**Lab Assignment 3**

**CIS 612 Big Data and Parallel Database Processing Systems**

**Sunnie Chung**

# Introduction

The main purpose of this project is to design and implement a program to collect a real time social media data from Twitter data stream, create a NoSQL Database in MongoDB, and process the JSON tweets into a relational database with NoSQL. It enables students to apply the knowledge learned in common logging structures of a social media server, how to collect a social media data stream in real time, JSON data processing with NoSQL MongoDB, and transformation strategies of semi-structured JSON data format to relational databases.

# Program architecture

The program is implemented as a command line script that executes the whole procedure from start to finish automatically. The program comprises of the following components listed below alongside their functionality:

**1.main.py**

-imports all the necessary libraries

-controls how and when all the other scripts execute

**2.streamlistener.py**

-handles authorization into Twitter using the Tweepy and Twitter API’s, and deals with the access tokens

-scrapes data from twitter by keywords to avoid unwanted data and exhausting your stream limit.

-monitor data streaming by streaming by returning a n error message when a stream is broken

**3.configparse.py**

-stores consumer API keys and access tokens hence protecting that data from unauthorized access.

-maintains a connection between the Twitter API stream and the MongoDB database.

**4.sql\_util.py**

-creates NoSQL tables from the JSON files in MongoDB database.

-insert records into the tables and provide the records for data processing

**5.mongodb\_util.py**

-create a MongoDB database for storing the tweets

-create separates collection for the Biden and Trump tweets

-make sure the tweets are unique by using an index

# Get started

The program besides being a command line executable will need access the Twitter API which will require you to register for Twitter’s developer account for access tokens. All other requirements besides are contained in requirments.txt file attached.

Installation pre-requisites

You can install the requirements by using the following steps:

1. Move your project to a predefined path of your preference.
2. Open the terminal in current directory
3. Cd to the Project root directory folder
4. Run the command:pip install -r requirements.txt

# How to run

Step 1: After signing up for Twitter’s Developer account and getting the access tokens for their API, replace

the ‘api\_key’ and ‘access\_token’ in the configparse.py file with actual API key and access token respectively.

Step 2: To execute the program run the following command I the command-line:

py main.py

When you run this command the following tasks are run:

1. The program logs into twitter social media server. User is notified in-case of an error in authentication
2. Scraping-the program scrapes Twitter for messages on presidential election with Donald Trump and Joe Biden using the keywords provided.
3. All tweets are saved locally into a MongoDB database. Only unique tweets are saved. Below is a screenshot of a Mongo database that was used in testing this program.

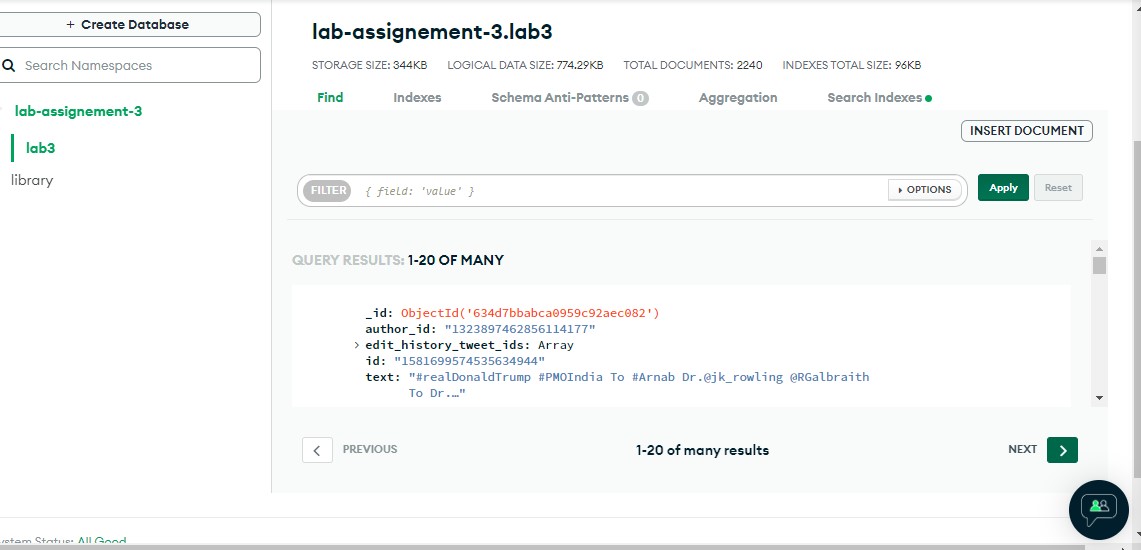


Fig 1.1: Mongo Database View after setup and Data Insertion

1. Fetching tweets-the program fetches all tweets for each president separately. This results in structured JSON data which the program saves in a json folder that resides in the public folder-located in the root folder. When this is executed you see the following in your predefined SQL database in the root-virtual environment folder.

