subreddit-analyser

A set of Python notebooks to scrape and analyse subreddit posts.
The Analysis is divided into three major steps:

1. Crawl Sub-Reddit data using the PushshiftAPI.
2. Preprocess the crawled data and perform feature-extraction.
3. Plot the data points and conclude from data.

Preparation of Data-set:

- I used the PushshiftAPI to crawl data.- The API has a get request with different query parameters and return JSON object containing the post information.
- Sample
  URL: https://api.pushshift.io/reddit/submission/search/?after=1577750400&subreddit=emacs&size=100&sort˙type=created˙utc&sort=asc&fields=author,author˙fullname,created˙utc,domain,full˙link,is˙crosspostable,link˙flair˙text,num˙comments,num˙crossposts,over˙18,permalink,score,selftext,title,total˙awards˙received
- After an analysis of different entries that I received from the API. I have used fields author, author˙fullname, created˙utc, domain, full˙link, is˙crosspostable, link˙flair˙text, num˙comments, num˙crossposts, over˙18, permalink,score, selftext, title, total˙awards˙received.

Data-set:

- I crawled and created data-set of subreddit posts from *r/emacs* and *r/vim*. The data-set represents the posts made on these subreddits by users from January 1st 2020 to March 31st 2020.
- The emacs-raw data-set is fairly small with 1353 rows which include posts that are deleted. To make this analysis more comprehensive I have filtered out the deleted posts and then the resultant vim-filtered has 1255 rows, which indicates only 98 posts were deleted.
- The vim-raw data-set is also fairly small with 1136 rows which include posts that are deleted. I filtered out the deleted posts and then the resultant vim-filtered has 1132 rows, which indicates only 4 posts were deleted. It is significantly lower from the number of deleted posts from the *r/emacs* subreddit.

Feature Extraction:

- After careful analysis of the data that is crawled, I identified the most relevant features that can help us understand the pattern and behaviour of users posting in the subreddits:
- One of the important thing to understand about user engagement in any social media to know if their posts have a positive sentiment or negative sentiment.
- I calculated sentiment from the title of the post and the content of the post, and populated the feature table with post˙sentiment and title˙sentiment using the VADER SentimentIntensityAnalyzer.
- VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.
- The other information about social media is captured by reddit˙score which is the (total number of upvotes - total number of downvotes).- I also kept the author of the post along with the total number of comments on the post.
- The date of creation of post is also kept as a feature for us to analyse this data from a time-series perspective.

Top 5 metrics in the feature CSV:

- **date˙created**: The date of creation of the post
- **author**: The author of the post. It is the username of the author
- **post˙sentiment** : The sentiment of post: Positive, Negative or Neutral
- **reddit˙score**: The Reddit score which is the *total number of upvotes - total number of down votes*
- **num˙comments**: The total number of comments on the post

I have also kept post and title columns to uniquely identify the post.

Data Visualization and analysis:

- Both datasets are comparable with about 1000 rows, which indicate roughly 10 posts per day average on both subreddits.
- The subreddit *r/emacs* is having slightly more data, so the number of posts are higher.

Sentiment Analysis:

After sentiment analysis I found that the posts on both subreddits have majority of positive or neutral posts.

