Homework 3: Python Web Crawler

Produced by: Aditi Rajagopal, Bradley Katcher, and Charlie Putnam[1]
Computing IDs: ar5vt, bk5pu, cmp2cz
Note: Sources are cited as hyperlinks through the document.

Introduction

We live in an age where commentary is abundant. The home for commentary and discussion for any topic under the sun is reddit.com. Users can create discussion forums dedicated to certain topics called "subreddits." For the scope of this project, we would like to scrape the data from a popular subreddit for College Football - commonly known as /r/CFB. Based on the data we pulled using python packages BeautifulSoup, and pandas, we were able to conduct exploratory data and answer some questions based on a dataset of the 100 most recent posts:

- When do posts occur during the day?
- What types of posts are the most common?
- Which fanbase posts most frequently?

Approach

The primary reasons we decided to pull data from Reddit rather than other websites that contain social data is the ability to avoid logging in as a user, and to avoid using an external API. We wanted to limit ourselves to scraping with common python libraries, specifically BeautifulSoup.

We decided to pull data from /r/CFB rather than other subreddits because of personal interest, and the time of year -- the college football season starts in 40 days, and there is quite a lot of interest in a broad range of topics related to college football (including recruiting, coaching changes, conference projections, etc). In addition, /r/CFB users can select their program affiliations via the Reddit "flair," or user-specified metadata (image, text).

Considering the questions that we wanted to answer, we decided to pull the following information from subreddit posts (as highlighted below): Post Title, Post Source, Direct Link, Post Type, Author, Flair, and Timestamp.

[1] We received permission from Professor Basit to work in a group of 3

Our project includes the requests, BeautifulSoup, and pandas libraries. We use requests to make the initial request to the user-specified[2] subreddit - we also provide a header to "declare ourselves" in the request. Without this header, web-scraping on Reddit will not work. Once we have a successful request, we use BeautifulSoup to both store and parse the data from the subreddit home page. As aforementioned, we parse the individual posts, and we also make sure to grab the link to the "next page" in the subreddit. This is because subreddits only include 23-25 posts per page. Using the "next page" link to parse more pages until we (roughly) hit a user-specified number of posts.

When analyzing our code, you'll notice that we pulled data from old.reddit.com rather than reddit.com - this is because the "old" Reddit UI more prominently features user flairs, uses pagination (BeautifulSoup does not work seamlessly with websites that use infinite scrolling), and is overall simpler in it's UI implementation (uses fewer elements).

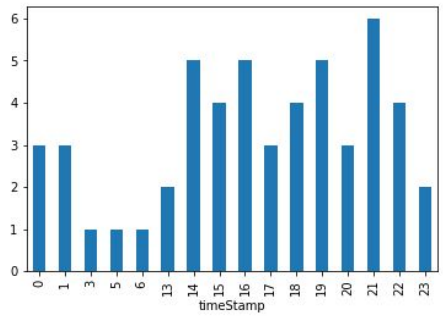
Useability

In creating this web-scraper, a driving force was usability and interoperability. On the client-side, we wanted the user to be able to scrape and store data from any subreddit with ease. In addition, we pulled data that would be useful to any subreddit user. On the development side, we wanted to minimize code redundancy (i.e creating a flexible SubRedditParse class).

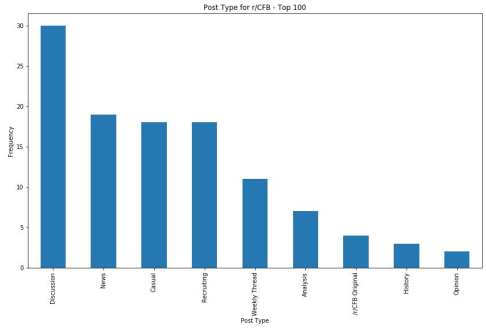
By using the pandas python library, and exporting the parsed data into a csv, we give the user an open-ended, but an easy way to chart and analyze their favorite subreddits (as we explored above). Some reasons users may find this program useful include:

- They can look at the timestamp data to gauge how active a subreddit as a deciding factor as to whether they want to join that subreddit
- They can look at the types of posts over time to decide what type of content get traction (using the "hot," and "top" sorting parameters)
- In the case of sports-related subreddits, they can evaluate what teams are the most active or most represented in posts

Data



Question One: "Time" - When do posts occur during the day? (timeStamp buckets are on the hour)

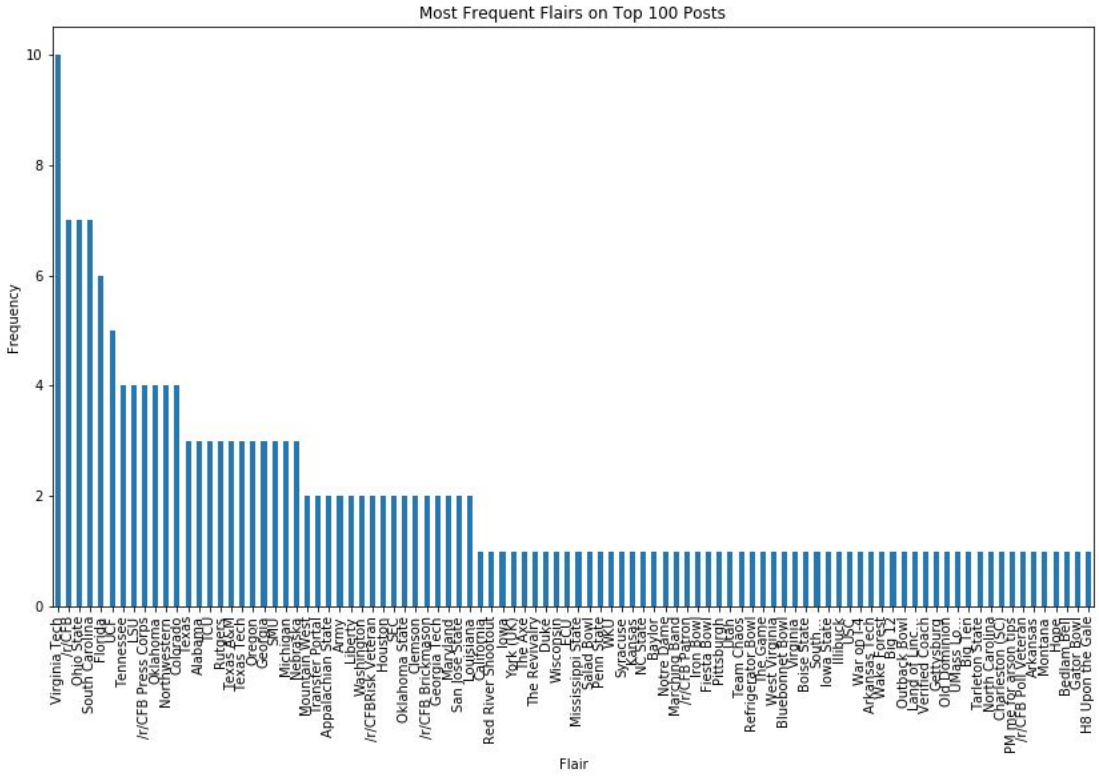


Question Two: "Content" - What types of posts are most common in the top 100?

[2] The code is written such that it can scrape from any subreddit, the user can select which to use (r/CFB in our case).

It seems that the top posts on Reddit's r/CFB subreddit occur between roughly 2pm-10pm, and the most common top posts are discussion posts, by a wide margin!

Question Three: "Flair" - Which fanbases post most frequently?



We were surprised to see that Virginia Tech had the highest representation! We were expecting to see flairs from a bigger school/a school with a more successful football program. (Go Hoos!)

Next Steps

In terms of next steps, we could:

- Broaden the dataset to look at upvotes and post comments/activity
- Incorporate sentiment analysis to evaluate the types of posts on a given subreddit (we have a prototype using the vaderSentiment package, but we have not evaluated the usefulness of the output data)
- Create a scoring mechanism to decide what subreddits are the most/least active