TWITTER SEARCH APP

BY:

ADVAITH RAO

AYUSH OTURKAR

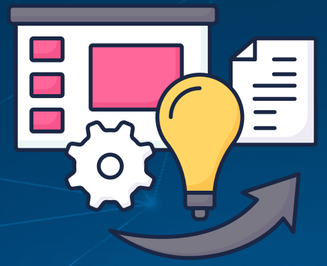
FALGUN GUPTA

VANSHITA GUPTA

COHORT PRESENTATION - COMPLETED

INTRODUCTION

- The goal of this project is to design and implement a search application for a Twitter dataset using both relational and non-relational datastores.
- The application will allow users to search for tweets by string, hashtag, and user, and also provide drill-down search features such as the ability to view other tweets by the same author or retweets of a particular tweet.
- To achieve this goal, the project will involve several steps including data collection, data storage design, cache implementation, and search application development.



METHODOLOGY

Data Exploration

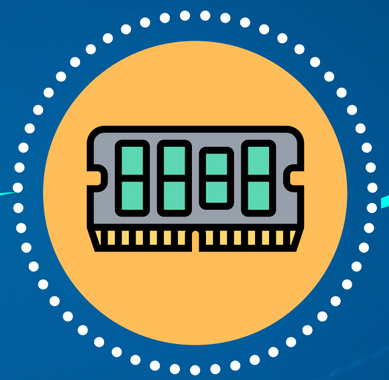
Involves thorough data exploration to understand the trend in the data and to get an idea how to structure the data load.

Data Load

Loading entire data in relational and non relational Database.

elasticsearch
Non Relational DB

PostgreSQL
Relational DB



Caching

Storing most queried tweet data in ram for faster retrieval of commonly queried inputs.

python™
Python Dictionary

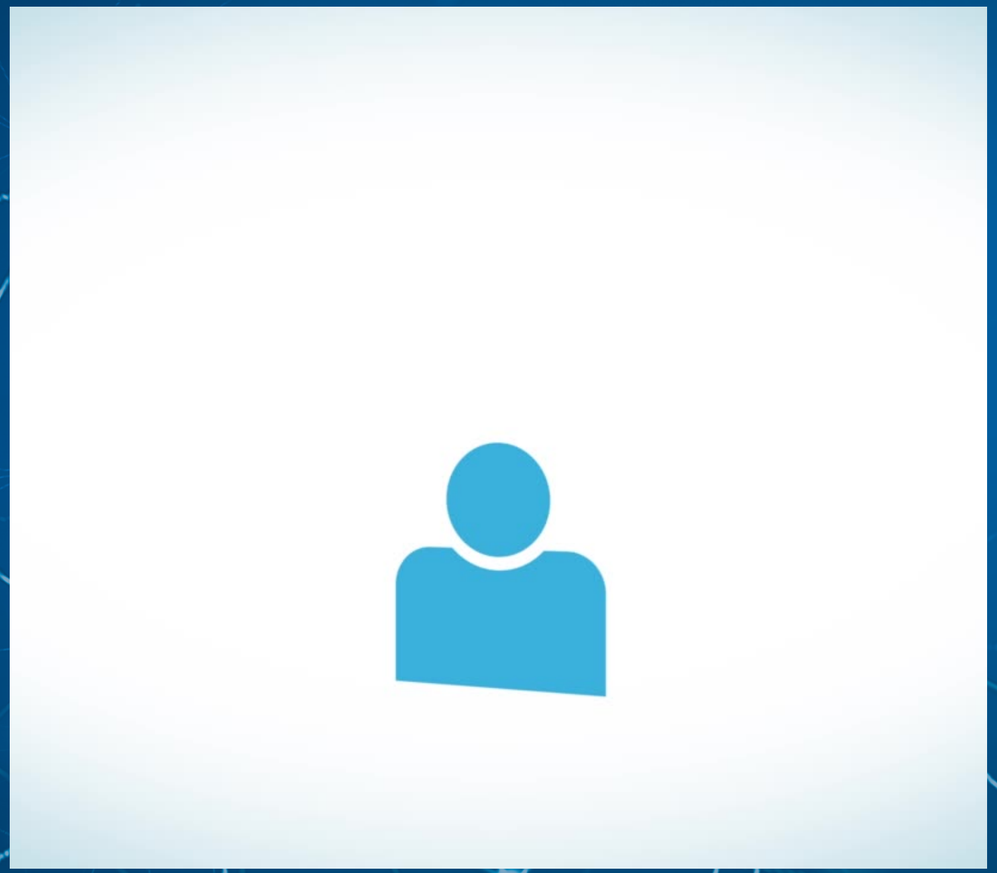
Search Application

The Crux of the entire application involving intelligent searches from cache and disk based on user query

Output on Frontend

Results an ouput on front end app.

