# Log in to Present

Team 5

16:954:694 Database Management Systems

Log in

**Cohort Presentation Done**

Neeti Patel (np912)
Rushabh Karnavat (rpk68)

Aalap Patil (akp177)
Abhishek Bahad (ab2446)

Home

Data Model

Caching

User Interface

Tweet

# Home

## Objective

This project aims to design and implement a search application for a Twitter dataset that allows users to perform searches based on accounts, hashtags, keywords, time, and top-level metrics.

GIF

Tweet

## Data Storage

The data storage uses two different database management systems – a non-relational DBMS (MongoDB) for storing the tweets and comments and a relational DBMS (MySQL) for storing the user details and retweets.



## About the dataset :

- The Twitter dataset is a JSON file that contains information about users and their tweets.
- There are a few embedded JSON objects in the dataset that contain further details about the tweet and user.

Home

Data Model

Caching

User Interface

Tweet

How is the data stored?

# Data Model

**# Modeling and Storage in MySQL**

The user data and retweets are stored in MySQL. There are 90336 users in the dataset. The user table contains the user ID, username, followers, location, etc. There are 61090 retweets in the dataset. The retweet table contains tweet ids and the users who retweeted the tweet.

**# Modeling and Storage in MongoDB**

The tweets and comments are stored in MongoDB. There are 36304 tweets and 13606 comments in total. The tweets collection contains information like author, tweet ID, text, likes, etc. The comments collection contains information like in reply to tweet ID, in reply to user ID, likes, text, etc.

**# Popularity and Hashtags**

In the tweets and comments collections, two new fields were added – Popularity and Hashtags. The popularity field has been defined using the number of replies (+3), likes (+5), followers (+10), quote tweets (+20), retweets (+20), verified (+10000), and comments (+3) by giving weight to each factor. The hashtags field contains all the hashtags from the tweet.

Home

Data Model

Caching

User Interface

**Tweet**

# Speeding it Up!

## How?

A Python dictionary is used to implement cache. Whenever a search is made, the value present in the search box is stored as the key, and the search results are stored as values in cache.

In the Least Recently Used (LRU) technique, the least recently used element gets evicted when the cache is full, and a new element is stored in the cache.
The size of cache is 10.

On every insert in cache, we save the dictionary in a file for checkpointing. We initialize the cache by retrieving the data from the file if the file exists or else initialize the cache as an empty dictionary.

Indexing is applied on the tweet ID in the comments collection and on the user ID in the tweets collection to speed up the retrieval process.

## Time Improvement

After caching, the time is getting reduced by a factor of $10^4$

Home

Data Model

Caching

User Interface

Tweet

## Team 5

How does our search application look, and what searches does it facilitate? Use this link to access the video
https://drive.google.com/file/d/1Q1zJ7xRhzcle_ZhuxsRMhqkKrq3tq5Bi/view?usp=sharing

# # Going an Extra Mile

Sentiment Analysis for the tweets

According to the text in the tweets, sentiment analysis is performed using the TextBlob library.

**Here is how it works…**

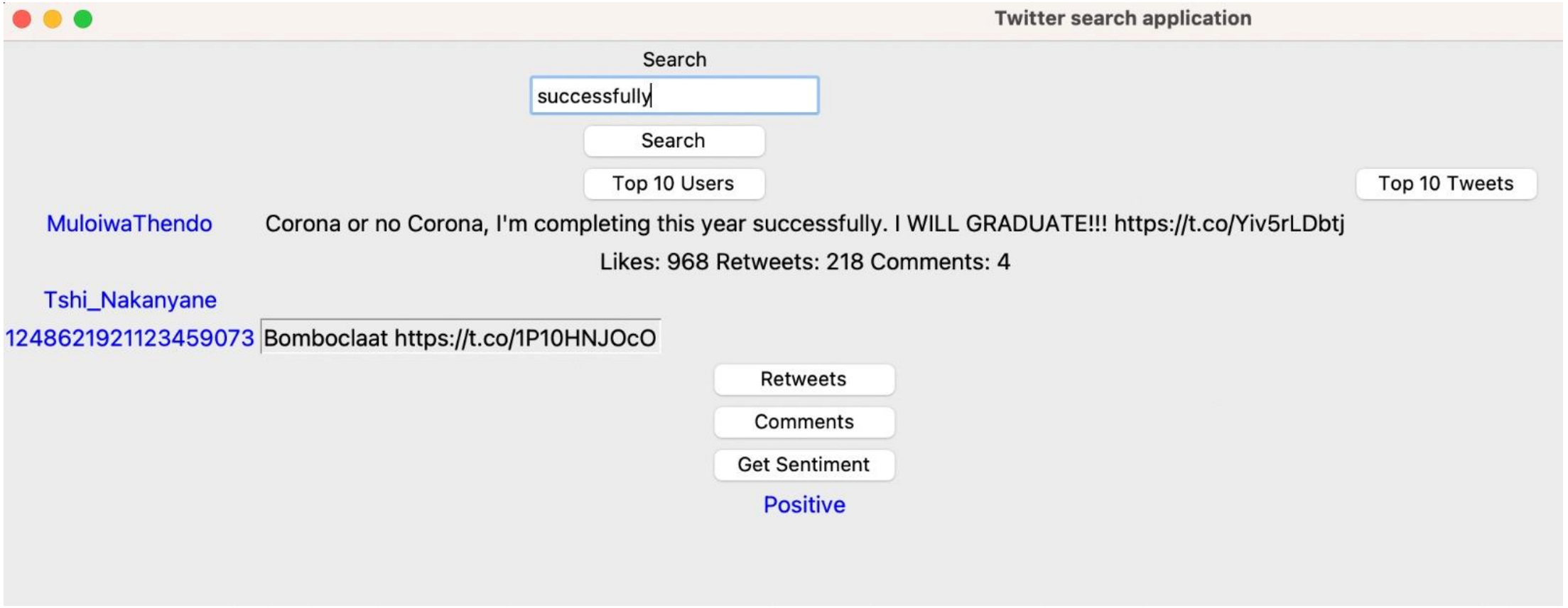
- The NLTK library is used to remove stop words from the text of each tweet.
- The sentiment polarity of each tweet is then calculated using TextBlob.
- Based on the polarity score, each tweet is categorized as either Positive, Negative, or Neutral.
- Finally, the sentiment and score values are stored in the database for each tweet.
- The search application returns the tweet's sentiment when the user presses the "Get Sentiment" button.

**It's not perfect…**

- It is a very basic Sentiment Analysis program, so it does not classify based on human emotions like joy, excitement, anger, etc.
- Tweets in other languages are classified as "Neutral" since the program cannot give a polarity score to words in other languages.
- The text in the tweets is often short so an accurate sentiment analysis is difficult.
- The program just evaluates tweets based on the content and not the context, which makes accurate predictions difficult.

# # Going an Extra Mile

Sentiment Analysis for the tweets



# # Going an Extra Mile

Sentiment Analysis for the tweets


Signing Out...
Team 5