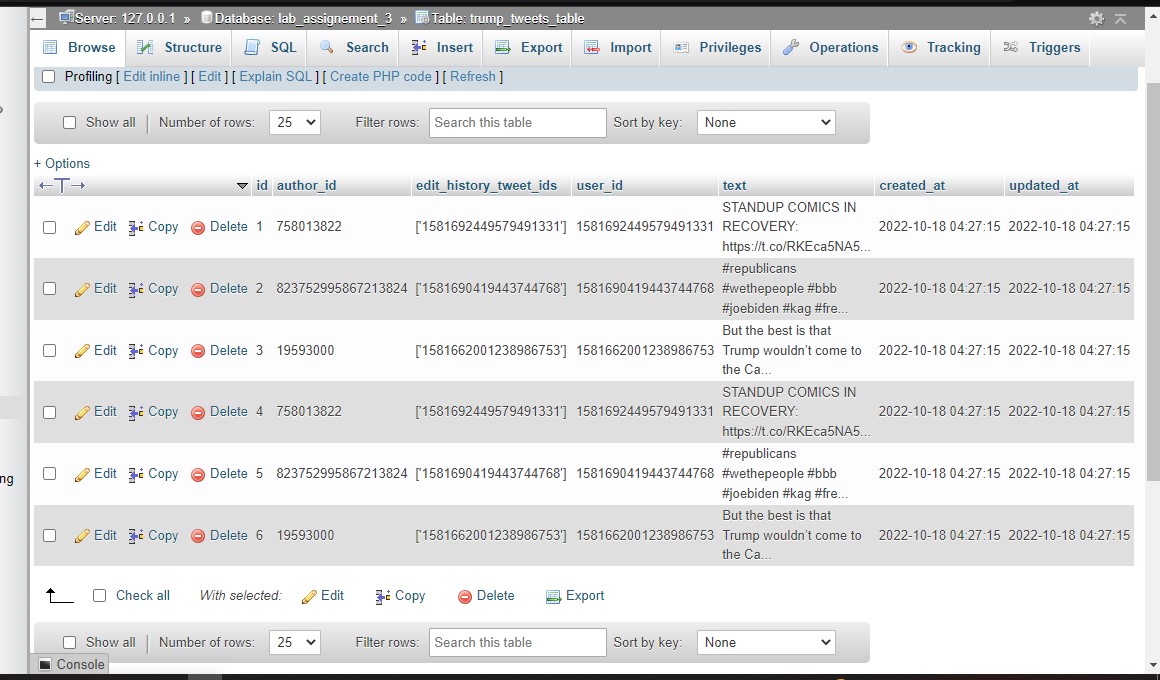


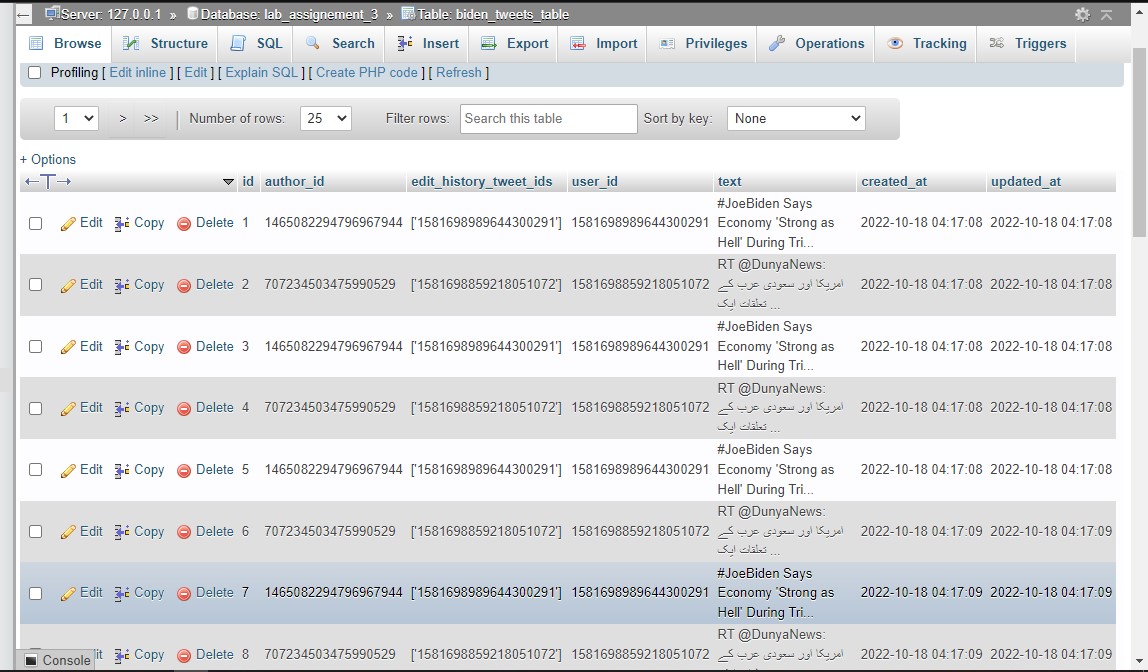
Fig 1.2: Donald Trump’s SQL Table

Fig 1.3: Joe Biden’s SQL Table

# References

<https://developer.twitter.com/en/docs/api-reference-index>

<https://docs.tweepy.org/en/v3.9.0/streaming_how_to.html>

th