# Reddit Scraper - Subreddit Statistics

Authors: Trevor Carpenter, Xiaoqi Na, Kane Wu, Jeanie Liu

Introduction:

The purpose of this project is to introduce a shiny application that will provide various statistics for the social media site Reddit.com. Reddit is a discussion and forum style website that is known for holding a widespread amount of information. Members of the Reddit community submit content to the site through text posts and images, which are then voted up or down by other members. Posts are organized by subject into user-created boards called "subreddits", which cover a variety of topics such as news, movies, video games, music, books, health, science, food, and more. Subreddits are generally denoted as the values following an r/ on the site, for example the “UC Davis” subreddit is r/UCDavis. Submissions with more up-votes appear toward the top of their subreddit and, if they receive enough up-votes, ultimately on the site's front page.

The Shiny Application “Reddit Scraper” provides statistics on any entered subreddit. Through these statistics, users can manipulate their own posts on the subreddit to try and have a higher probability of a post gaining more upvotes and ultimately being seen by more users. The various pieces of statistical data gathered by the application can be useful for advertisers, activists, and general community members who want to share their thoughts with a greater audience.

Source Data:

The data from the project comes from an individual subreddit’s website. When the user enters in a subreddit they would like information on the application links to the url:

https://www.reddit.com/r/ + the subreddit name

The application then scrolls to the bottom of the page in order to load more material 15 times, in order to get data for the top 15 pages worth of reddit posts. This process takes approximately 30 seconds, 2 per iteration so that the page has time to load the new posts before scrolling again. The html from the page is then scraped and parsed.

An interesting note about the way Reddit organizes their html is that they do not have user friendly names for their html classes. As a result, parsing through the html was initially difficult until the correct class names were found, although this was a strong exercise for using the inspection tool for web scraping. The html is parsed into nodes of `<div class = “Post”>`, and then a for loop is used to parse out the parts of each individual post. The Content

and Pictures parts were split from a more generalized “General Content” type. Posts were divided as follows:

| Content Type | HTML Tag | HTML Class |
|---|---|---|
| Title | h3 | _eYtD2XCVieq6emjKBH3m |
| Upvotes | div | _1rZYMD_4xY3gRcSS3p8ODO |
| “Promoted” Label | span | _2oEYZXchPfHwcf9mTMGMg8 |
| General Content | div | STit0dLageRsa2yR4te_b |
| *from General Content* | | |
| Content Text | p | any |
| Content Picture | img | any |

For the data used, the application does not need the actual content of the Promoted Label or Content Picture, as these would simply be the word “Promoted” or a link to the local picture on the site’s server. Thus, only the boolean qualitative variable is added to the data table when these tags had content.

After this web scraping, the final format of the data is a table of posts similar to the following example of the first five entries of the data table from r/Sacramento:

| | Title | Upvotes | Content | Pics | Promoted |
|---|---|---|---|---|---|
| 1 | Please help Sacramento Food Bank | 305 | No Text | FALSE | FALSE |
| 2 | Raleys will be offering a discounted bag of pre-selected food ($20) for seniors and at risk people sheltering in place. Details in the link. | 269 | No Text | FALSE | FALSE |

| | | | | | |
|---|---|---|---|---|---|
| 3 | Become a digital marketing professional. Gain hands-on strategy and advertising experience including campaign development, SEO, and more. Classes starting soon at UC Davis Digital Marketing Boot Camp — apply today! | 1 | No Text | FALSE | TRUE |
| 4 | Now I have an excuse... | 591 | No Text | TRUE | FALSE |
| 5 | Shout out to the Madison Bingo Hall in Fair Oaks for hosting +/-100 elderly folks last night even though they're all recommended to be home, even in the face of news a caller (volunteer) was tested positive. It's just THIS sort of thinking we need rn! /s | 83 | They can be reached to affirm this amazing decision at either 916-965-7800 or via email at fastpitch@allstartournaments.com. It's time to start holding irresponsible businesses accountable! | FALSE | FALSE |
| ... | | | | | |

In the data table, we related sentiments according to titles. There are five items total in the table: Title Sentiment, Title, Content, Pictures, Upvotes. This is done by relating the words in each title to the "Bing" sentiments. A boxplot of the number of upvotes related to the general sentiment of the words is also created for the user to analyze.

Users can also see the relationship between the number of upvotes with the contexts which those reviewers posted on the web. We used the shiny application to generate the histogram and word cloud of frequency words in the contexts from randomly different ranges of upvote numbers. Hence, by using this shiny application, users can more clearly find frequent words in different upvotes intervals.

Users can also visually see what factors are related to the number of upvotes gained on reddit. A scatter plot is generated showing the effect of title length on average number of upvotes for a given subreddit. A barplot is generated showing the effect of picture status (post contains a picture or not) on the number of upvotes. Both of these plots are created from simple data table manipulation and plotly.

<u>User Guide:</u>

This application utilizes a local port, Port 4565 in order to create a server, so that port must be clear in order to utilize the application.

To begin, there are two panels at the top left of the application. The first is the subreddit to get data from, and the second is where you can enter a “wait time”. The wait time is optional, however if you would like to gather more data, enter an approximate time, in seconds, you are okay waiting for that data for. If you do not specify a wait time, the program, depending on the internet speed and other factors, should generate the data in around 5 seconds. Then, hit the “Go!” button to begin.

Once the data has been gathered, many different statistics and plots will appear from the website inside of the different tabs. These are divided into five sections: Title, Content, Sentiment, Picture, and Ads.

In the “Title” tab, a scatterplot comparing title length and upvotes is shown. This can be used to see if there is a trend in using longer or shorter title lengths in your post, and if that will provide more upvotes.

In the “Content” tab, you can use the sliders to set a range of the upvotes you are looking for. Within that range, the word cloud and histogram are two graphics that will show the most frequent words used in the subreddit.

In the “Sentiments” tab, first there are three box-whisker plots displayed. These plots have outliers removed so that you can clearly see the difference in means and variances of the upvotes when positive, negative, or neutral posts are submitted. Below is a list of the actual data point entries as examples, that can be filtered using the dropdowns and search bars. Each data entry, which are the real posts from the queried subreddit, are marked as “positive”, “negative”, or “neutral”. Users can enter any keywords in the upper right corner space after “Search:” to see all information that contains the keywords they enter.

In the “Picture” tab and the “Ad” tab, there are bar charts comparing the average number of upvotes for containing a picture or being an Ad or not. This is useful both for general users and promoters wanting to know how to get their advertisement or post seen by more users.