CS685 Course Project
Stack-Exchange Miner
Group 26

1 Group members

1. Aaryan Srivastava (180007, aaryans@iitk.ac.in, /aaryans941)
2. Ashwin Shenai (180156, ashwins@iitk.ac.in, /ashwin2802)
3. Utkarsh Gupta (180836, utkarshg@iitk.ac.in, /utkarshg99)
4. Varun Goyal (180850, govarun@iitk.ac.in, /govarun)

2 Project Introduction

The Stack Exchange network comprises of over 173 Q&A communities, including popular ones like Stack Overflow, the largest and most trusted developers to learn and share their knowledge. It is now one of the most popular forums on the internet, with over 3.2 Million questions and over 430 Million monthly user visits.

The project aims to analyse data and study interesting insights of various stack exchange forums, ranging from hinduism to data science, using their data dumps. Since the schema for various stack exchange data dumps is similar, the goal is to develop generic data analysis irrespective of the specific data we examine.

3 Running the project

The details are listed in a separate file, named ’README.md’ within the project repository. Please refer to the same, details omitted to keep the report as short as possible.

4 Data Description

Data available was for up until 6th September 2021. The data is downloaded using the following link.
Data format: 7zipped
Schema can be found within the repository in a file named ‘schema.md‘ and is also publicly available here.
Omitted complete details here to keep the report short.

5 Data Extraction

- We have automated the entire process (downloading, extracting and pre-processing) via JavaScript, the code is present in the file ‘index.js‘
- We first verify if all the files required for analysis are available for the given Stack Exchange. If they are available, we download all the 8 required files, namely, ‘Badges.xml, Comments.xml, PostHistory.xml, PostLinks.xml, Posts.xml, Tags.xml, Users.xml.‘ in form of a .7z archive.

- This archive was then unzipped.
- We then process the extracted files, as described subsequently.

5.1 Challenges Encountered

- The data source for scraping had some inconsistencies.
  - For example, some catalogue files were not listed, whereas they were actually available for scraping.
- To ensure no loss of data, we check all the files by calling API requests separately for all the files. We took note of all the metadata for the file then.
- Now, the developer could easily choose which database to work on, by looking at the meta data of the database.

6 Data Preprocessing

- The data was available in `.xml` format. We converted the data to to make it more intuitive.
- Also for advanced analysis, `json` files are supported by many python libraries.

7 Project Structure

```
StackExchange-Miner/
├── Scraping/
│   ├── main.tsv
│   ├── xmls-v.csv
│   ├── xmls.csv
│   └── xmls.json
├── Scripts/
├── Data/
│   ├── Extracted
│   │   └── hinduism.stackexchange.com/
│   │       ├── Badges.json
│   │       ├── Comments.json
│   │       ├── PostHistory.json
│   │       ├── PostLinks.json
│   │       ├── Posts.json
│   │       ├── Tags.json
│   │       ├── Users.json
│   │       └── Votes.json
│   └── hinduism.stackexchange.com.7z
├── Results/
├── README.md
├── Report.md
├── index.js
├── requirements.txt
├── schema.md
├── sites.xml
└── start.sh
```

```
StackExchange-Miner/
└── Scripts/
    ├── active_users.py
    ├── association_rule.py
    ├── badges.py
    ├── comments.py
    ├── fastestgun.py
    ├── main.py
    ├── main_.py
    ├── map_reduce.py
    ├── posthist.py
    ├── postlinks.py
    ├── posts.py
    ├── question_time.py
    ├── tag_pred.ipynb
    ├── tag_pred.pkl
    ├── tag_prediction.py
    ├── tags.py
    ├── users.py
    ├── utils.py
    ├── votes.py
    ├── voting_reputation.py
    └── word_cloud.py
```

```
Results/
└── hinduism.stackexchange.com/
    ├── Badges/
    │   └── badges.results.json
    ├── Comments/
    │   └── comments.results.json
    ├── Fastestgun/
    │   └── fastestgun.json
    ├── PostHistory/
    │   └── posthist.results.json
    ├── PostLinks/
    │   ├── postlinks.results.json
    │   ├── postrel.graph.json
    │   └── static_graph.html
    ├── Posts/
    │   ├── post_graph.html
    │   ├── posts.json
    │   ├── posts.users.json
    │   └── user_graph.html
    └── Users/
        ├── profiles.results.json
        └── users.results.json
```