DATASET



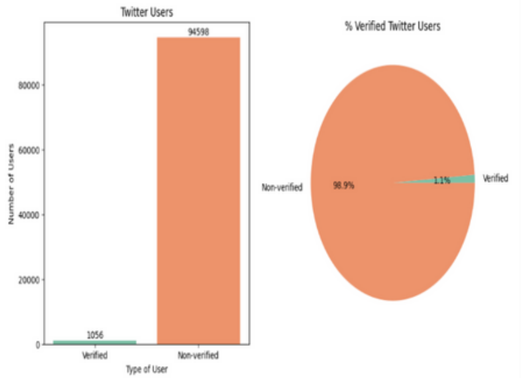
Key Snapshot

- Both static data file is used i.e. corona-out-2.zip and corona-out-3.zip
- Min Date-time of the tweet is : 2020-04-12 18:27:25+00:00
- Max Date-time of the tweet in is: 2020-04-25 14:48:38+00:00
- Total unique users in scope: 95,625
- Total tweets: 120,358

DATA EXPLORATION

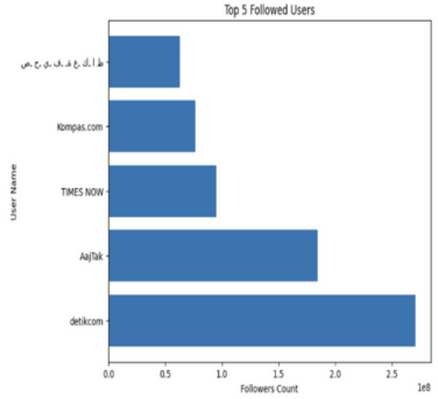
User Based

Split between verified and unverified users:



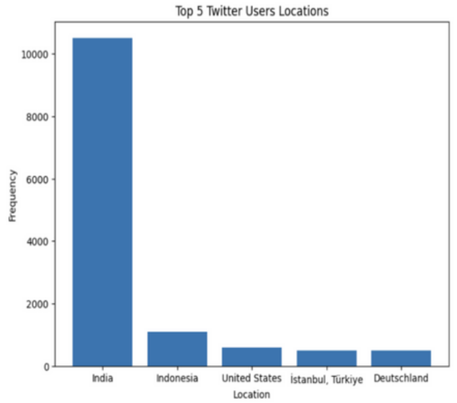
Takeaway: Around 98.9% of the twitter users are non-verified with just 1.1% of users enrolled as verified users.

Top 5 followed Twitter Users:



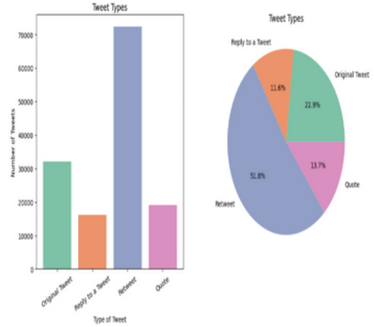
Takeaway: According to the data, maximum followers are for detikcom followed by Aajtak and Times now which are all digital media companies

Top twitter user location can be plotted as below:



Takeaway: Most twitter users in this data source are from India, followed by Indonesia and the United States.

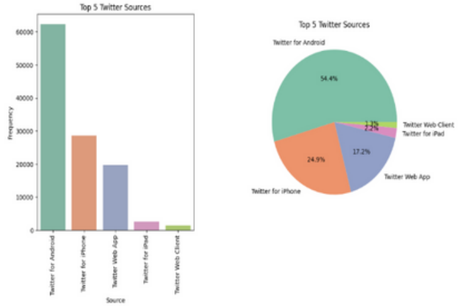
Below is the distribution of different types of twitter posts:



Takeaway: Maximum number of tweets in the data are retweets followed by original tweets, quotes and replies to different tweets.

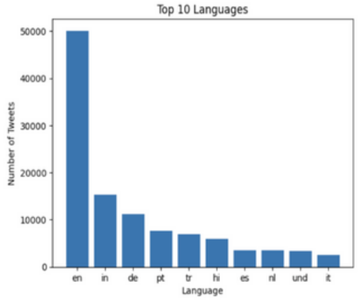
Tweet Based

Top 5 twitter sources used by Users are:



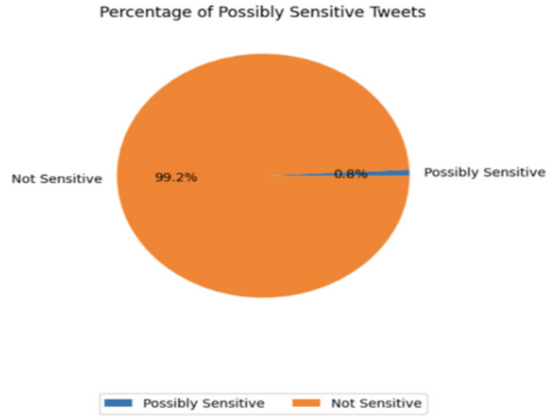
Takeaway: Top twitter sources are twitter app for android followed by Twitter app on IOS. Very few users use twitter on ipad and web client.

Top 10 twitter languages used by users in the data are:



Takeaway: English is the most used language on Twitter. German and Portuguese are also used by many twitter users followed by other languages.

