Mini Project-2 Report

System Overview

The system is a MongoDB based document store built using python language. The system has 2 phases. In phase-1 of the system, I have built a program (i.e.- load-json.py) that loads the json file data given to us by reading it and then it connects to mongoDB using the given port number and creates a DB called 291db. It then rebuilds the collection called articles, dropping the already existing one and then it inserts data in batches of 2000 so that the program runs smoothly. In phase-2 of the system, I have built a program (i.e.- phase2_query), this is where the user will interact with the document store by exploring the features of the program like searching news and articles, identifying common words based upon articles and many more through a menu-interface. It uses the database created in phase-1 of the system and then it uses the data to give out results. The system also handles various errors, handles ties and case insensitivity of the input

User Guide-

In the system,

Phase-1:

Run "python3 load-json.py chunk_1.json 27017" in the terminal.

Here chunk_1.json is the database file given to us and 27017 is the default port where we are running operations.

Phase-2:

Run "python3 phase2_query.py" in the terminal.

After running this, the user is prompted to enter the port number to connect to and run the operations.

Then the user is taken to the menu page from where the different functionalities of the program are displayed like:

1. Most Common Words by Media Type

2. Article Count Difference Between News and Blogs

3. Top 5 News Sources by Article Count (2015)

4. 5 Most Recent Articles by Source

5. Exit the Program

Choose from [1-5]

The user has to input from [1-5] in order to access that feature.

1. **Most Common Words by Media Type**
   - **Enter news or blog**
   - **The system prints a list of the top 5 most common words for that media type, along with the number of times each word appears.**

2. **Article Count Difference Between News and Blog**
   - **Enter a valid date in the correct format like "September 1, 2015".**
   - **The program shows how many news and blogs were published on that day along with the count and then compares which one is higher.**

3. **Top 5 News Sources by Article Count (2015)**
   - **No input required here**
   - **Displays top 5 news sources that published the most articles during the year 2015 along with the number of articles.**

4. **5 Most Recent Articles by Source**
   - **Enter a source name but please input the exact spacing between the words.**
   - **Displays the top 5 most recent articles from that source including the date on which it was published.**

5. **Exits the program**

Break-down Strategy:

Group-member(s)	Work Items	Time spent	Progress made
Mann Patel	Phase-1, Phase-2, Testing, Design document	17 hrs	Created phase-1, Creating functions for phase 2 including the whole system, did testing and created the document.

For the communication part and break-down strategy, since I was the only one in the group, no communication tools were used in this assignment and I was the only one doing the whole project.

No assumptions were made regarding any coding done outside of the requirements.