SI 507 Final Project

Chendi Zhao

Uniqname: zhaocd

**I.Overview**

This project is a tour guide system that utilizes data from Reddit (<https://www.reddit.com> ) to provide information about Michigan university subreddits. The system provides an interactive command line interface for users to retrieve and explore data related to university subreddits. The data is presented in the form of a data frame, line chart, and summary information such as recent posts, top posts, and most commonly used words in post titles or text. The system also offers options to save the data and combine it with previously scraped data.

**II. Project Code**

Github repository: <https://github.com/SandyZ98/SI-507-Final-Proj>

Included in this repository are the data source cache, consisting of one .json file, text files in the form of one .txt file, and source code files in the form of six .py files for the SI 507 final project. The current system does not require API keys (<https://www.reddit.com/dev/api> ) as most of the data has already been stored in the cache. However, a Reddit API must be applied to access the data, and if a new key is not applied, the system will use the default one. This program uses Python packages such as requests, json, pandas, praw, datetime, nltk, ssl, matplotlib, string, and webbrowser. The majority of these packages are built-in. To execute the program, simply run RedditInterface.py and follow the interactive instructions provided.

III.Data Source and Structure

Reddit, one of the most widely used social platforms in the US, is the source of the data. Once the data is obtained, it can be saved as a json file and processed as a Python dictionary. The post information that is scraped includes the title, post text, user ID, post score, upvote ratio, total number of comments, creation date, post URL, post flair, and subreddit. Figure 1 provides a snapshot of the cache file as evidence of caching. The size of the data depends on the reference time that the user wants to retrieve from. For example, the combined\_example.json file in the repository contains around 2000 posts.

A screenshot of a computer screen

Description automatically generated with low confidence

Figure 1

IV.Interaction and Presentation Options

The system utilizes command line prompts for user interaction. Users are prompted to input the starting date and the university subreddit they want to retrieve data from using the UserInterfaceReddit function, which offers a list of five university names to choose from. Data can be presented in the form of a data frame, and users can view the trend of post numbers over time through a line chart. Users are then directed to choose between viewing data from post titles or the main post text. This selection is facilitated by the TitleInterface and PostInterface functions. Within both interfaces, users can opt to view either the data or a summary of it, such as the five most recent posts, the top five posts with the most comments, or the most commonly used words in the title or text. Additionally, the frequency of the words can be displayed in a bar chart if desired. Next, users can choose to save the data as a single file or combine it with previously scraped data using the CombineInterface. Finally, users can choose to write the data out as a json file or not. At every prompt, users have the option to exit the system.

V.Demo Link