Sensitive post count:



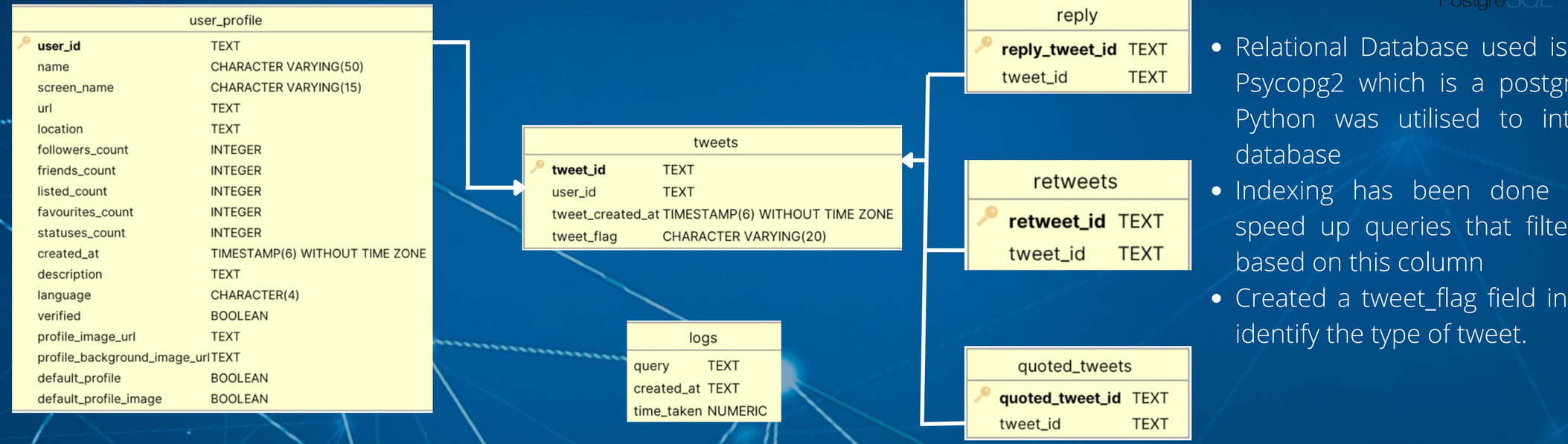
Takeaway: Around 0.8% of the total posts are marked as sensitive by twitter and the majority of the post are neutral in tone.

Wordcloud for hashtags:

Takeaway: Trending hashtags of the static tweeter data used are related to corona, covid-19 and lockdown.

DATA LOAD

Relational



- Relational Database used is Postgresql and Psycopg2 which is a postgresl adapter for Python was utilised to interact with our database
- Indexing has been done on tweet_id to speed up queries that filter or join tables based on this column
- Created a tweet_flag field in tweets table to identify the type of tweet.

DATA LOAD

elasticsearch

Non Relational

- Stored tweets data with the quoted tweets and retweets data in the Elasticsearch database
- Cleaned and formatted appropriately for indexing into Elasticsearch
- Stored 5000 JSON documents per index
- Elasticsearch calculates a relevance score for each search result based on frequency and proximity of search terms in the document
- Kibana for lucene query testing againt Elasticsearch indices
  -

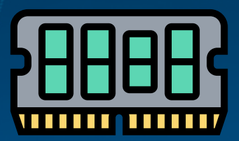
CACHING

Eviction Policy

1. Tradeoff: Up-to-date data Vs Search Time
2. Scalability: Does this app scale up well?
3. Are all cached items same?

- Least Recently Used
  - Maximum Items: 1024
  - Priority: Popular searches

- Time-To-Live
  - Based on the Search Time
    - 5 mins for < 5sec searches
    - 5 mins + Search Time for > 5sec searches
  - Priority: Time-consuming searches



CACHING

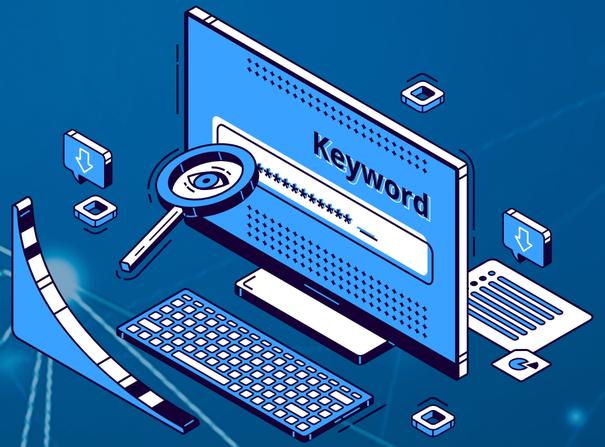
Cache Update Mechanism

- LRU update frequency: After every search request (Asynchronous)
- TTL-based update frequency: CRON scheduler to periodically(10 mins) remove stale data

Results

	Username	Tweet String	Hashtags	Date Range	Time Taken(Sec)
Without Cache	**Charles%**	**virus**	**corona**	**04/12/20 to 04/26/20**	**2.8550**
With Cache	**Charles%**	**virus**	**corona**	**04/12/20 to 04/26/20**	**0.0010**

SEARCH APPLICATION



ARCHITECTURE DIAGRAM

Front End (Flask)

Core Search Engine App

Cache (Python Dictionary)

