# README for Course project for CSE 6240: Web Search and Text Mining, Spring 2020

**Data Extraction:**

The post submissions were limited to the top 1000 posts each year since the inception of the two subreddits until 2020. For example, if a subreddit was created in July 2010, the top 1000 posts for each year until July 2020 were scrapped. Pushift.io was used to get the top links. Using Reddit’s API PRAW, the comments under each of these posts were scrapped. For efficient usage of time, only the comments were scrapped, and not the individual replies to the comments. Along with the posts and submissions, we also extracted the user information, the comment author information, and creation of the post/comment. Once, the offending users were identified, the top controversial subreddits they post in were also identified. To comply with Reddit’s API request limit of 30 requests per minute, multiple Reddit accounts were created and multiprocessing was used to better utilize time.

**Data Processing:**

**Dataset:** For each community, there are two datasets that can be processed. Subreddit\_posts.csv contains the metadata of posts and links to comments made on the posts. Subreddit\_comments.csv contains the metadata of the comments made on each of the posts in subreddit\_posts.csv

**Data Pre-Processing:** Data\_Processing\_1.ipynb has been used for AskReddit, which was used in the midterm-report. The rest of the datasets are processed using Data\_Processing\_2.ipynb. Data\_Processing\_1.ipynb has 2 sections. In the first section, data from AskReddit is mounted and processing is done. The same goes for the second notebook.

It starts with uploading the datasets onto the workspace. Unnecessary columns are dropped. The columns indicating date-time are converted to date-time objects from Unix Epoch time for easier sorting. The comments are then processed to remove non-letter alphabets. The comments section is stored as a list of words contained in the comments of each post. Some extra columns are added to keep track of misogynistic words. Statistics of the dataset are displayed.

**Functions:**

*Fill\_table():* To fill the table with frequency counts of misogynistic words for each category.

*Plot\_graph():* Used to plot the graph of date-time vs frequency of misogynistic words used per category

*Display\_most\_common\_words():* Display a table of the top 5 most frequently used words for each category

*Display\_perc\_table():* Display the table of percentage of misogynous posts per category

*Plot\_above\_18\_graph():* Display the frequency of mis. Words encountered by over 18 and under 18 age groups

*Plot\_score\_\_negativity():* Plot the graph of fraction of mis. Words in the comments section vs the score of the post

**Comment Processing:** In this section, the comments from subreddit\_comments.csv are processed to find the frequency of misogynous words used. And then, the table is grouped by the username. The users are then sorted according to the frequency of misogynous words used.