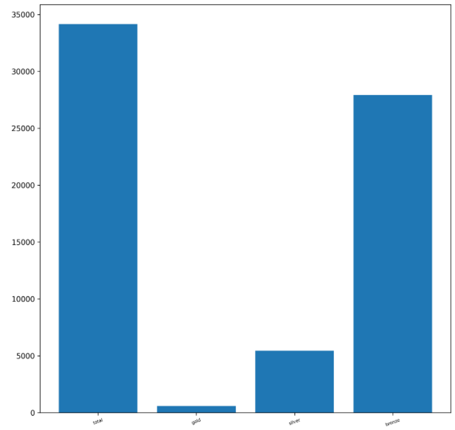
```
Results/
└── hinduism.stackexchange.com/
    ├── ARM_badges_fits.csv
    ├── ARM_badges_mined.csv
    ├── ARM_tags_fits.csv
    ├── ARM_tags_mined.csv
    ├── MapReduce_AboutMe_Users.json
    ├── MapReduce_Body_Posts.json
    ├── MapReduce_Title_Posts.json
    ├── WordCloud_Posts_Body.png
    ├── WordCloud_Posts_Title.png
    ├── WordCloud_Users_AboutMe.png
    ├── active-users.png
    ├── active_users.json
    ├── question-time.png
    ├── question_time.json
    ├── voting-reputation.png
    ├── Tags/
    │   ├── counts.csv
    │   └── tags.results.json
    └── Votes/
        ├── special.posts.json
        └── votes.results.json
```

8 Results and Inferences

To keep the report short and sweet, we provide results and inferences from only one Stack Exchange forums as an example of depicting potential analysis from its data. For this report, we look at the data from the Hinduism stackexchange forum. We have also provided results for the Hinduism, Ethereum, Crypto, Datascience, Space and Islam stackexchange forums in separate .pdf files along with the project to highlight our diverse and generic analysis methods. Please find these in the `Results/` folder.
It is worth noting that this report explains various observations, insights and results as part of any single Stack Exchange forum and displays representative plots only, leaving out repetitive tabular results for the sake of brevity. The complete results can be found in their respective files submitted for various stackexchange forums as mentioned above.