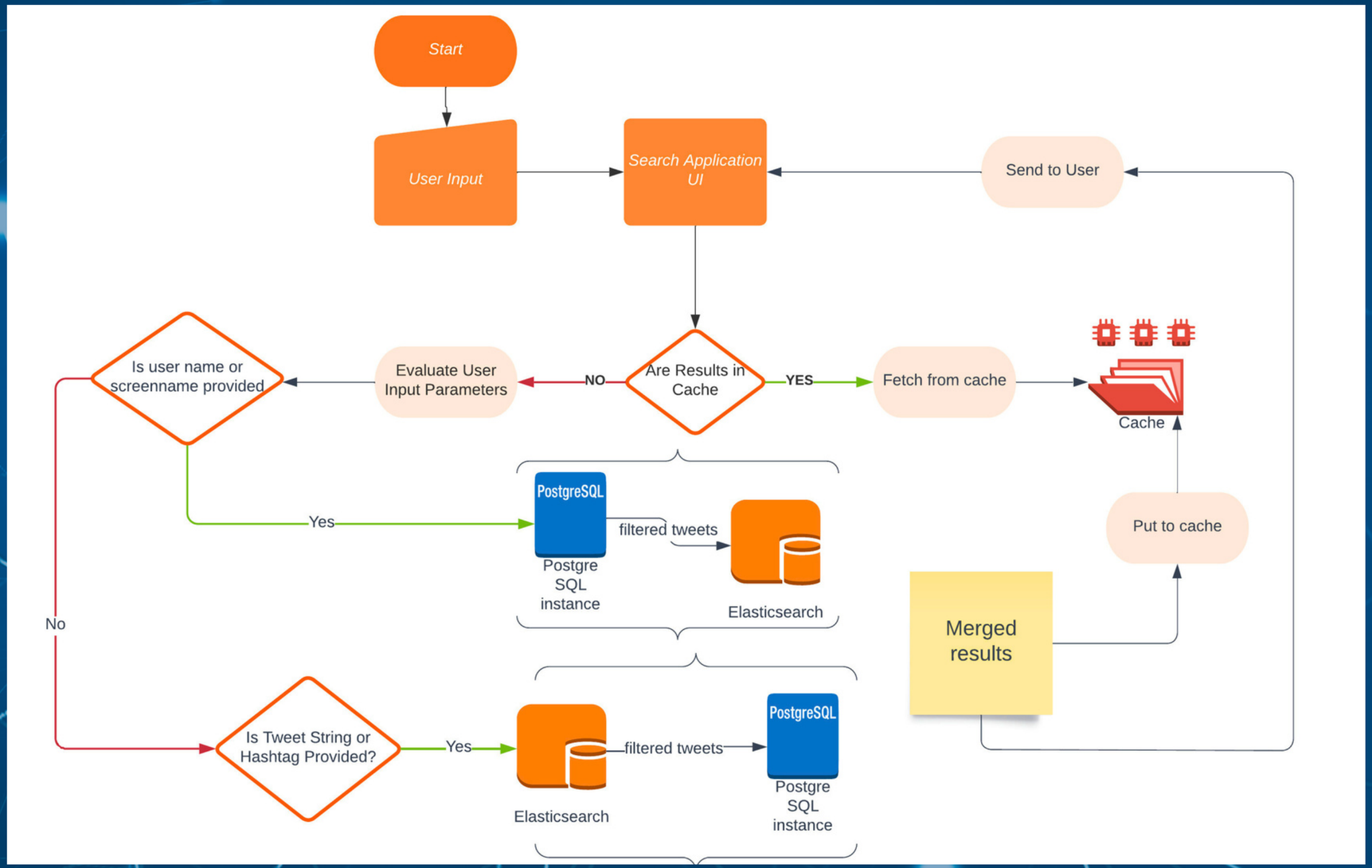
elasticsearch

Non Relational DB (Elasticsearch)

PostgreSQL

Relational DB (Postgres)

