SI 507 Final Project
Winter 2023
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.

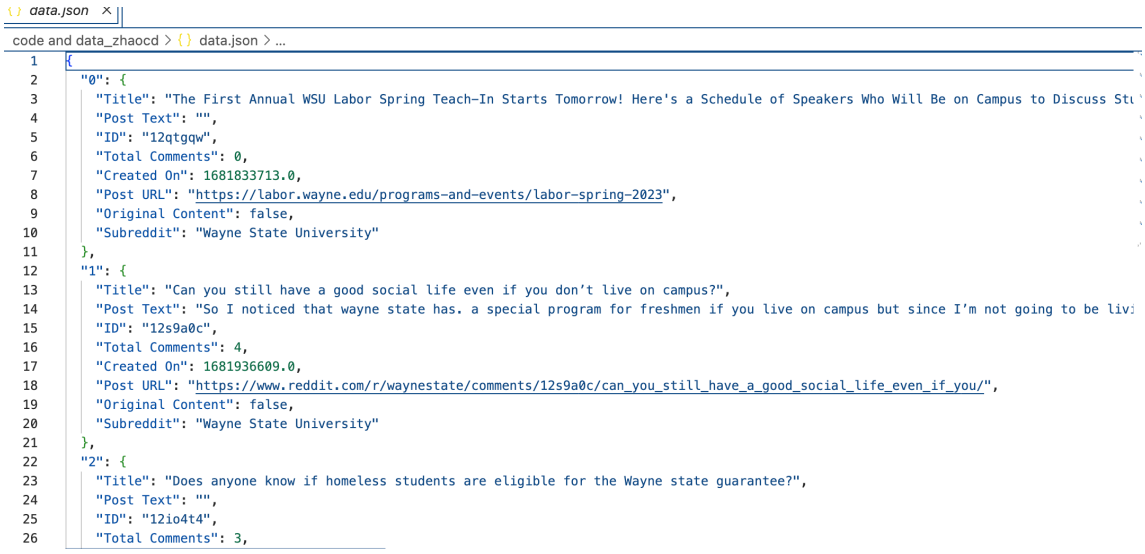
II. Project Code

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

Included in this repository are the data source cache, consisting of two json files and six source code files in the form of.py files for the SI 507 final project. The current system require API keys (https://www.reddit.com/dev/api) and 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 os, json, pandas, praw, datetime, nltk, ssl, matplotlib, string, and webbrowser. The majority of them are built-in packages. 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 single 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 and Figure 2 provide snapshots 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_posts.json file in the repository contains approximately 1500 posts.

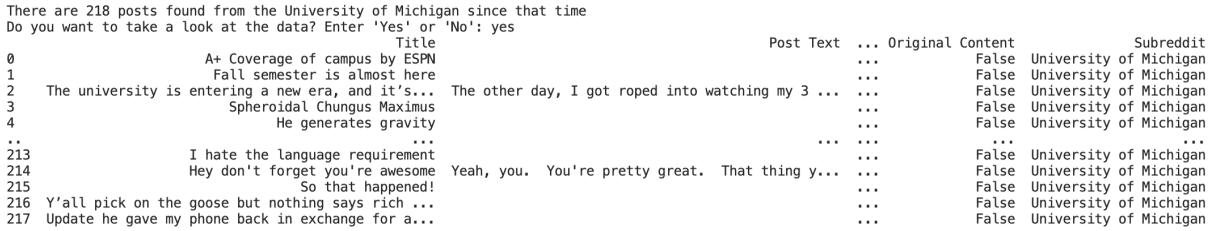


data.json

code and data_zhaocd > {} data.json > ...


```
{
  "0": {
    "Title": "The First Annual WSU Labor Spring Teach-In Starts Tomorrow! Here's a Schedule of Speakers Who Will Be on Campus to Discuss Stu
    "Post Text": "",
    "ID": "12qtgqw",
    "Total Comments": 0,
    "Created On": 1681833713.0,
    "Post URL": "https://labor.wayne.edu/programs-and-events/labor-spring-2023",
    "Original Content": false,
    "Subreddit": "Wayne State University"
  },
  "1": {
    "Title": "Can you still have a good social life even if you don't live on campus?",
    "Post Text": "So I noticed that wayne state has. a special program for freshmen if you live on campus but since I'm not going to be liv
    "ID": "12s9a0c",
    "Total Comments": 4,
    "Created On": 1681936609.0,
    "Post URL": "https://www.reddit.com/r/waynestate/comments/12s9a0c/can_you_still_have_a_good_social_life_even_if_you/",
    "Original Content": false,
    "Subreddit": "Wayne State University"
  },
  "2": {
    "Title": "Does anyone know if homeless students are eligible for the Wayne state guarantee?",
    "Post Text": "",
    "ID": "12io4t4",
    "Total Comments": 3,
```


Figure 1




```
There are 218 posts found from the University of Michigan since that time
Do you want to take a look at the data? Enter 'Yes' or 'No': yes
                                                 Title                                                Post Text  ... Original Content               Subreddit
0                                  A+ Coverage of campus by ESPN                                                ...            False  University of Michigan
1                                   Fall semester is almost here                                                ...            False  University of Michigan
2    The university is entering a new era, and it's...  The other day, I got roped into watching my 3 ...  ...            False  University of Michigan
3                                      Spheroidal Chungus Maximus                                                ...            False  University of Michigan
4                                            He generates gravity                                                ...            False  University of Michigan
..                                                            ...                                                ...  ...              ...                     ...
213                            I hate the language requirement                                                ...            False  University of Michigan
214                            Hey don't forget you're awesome  Yeah, you.  You're pretty great.  That thing y...  ...            False  University of Michigan
215                                        So that happened!                                                ...            False  University of Michigan
216  Y'all pick on the goose but nothing says rich ...                                                ...            False  University of Michigan
217  Update he gave my phone back in exchange for a...                                                ...            False  University of Michigan
```


Figure 2

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 six university names to choose from. The syetem will then convert the date to a Unix timestamp, which represents the number of seconds since January 1, 1970. 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. After that, 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 (see examples in Figure3). 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.

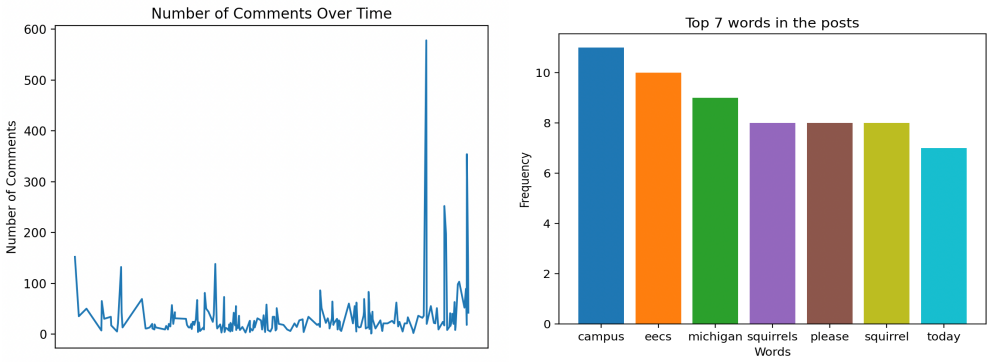


Figure 3 the university of Michigan (08/01/2022-now)

V. Demo Link

https://drive.google.com/file/d/1hK9MGyQFBP61cgnh2z1V_jx-Fz8FLH7E/view?usp=sharing

In the demo video, data is scraped only from the last two weeks, considering the time constraints of the video.