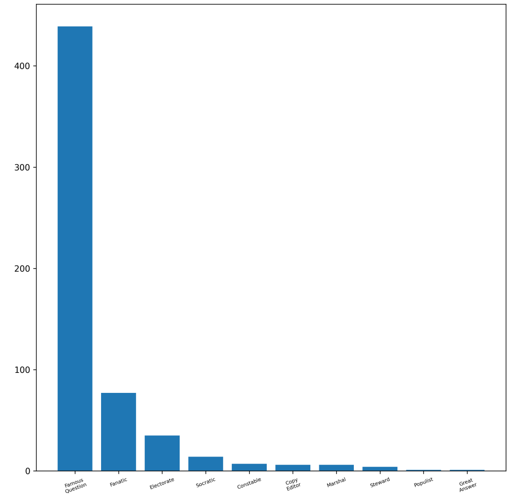
8.1 Badges

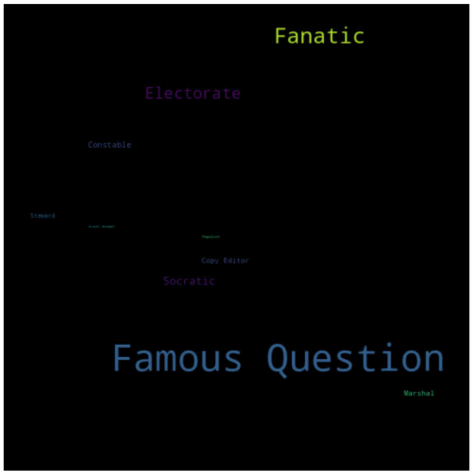
Stack Exchange forums provide badges members for being especially helpful through their questions and answers and even awarded to special posts and answers. Badges come in three tiers; gold, silver and bronze. We look
While our analysis across different stack exchanges we also found out that the top badges are quite similar across them. We first plot the number of total, gold, silver and bronze badges awarded to community users. It is worth noting that each user can get multiple badges, too. It can be observed from the plot that gold badges aren't easy to get, whereas bronze are quite common.



8.1.1 Insights on Gold badges

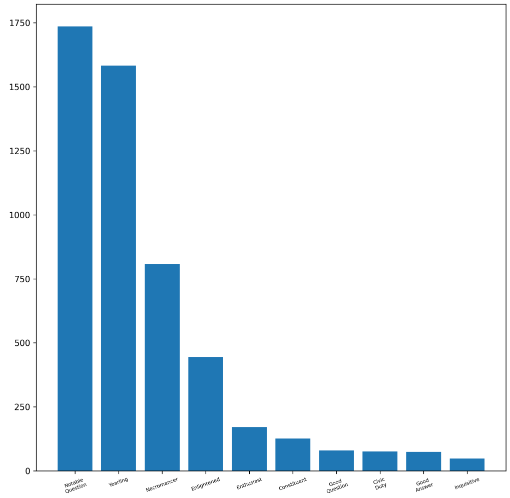
We plot the top 10 most awarded gold badges and depict the most popular ones in a word cloud representation. We notice that 'Famous Question' is the most popular badge (by a huge margin) which is given out to a question with over 10,000 views. It is also pretty evident from the word cloud and frequencies that gold badges are pretty rare.

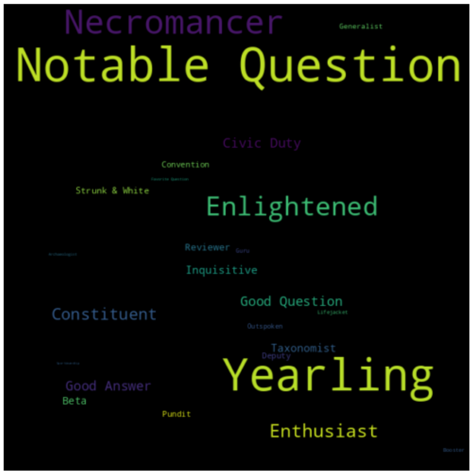




8.1.2 Insights on Silver badges

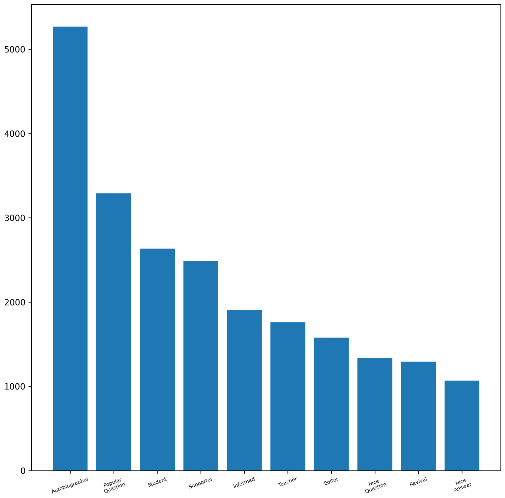
We plot the top 10 most awarded gold badges and depict the most popular ones in a word cloud representation. We notice that 'Notable Question' is the most popular badge which is given out to a question with over 2,500 views. Though it is closely followed by the 'Yearling' badge which is awarded to users active for more than a year, with at least 200 reputation. It is also pretty evident from the word cloud and frequencies that silver badges are not as uncommon as gold badges.