FLOW DIAGRAM

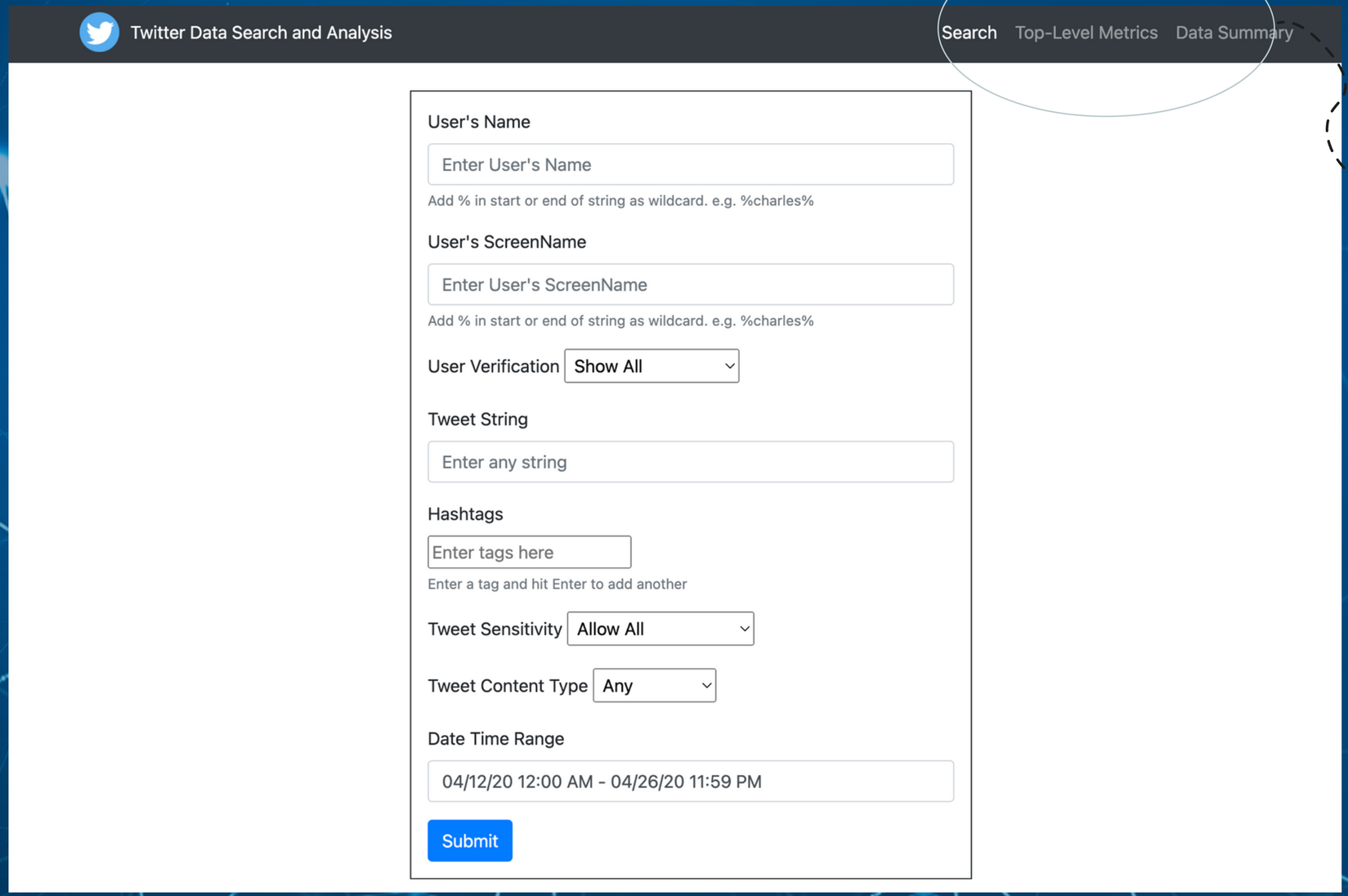


ELASTICSEARCH QUERIES AND RELEVANCE SCORE

- Elasticsearch provides a full Query DSL (Domain Specific Language) based on JSON to define queries.
- Tweet Data is searched on all indexes in Elasticsearch using an index pattern.
- For our project, results are ranked by relevance score based on 2 criteria: Text, Hashtags.
- Boosting: We also experimented with Unequal Weights that can be provided on each criterion.

UI WALKTHROUGH

TAB 1: SEARCH



Accessible Tabs

SAMPLE RESULTS WITH RELEVANCE SCORE

- Parameters:
  - Text: Corona
  - Hashtags: Biodiversity
  - Others: Username, Datetime Range

Show 10 entries Search:

	hashtags	_score	p
Ep1 Thanks to @fluglehrer @acrane79661230 @McnellisWilliam #corona #biodiversity	[corona, biodiversity]	2.607589	0
	NaN	1.788462	N
CMXLn0D	[NewProfilePic]	1.696595	0
id19 https://t.co/6i0VBNPmgW	[Corona, covid19]	1.671364	0
rona #COVID real mortality rates?	[COVID19, CoronaUpdate, corona, COVID]	1.666757	0
d van verzet! #CoronavirusLockdown #coronavirusNederland #Corona	[CoronavirusLockdown, coronavirusNederland, Corona]	1.659729	N
//t.co/rdB8LZpXOh\nvia TheFashnCollctr\n#VoteBlueNoMatterWho	[coronapocolypse, VoteBlueNoMatterWho]	1.653886	0
'Jullie ... - De Standaard Mobile https://t.co/0Sf18GDQPP	NaN	1.591029	0
recent investigation shows. https://t.co/TJ7TFRRmE8	NaN	1.586050	0
Rentenpaket trifft kaum einen Rentner! Umweltpaket trifft wied... https://t.co/4ww5OAv5Da	NaN	1.575789	N

Showing 1 to 10 of 36 entries Previous 1 2 3 4 Nex

DRILL DOWN - BASED ON TWEET

a) On tweet_id

Show 10 entries

	tweet_id	text
2	1249405407904358400	Tidur biar
1	1254023458079404032	मोदी और यो
3	1249404506963677184	In Turkey,
4	1254053635983585280	Okay, I am
5	1254028911291367424	अजय देवगन
6	1254032029005553664	Guys plz s
7	1254025545735692288	The depar
8	1254038275129409536	संक्रमित इला
9	1254046298107334656	Appeal to
10	1254038001044291584	Merkel jub

Showing 1 to 10 of 2,824 entries

b) On tweet metadata (counts)

Search:

retweet_count	quoted_count	reply_count
81	7	10
132	0	0
62	0	0
53	0	0
47	0	0
46	0	0
38	0	0
21	0	0
20	0	0
18	0	0

a) Clicking on Tweet_id : Provides Retweets, Quotes, Replies for given tweet
b) Clicking Metadata Counts: Provides tweets of that category

Total Time taken for search to run= {1.9826958179473877} seconds.

