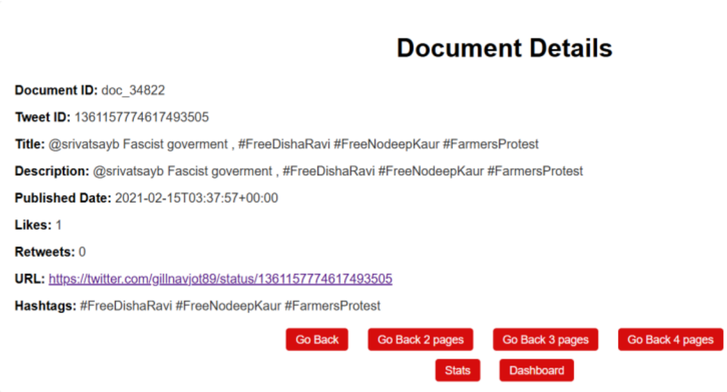
Found **14715** results...

@srivatsayb Fascist goverment , #FreeDishaRavi #FreeNodeepKaur #FarmersProtest...
doc_details?docId=doc_34822&search_id=97021&param2=2
2021-02-15T03:37:57+00:00 — @srivatsayb Fascist goverment , #FreeDishaRavi #FreeNodeepKaur #FarmersProtest

@LicypriyaK Crual goverment #FarmersProtest...
doc_details?docId=doc_29369&search_id=97021&param2=2
2021-02-16T08:18:15+00:00 — @LicypriyaK Crual goverment #FarmersProtest

1.4. Tweet display.

Increased user interaction by implementing a feature that allows users to view the full tweet by simply clicking on its title. We included all the information we considered relevant to the tweet, such as the number of likes, retweets or hashtags, as well as the URL and the date the tweet was published.



Document Details

Document ID: doc_34822

Tweet ID: 1361157774617493505

Title: @srivatsayb Fascist goverment , #FreeDishaRavi #FreeNodeepKaur #FarmersProtest

Description: @srivatsayb Fascist goverment , #FreeDishaRavi #FreeNodeepKaur #FarmersProtest

Published Date: 2021-02-15T03:37:57+00:00

Likes: 1

Retweets: 0

URL: https://twitter.com/gillnavjot89/status/1361157774617493505

Hashtags: #FreeDishaRavi #FreeNodeepKaur #FarmersProtest

2. Web Analytics

Our web analytics implementation encompasses a diverse set of features and insightful visualizations to unravel the intricacies of user behavior.

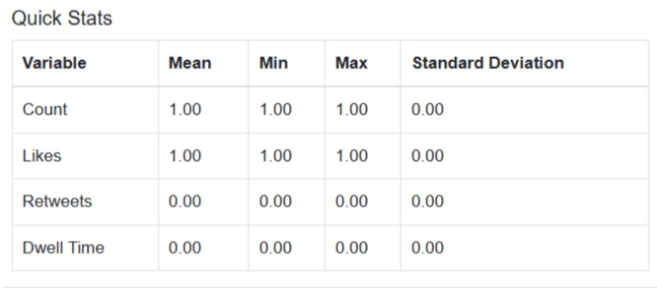
2.1. Stats.

The stats are divided into four separate tabs and each tab is composed of different metrics.

2.1.1. Visited Documents.

2.1.1.1. Quick Stats.

These are statistics for Count, Likes, Retweet and Dwell time that calculate the mean, min, max and standard deviation of each variable mentioned above.



Quick Stats

Variable	Mean	Min	Max	Standard Deviation
Count	1.00	1.00	1.00	0.00
Likes	1.00	1.00	1.00	0.00
Retweets	0.00	0.00	0.00	0.00
Dwell Time	0.00	0.00	0.00	0.00

2.1.1.2. Engagement Metrics.

We build compute this measures in order to see how much users interact with documents.