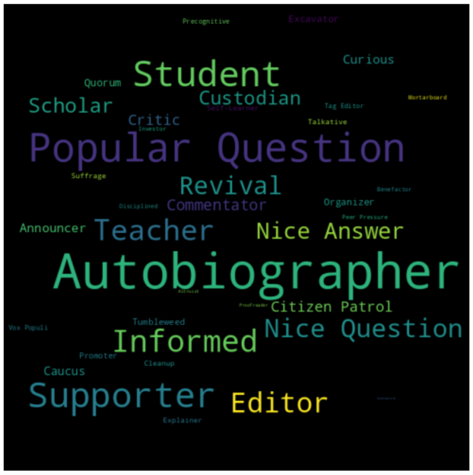




8.1.3 Insights on Bronze badges

We plot the top 10 most awarded gold badges and depict the most popular ones in a word cloud representation. We notice that 'Autobiographer' is the most popular badge (by a decent margin) which is given out to any user with a complete 'About Me' section in their profile. It quite evident from the top badge, word cloud and frequencies that bronze badges are very easy to obtain and thus are quite common.



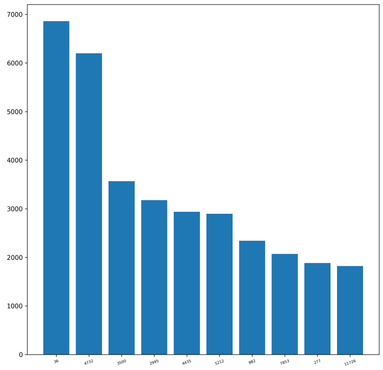


8.2 Comments

Comments are an essential part of the stackexchange community, allowing users to put notes under questions asking for clarifications, without explicitly answering them and under answers, requesting updates to tailor them as per need. We observe that the ratio of max to average values for 'Comments per post' is around 17 (Max: 73, Avg: 4.3) whereas that of 'Comments by a user' stands at about 37.5 (Max: 6860, Avg: 37.5). Expectantly, these ratios imply skewed distribution of comments in both scenarios.

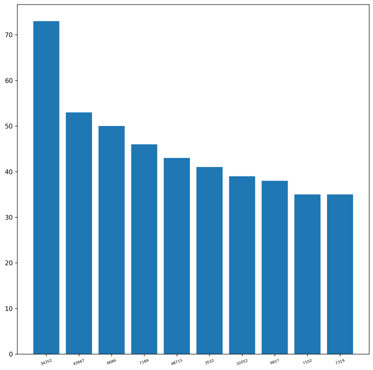
8.2.1 Top users by number of comments

We see that there are a few users who are particularly active and make a lot of comments.



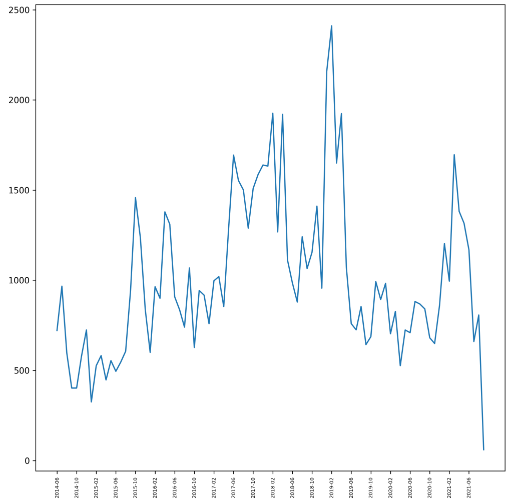
8.2.2 Top post by number of comments

There are long discussions in the comments section or people are just asking for more clarification.