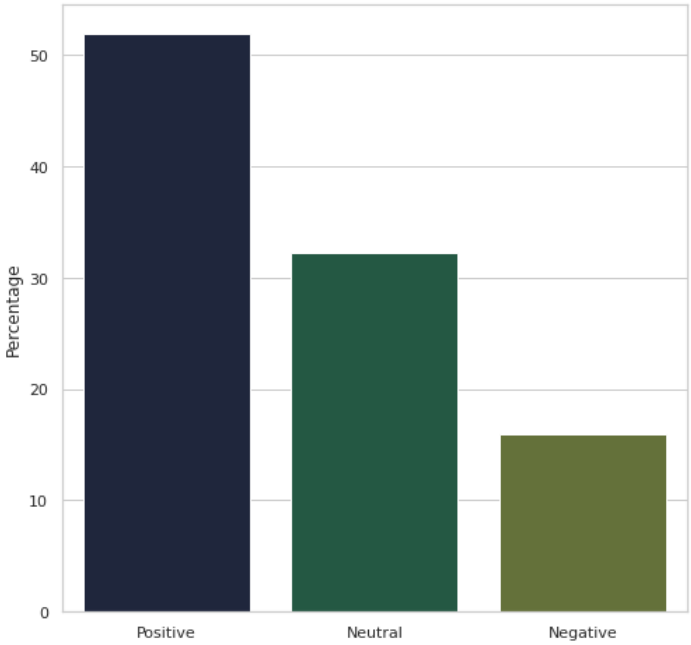

Figure 1 : Sentiment analysis of posts of r/emacs

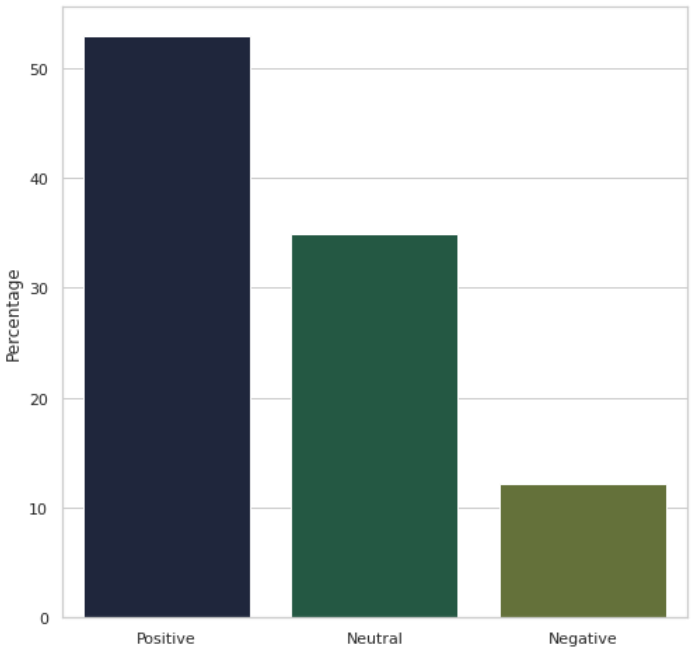

Figure 2 : Sentiment analysis of posts of r/vim

Engagement of users:

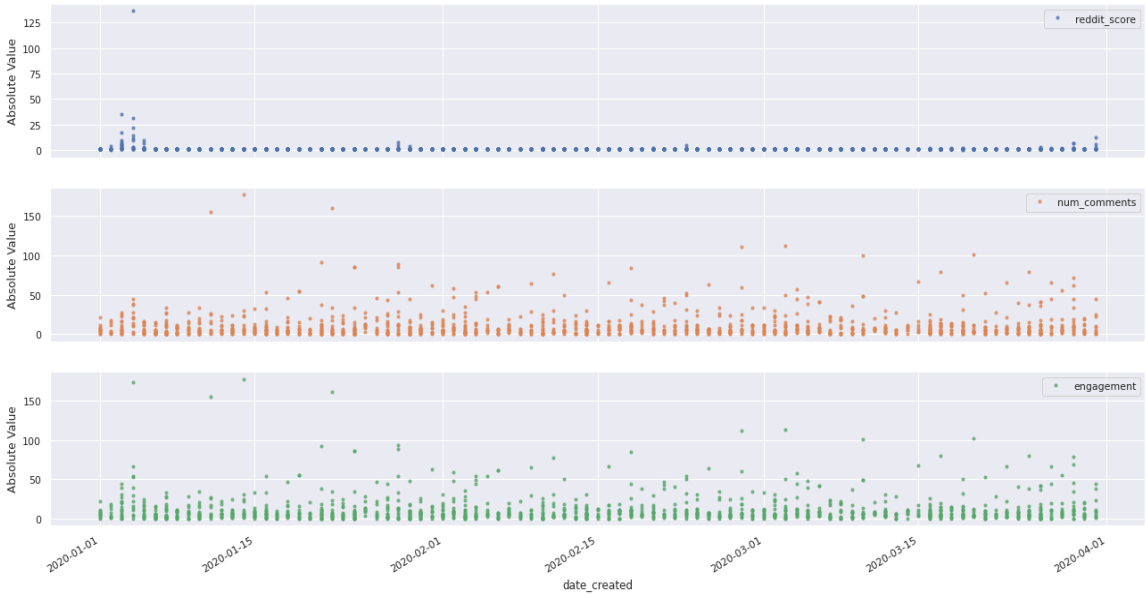

Figure 3 Time-series plot of reddit˙score, num˙comments and engagement for r/emacs