- **Engagement Score.**

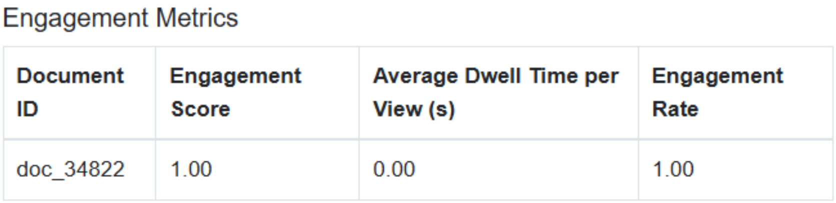
It is a weighted combination of likes, retweets and dwell time, we compute it using the following formula: $Engagement\ = likes\ +\ (retweets\ 2)\ +\ (dwell\ time\ 10)$.

- **Average Dwell Time per View (s).**

Average time users spend in the document, we used the following formula: $Avg. Dwell\ Time\ = \frac{dwell\ time}{count}$.

- **Engagement Rate.**

Measures the percentage of users interacting with the document. The formula applied in this case is: $Engagement\ Rate\ = \frac{likes + retweets}{count}$.



Engagement Metrics

Document ID	Engagement Score	Average Dwell Time per View (s)	Engagement Rate
doc_34822	1.00	0.00	1.00

2.1.1.3. Popularity Metrics.

Metricces to determine the popularity of a document.

- **View Share.**

Percentage of visits of the document in relation to the total number of documents. The formula used is: $View\ Share\ = \frac{count}{Total\ visits} * 100$.

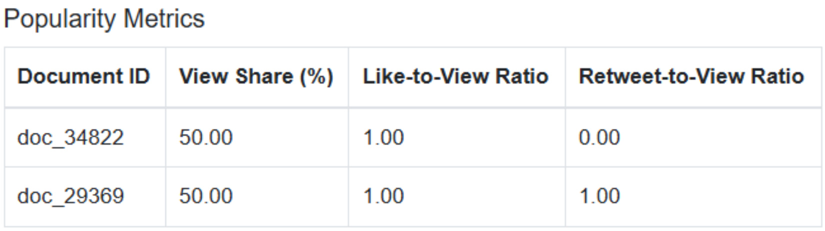
- **Like-to-View Ratio.**

Relationship between Likes and Views. We computed this using:

$$Like\ to\ View\ Ratio\ = \frac{likes}{count}$$

- **Retweet-to-view Ratio.**

Relationship between Retweets and Views, it was computed using: $Retweet\ \ to\, View\ Ratio\ = \frac{retweets}{count}$.

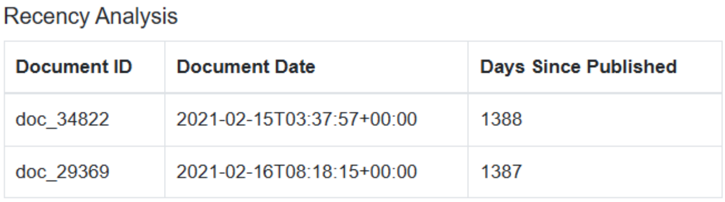


Popularity Metrics

Document ID	View Share (%)	Like-to-View Ratio	Retweet-to-View Ratio
doc_34822	50.00	1.00	0.00
doc_29369	50.00	1.00	1.00

2.1.1.4. Recency Analysis.

It is the history of the documents that have been accessed.



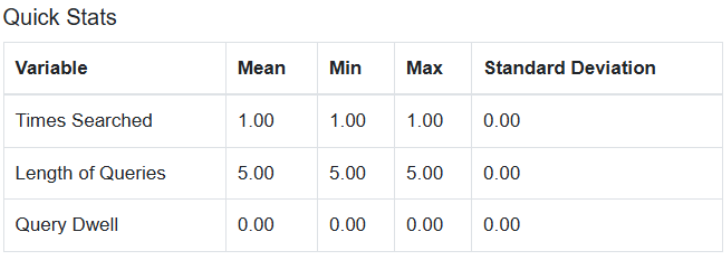
Recency Analysis

Document ID	Document Date	Days Since Published
doc_34822	2021-02-15T03:37:57+00:00	1388
doc_29369	2021-02-16T08:18:15+00:00	1387

2.1.2. Searched Queries.

2.1.2.1. Quick Stats.

These are statistics for Times Searched, Length of Queries and Query Dweel that calculate the mean, min, max and standard deviation of each variable mentioned above.