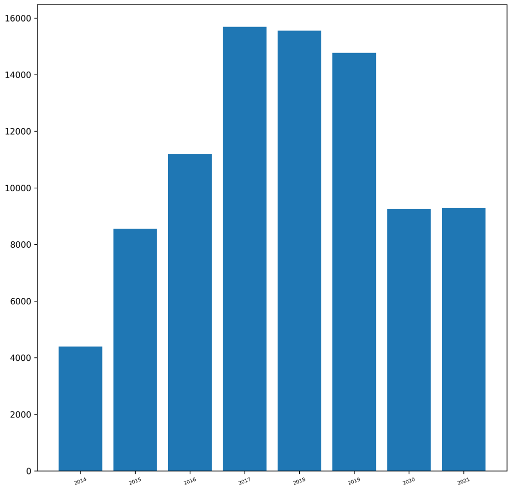
8.2.3 Number of comments by month

There is no particular trend in the comments observed by month. We observe that in Feb '18, and Feb '21 the number of comments peak, but in Feb '20 there is a dip. Hence, no conclusion can be made.



8.2.4 Number of comments by year

We see that rise in number of comments is steady till 2017, that means the users became more active. But, we can see a steady trend after that.

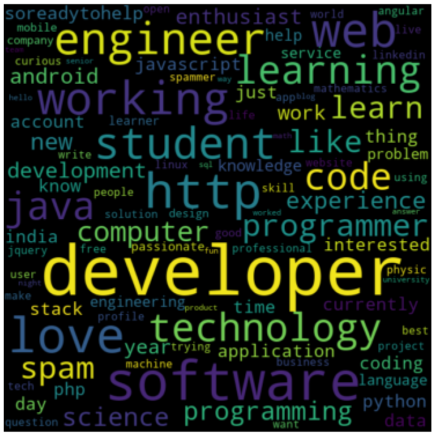


8.3 Users

Users are the key aspect of the stack exchange community ideology. Users can gain reputation by getting up-voted by other users based on their helpful posts, answers and comments.
Users have a profile, which consists of various sections like an 'About me' section describing themselves in short. We mined the text in these sections and analysed its data.