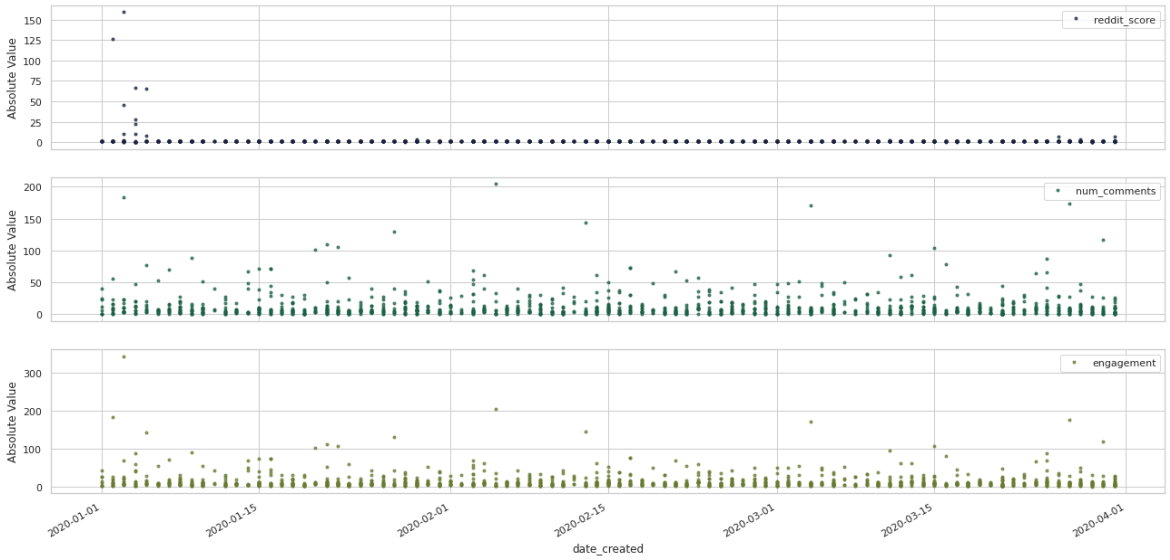

Figure 4 Time-series plot of reddit˙score, num˙comments and engagement for r/vim

In both datasets we see very high reddit scores in month of January, and low scores in February and March. It could be an issue with data from the Pushshift API as similar trend is not observed in the number of comments.

We also don't observe a positive correlation between the reddit scores and number of comments of the posts, which is often observed in other social media like Facebook. Reddit is mostly used for sharing new information in form of memes or asking doubts. This plot validates the reddit use-cases.

We see more comments in the posts of *r/vim* which indicates that users are more active in that subreddit.

We can notice spikes of reddit scores and comments which follows a general rhythm, it indicates that users get more active around certain days of week on both subreddits.

Top Contributers:

I have calculated the top contributor by the total number of posts shared by them.

They can also be ranked by how engaging their posts are, how many positive negative posts etc. That is future work for this project.