Quick Stats

Variable	Mean	Min	Max	Standard Deviation
Times Searched	1.00	1.00	1.00	0.00
Length of Queries	5.00	5.00	5.00	0.00
Query Dwell	0.00	0.00	0.00	0.00

2.1.2.2. Popularity Metrics.

These metrics assess how common or popular the queries are among users.

- **Search Share.**

Ratio of searches for a query to total number of searches performed. The formula applied in this case is: $Search\ Share\ = \frac{times\ searched(query)}{\text{total number of searches performed}}$.

- **Time Rate.**

Time spent on each query.

- **Success Rate.**

To compute the percentage of satisfactory queries, we used two different metrics: one based on terms and the other on time.

For computing the number of queries that are satisfactory, based on terms, we use the following procedure: for each query, check if the previous query shares 70% of its terms (threshold) with the current query. If so, mark the previous query as unsatisfactory, as it implies you are searching for the same thing.

For the time-based metric, we check if the dwell time for a query is less than 5 seconds. If so, the query is considered unsatisfactory.

Finally, we computed the rate as following:

$$Success\ Rate\ = \frac{Queries\ that\ produced\ satisfactory\ results}{Number\ of\ queries} * 100$$

2.1.2.3. Difficulty Metrics.

In this section, we evaluate the difficulty of the query according to the number of terms it contains.

- **Percentage of Short Queries.**

$$Percentage\ of\ Short\ Queries\ = \frac{query\ with\ length\ \leq 2}{Total\ queries} * 100$$

- **Percentage of Long Queries.**

$$Percentage\ of\ Long\ Queries\ = \frac{query\ with\ length\ \geq 5}{Total\ queries} * 100$$

2.1.3. HTTP.

2.1.3.1. Request Metrics.

Count the number of the differents requests, in order to know if it is a GET or POST request.

2.1.3.2. Click Metrics.

In this section we show the number of clicks in the differents elements, so if the click is to the index page, the results, dashboard, etc.

2.1.3.3. Session Analysis.

For the Session Analysis we want to print the number of different sessions per each day.

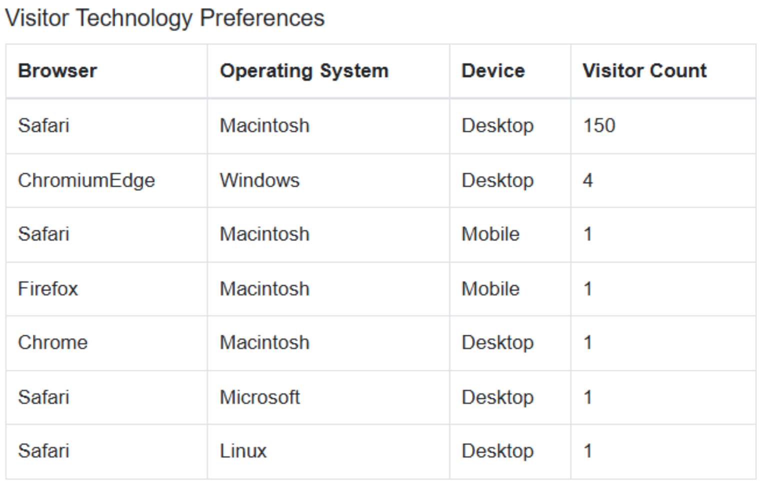
2.1.4. Visitors' History.