Retweets of this tweet:

Show 10 entries Search:

	tweet_id	text	hashtags	possibly_sensitive	media_type	media_url	user_id
1	1249405459871612928	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	117258830169
2	1249407185190187008	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	101964074152
3	1249407159051227136	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	10853243883(
4	1249407155628728320	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	117685922178
5	1249407115870887936	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	12488186334(
6	1249407069066682368	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	74342692
7	1249407046115418112	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	120837355408
8	1249407015215960064	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	863079014
9	1249406917442564096	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	2285814697
10	1249406821858598912	RT @andihiyat: Tidur biar besok punya tenaga buat ngelawan corona	NaN	NaN	NaN	NaN	119201757032

Showing 1 to 10 of 81 entries Previous 1 2 3 4 5 ... 9 Next

Quoted Tweets of this tweet:

Show 10 entries Search:

	tweet_id	text	hashtags	possibly_sensitive	media_type	media_url	user_id
1	1249406791462440960	Tapi lagi gabisa tidur ini mas	NaN	NaN	NaN	NaN	808095308

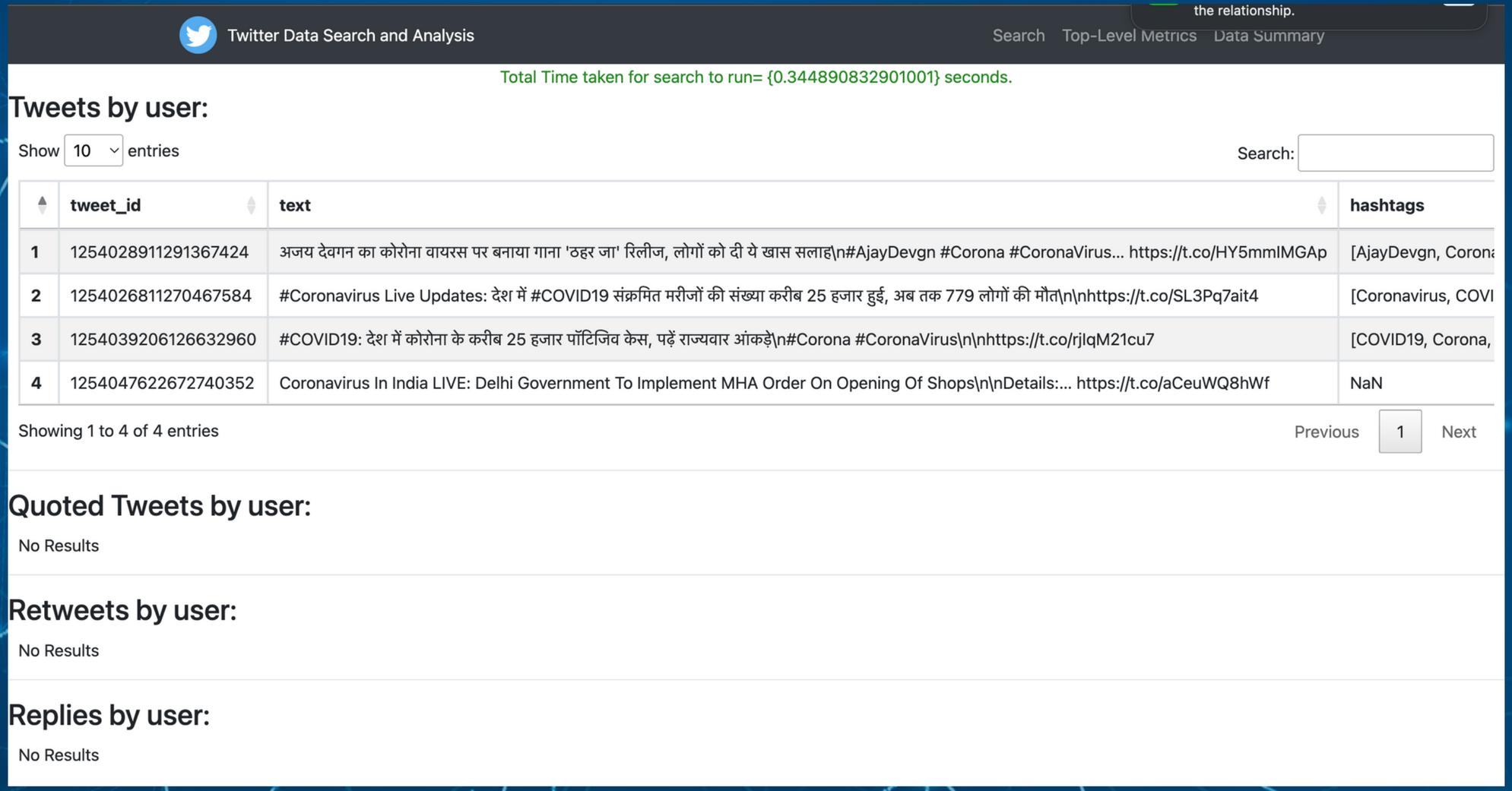
DRILL DOWN - BASED ON USER

Search:

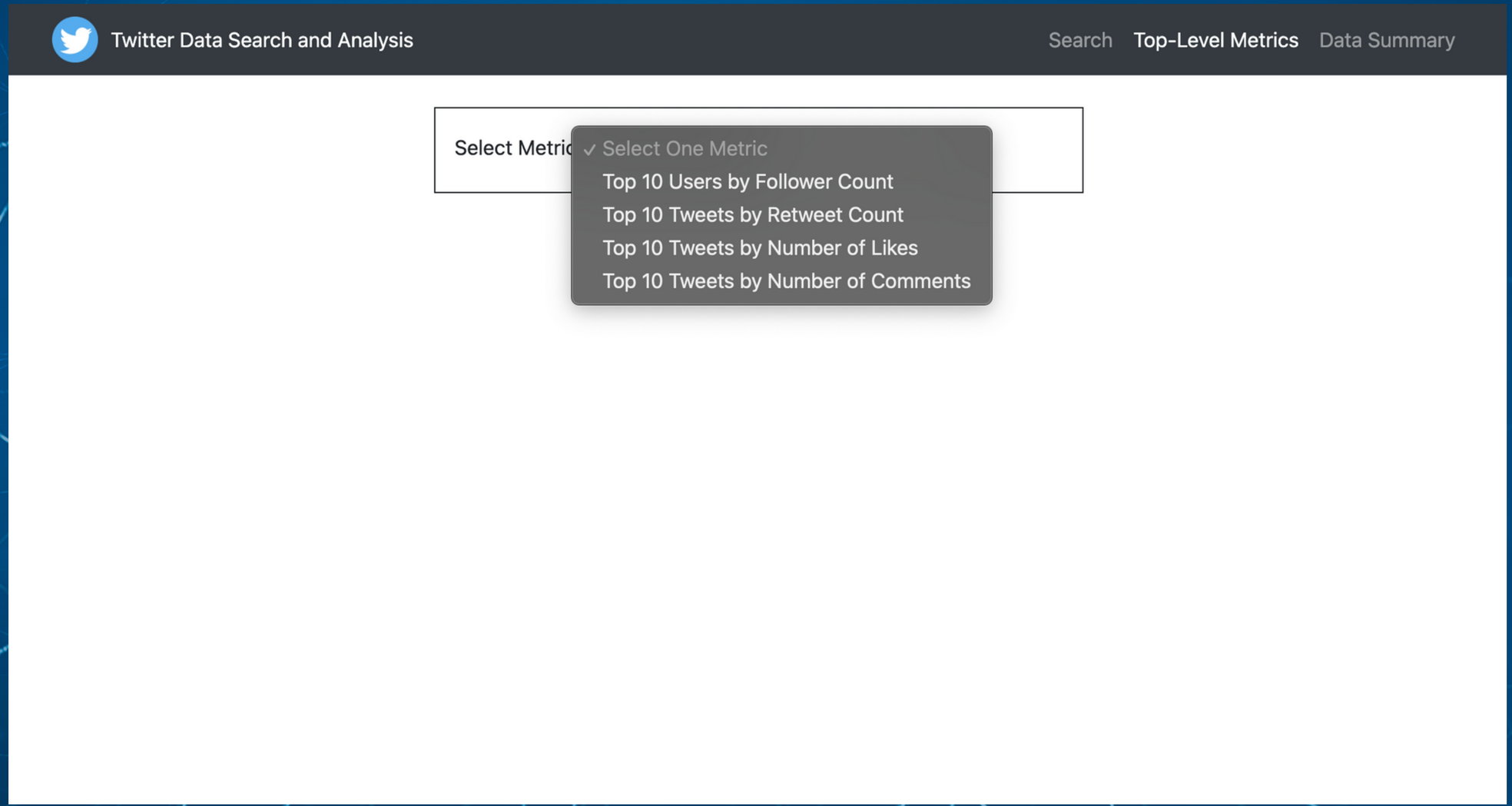
	user_id	tweet_created_at	tweet_flag	name	screen_name	verified
	886934161	2020-04-12 18:33:56	original_tweet	andihiyat	andihiyat	False
	42606652	2020-04-25 12:24:25	original_tweet	AajTak	aajtak	True
	764617494761971714	2020-04-12 18:30:21	original_tweet	ADEM YAVUZ ARSLAN	ademyarslan	True
	57947582	2020-04-25 14:24:20	original_tweet	Anubhav Sinha	anubhavsinha	True
	39240673	2020-04-25 12:46:05	original_tweet	ABP News	ABPNews	True
	70617758	2020-04-25 12:58:29	quoted_tweet	Anjali_Sharma	TribeccaAngie	False
	66018693	2020-04-25 12:32:43	original_tweet	Abdul Majeed Khan Marwat	koolkopper	False
	42606652	2020-04-25 13:23:18	original_tweet	AajTak	aajtak	True
	24516819	2020-04-25 13:55:11	original_tweet	Amalorpavanathan	amalorj	False
	1189660560343863300	2020-04-25 13:22:13	original_tweet	Artemi, Patriotin von Geburt	artemi_ecrit	False

