CS 4525 Project Proposal

Nathan Annoh-Kwafo

November 2022

1 Project Aim

Extract data from sample datasets and insert the data into a well designed SQLite database using the techniques obtained from this class and its prerequisite. The data in the designated database via SQLite queries will be used for basic statistical analysis and visualized.
In particular but not limited, comparing and documenting devices used by twitter users, their geographic location and the time periods in years and months that recorded a large number of tweets.

2 Tools and Technologies

The choice of technologies below are subjected to changes.

- Python

  The entire project will primarily be built and manipulated in python.

- SQLite

  SQLite will be the initial database management system used to store the extracted data from datasets.

- Draw.io

  This graph design tool will be used to draw the required database schema and ER diagram.

- Visual Studio Code

  This will be the desired IDE used for the Project.

- GitHub

  GitHub is the preferred version control system used to keep track of the project's changes.

- MongoDB

  MongoDB will be the selected advanced database system used.

- Microsoft Power BI

  This final piece of technology will be used to display statistical data obtained from database queries.

- Data sources: CS4525 sample Twitter datasets

3 Overview

The main project python script extracts the compressed dataset files and loads its JSON contents into a python object. The loaded data is then inserted into an SQLite database for analysis using standard SQL queries. The results are then recorded and visualized using Microsoft Power BI. The entire process described above is then repeated with a large dataset using MongoDB as the database.