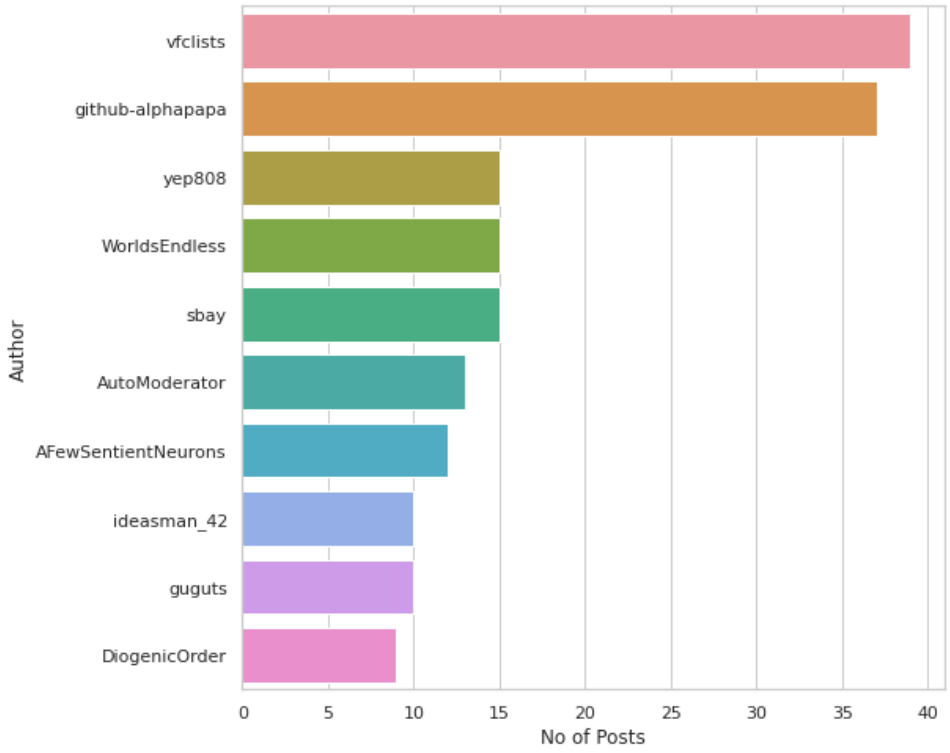

Figure 5 Top contributers on r/emacs

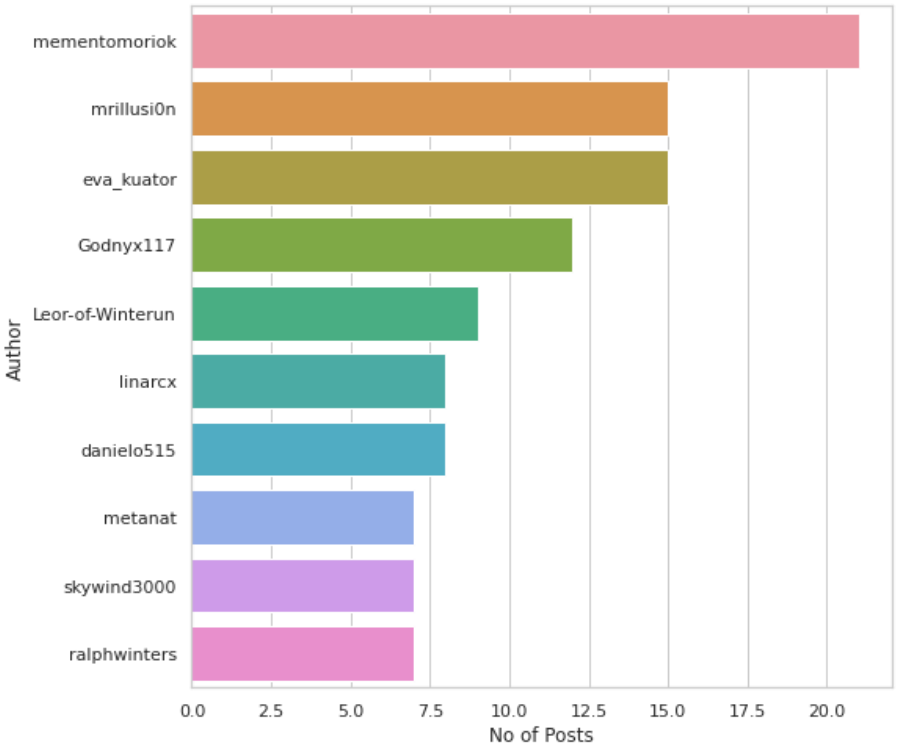

Figure 6 Top Contributors on r/vim

In both subreddits we a found a lot of authors who have posted more than 5 posts during the last three months.

Interestingly two authors are standing out for r/emacs while r/vim has more people writing posts.

Top keywords:

Interestingly the top keyword for r/emacs was emacs and for r/vim was vim.

It is also a positive control on our validation as it was an expected result.

Words like org, package, suse, mode are more common for r/emacs.

Words like gt, lt, plugin, line are more used in r/vim.

Figure 7 Top keywords for r/emacs

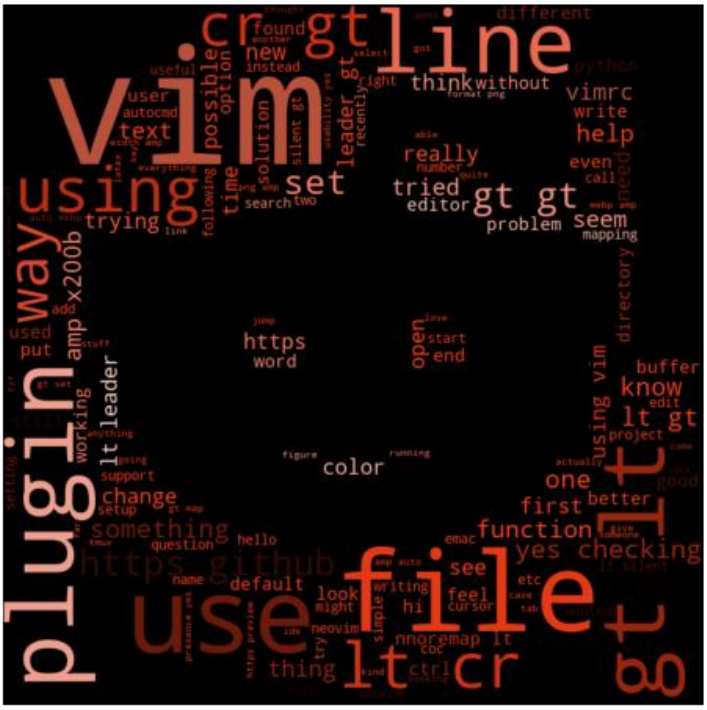

Figure 8 Top keywords for r/vim

Conlcusion and Future work:

- The number of posts on both subreddits are very similar averaging to 10+ posts per day.
- *r/vim* has more engagement of users in terms of number of comments, more number of top contributors.
- *r/emacs* has more number of reddit scores in general.
- *Both subreddits have 0 over'18 posts.*
- *r/vim has less number of deleted posts compared to r/emcas over the time period Jan-20 to Mar-20.*

- I plan to do more analysis, and draw more plots to get more conclusions.
- I want to convert these Py notebooks to a rest API which can accept some information related to subreddit name, date range, perform analysis and generate plots for users to infer meaning from this data.