Previous 1 2 3 4 5 ... 283 Next

Clicking on username, userscreenname, userid provides all other tweets by user



TAB 2: TOP LEVEL METRICS



Example: Top 10 Users by Followers Count

Twitter Data Search and Analysis

Search | Top-Level Metrics | Data Summary

Select Metric: Top 10 Users by Follower Count

Total Time taken for search to run= {0.2290036678314209} seconds.

Show 10 entries

Search:

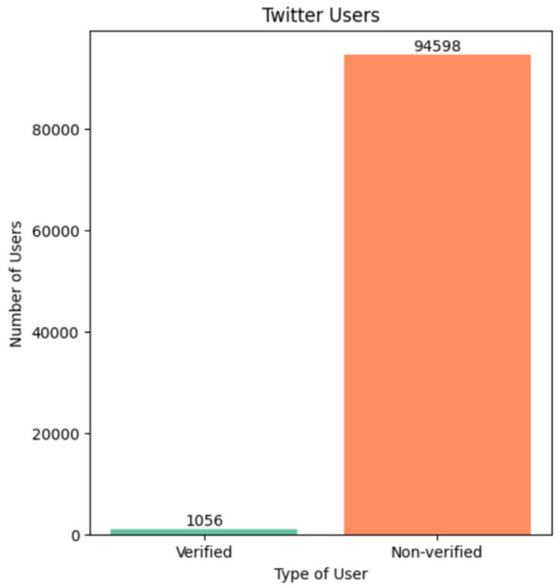
	user_id	name	screen_name	url	location	followers_count	friends_count	listed_count	favourites_cour
1	69183155	detikcom	detikcom	http://www.detik.com	Jakarta, Indonesia	15928061	28	13299	313
2	62513246	J.K. Rowling	jk_rowling	http://www.jkrowling.com	Scotland	14608046	721	37917	27353
3	42606652	AajTak	aajtak	http://www.aajtak.in	India	9706667	416	3516	16782
4	39240673	ABP News	ABPNews	http://abplive.com	India	9563509	248	3942	99
5	240649814	TIMES NOW	TimesNow	http://www.timesnownews.com	India	9499855	391	5110	4
6	56304605	Rajdeep Sardesai	sardesairajdeep	http://rajdeepsardesai.net/	New Delhi	8947464	568	8994	7598
7	24744541	Le Monde	lemondefr	https://www.lemonde.fr	Paris	8808784	628	36505	1609
8	55507370	tvOneNews	tvOneNews	http://tvonenews.com	Indonesia	8787649	50	8737	4347
9	23343960	Kompas.com	kompascom	http://www.kompas.com	Indonesia	7678373	23	10619	114
10	15016209	NTV	ntv	https://www.ntv.com.tr/	Turkey	7429200	29	5436	3

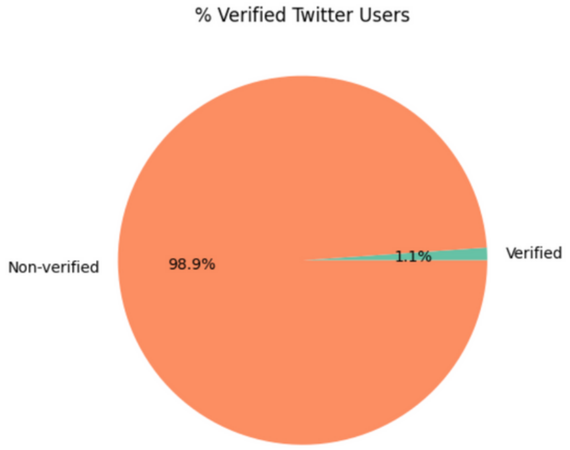
Showing 1 to 10 of 10 entries

Previous 1 Next

TAB 3: LINK TO DATA ANALYSIS

```
].sum()
num_nonverified = num_users - num_verified

# Plot bar chart
ax1.bar(["Verified", "Non-verified"], [num_verified, num_nonverified], color=palette)
ax1.set_title("Twitter Users")
ax1.set_xlabel("Type of User")
ax1.set_ylabel("Number of Users")

# Add data labels to bar chart
for rect in ax1.patches:
    x = rect.get_x() + rect.get_width() / 2
    y = rect.get_height()
    ax1.text(x, y, f"{y:.0f}", ha="center", va="bottom")

# Plot pie chart
types = ["Verified", "Non-verified"]
counts = [num_verified, num_nonverified]
ax2.pie(counts, labels=types, colors=palette, autopct="%1.1f%%")
ax2.set_title("% Verified Twitter Users")

# Show plot
plt.show()
```





Twitter

THANK YOU