2.1.4.1. Traffic Timing Metrics.

These metrics help you understand when visitors are most active. It is granulated by date, week and month.

2.1.4.2. Visitor Technology Preferences.

T hese metrics indicate the browsers visitors use to access the website. The table presents all possible combinations of visitor preferences and the corresponding number of visitors for each combination.



Visitor Technology Preferences

Browser	Operating System	Device	Visitor Count
Safari	Macintosh	Desktop	150
ChromiumEdge	Windows	Desktop	4
Safari	Macintosh	Mobile	1
Firefox	Macintosh	Mobile	1
Chrome	Macintosh	Desktop	1
Safari	Microsoft	Desktop	1
Safari	Linux	Desktop	1

2.2. Dashboard.

Crafted an engaging graphical representations to illustrate different analysis, providing a nuanced perspective on user preferences.

2.2.1. Visited Documents.

In this section, we introduce two different plots. The first analyzes the visits, number of likes, retweets, and dwell time spent on each document. The second shows the publication date of the visited tweets, with the dates sorted.

2.2.2. Searched Queries.

In a new section, we included two dasboards:

- Quick Stats: This graph displays key numercial variables of queries, such as length, dwell time or number of time that the query is searched,offering a quick overview of the data.
- Terms Searched Frequency: This plot shows the frequency of words appearing in the tweets, highlighting the most commonly searched terms.

2.2.3. HTTP.

For the http we built the following dashboards:

- Method Distribution: This plot shows the number of instances for each method used, giving an overview of the distribution of different methods in the session.
- IP Address Distribution: This plot displays the distribution of different IP addresses, showing how the data is spread across various sources based on IPs.
- Clicks: This plot shows the number of clicks per element, helping to identify which elements were clicked the most by users.
- Time Distribution (Per Minute): This plot displays the number of clicks per minute, providing insight into how click activity is distributed over time.

2.3.4. Visitors’ History

In this section, we present several plots that provide insights into various aspects of visitor activity:

- Browsers Analysis: This plot shows the number of visitors using different browsers, helping to understand the distribution of browser usage.
- OS Analysis: This plot displays the number of visitors using different operating systems, providing insights into the OS preferences of visitors.

- Device Pie Chart: This pie chart identifies the proportion of access from different devices, such as mobile, tablet, or desktop.
- IP Address Pie Chart: This pie chart identifies the distribution of different IP addresses, showing how visits are distributed across various sources.
- Visitors Over Time: This plot shows the number of visitors over time, illustrating trends in traffic and identifying peak visitation periods.

3. Data Storage and management.

Documents are saved in the corpus using the defined Document class. We use the StatsDocument and ResultItem classes to temporarily store the retrieved documents for stats.html and doc_details.html.

For sessions and visitors, we created .txt files and the functions load_sessions_from_file() and read_visitors_file(), respectively, to store the data outside the search engine and generate the stats and dashboards. Both have their respective classes to implement the code more easily and efficiently.

We create an instance of the HTTPAnalytics class, in which we store the queries in a dictionary of the Query class, along with the requests and clicks. To track each request, we create the track_request() method. The HTTPAnalytics class also stores the current visitor.

4. Future Works

The following outlines potential avenues for future development, promising to elevate both the user interface and web analytics capabilities of our search engine.

- Enhanced User Interface: Implement user feedback mechanisms to continuously refine and improve the user interface based on user interactions and preferences.
- Real-Time Updates: Explore the feasibility of implementing real-time updates to keep users informed of the latest relevant information.
- Machine Learning Integration: Investigate the integration of machine learning models to enhance search result relevance and improve the overall user experience.
- Extended Analytics: Expand the scope of web analytics by incorporating additional metrics, such as click-through rates, popular search queries, and user demographics.

- Enhanced Sentiment Analysis: Enhance the sentiment analysis feature by incorporating advanced natural language processing techniques for more accurate emotion prediction.