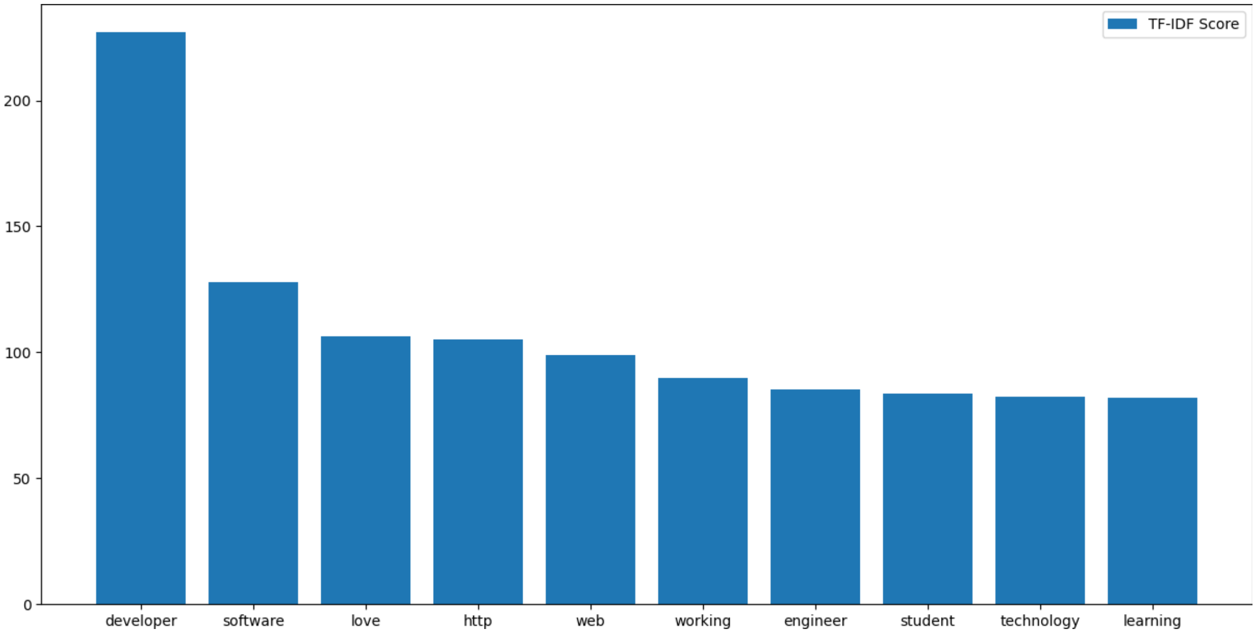
8.3.1 Word Cloud of Users' About me

We mined the text from the 'About me' section of various users and analysed it. Surprisingly users on a forum dedicated to Hinduism, the word cloud and frequencies indicate popular words like 'developer', 'engineer', 'technology'. This can be owed to the fact that stack exchange users are mainly developers and user accounts are common across different stackexchanges as developers are more versed with stackexchange and its functionalities compared to say, an average literature person.



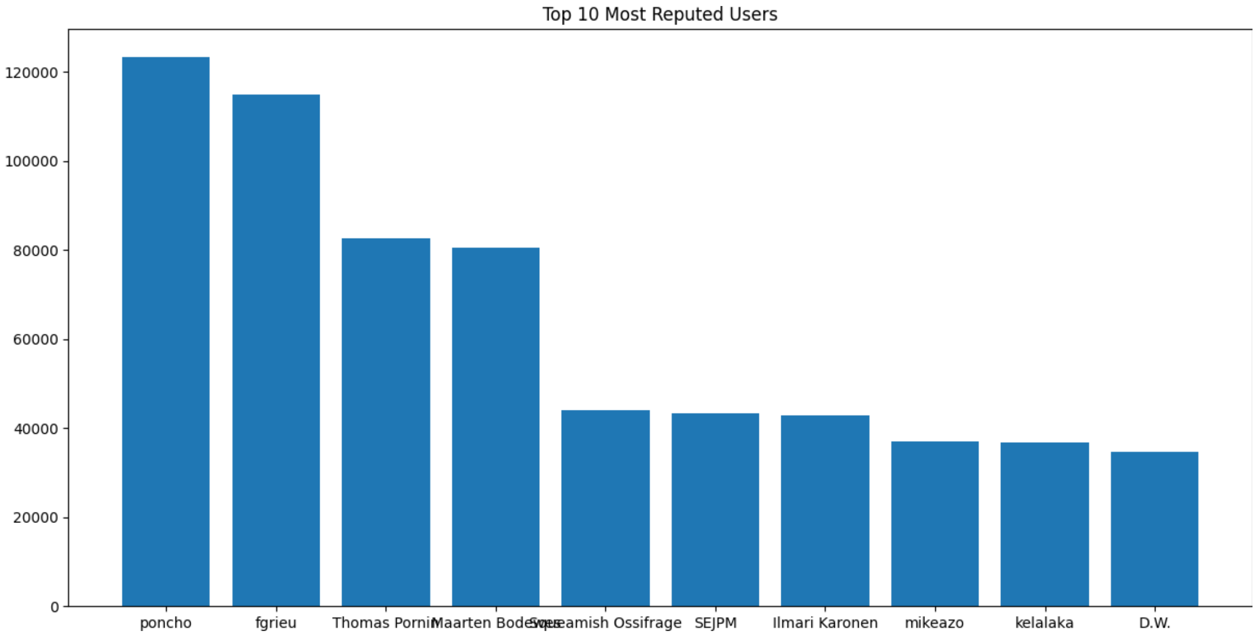
8.3.2 Map Reduce of Users' About me

For calculating the MapReduce of Users' 'About me' section we first tokenized the posts' titles, and then we lemmatized the tokens. Then the TF-IDF score was calculated for every term in every document (here, the 'About me' section). To rank the terms we take the sum of TF-IDFs across all the documents.



8.3.3 Top 10 users by reputation

"Total Number of Users": 18481, "Total Reputation Points": 1786791, "Average User Reputation": 96.68 We see that the max reputation of a user is around 90,000 which is a lot more than the average user reputation. This behaviour is expected as an average user just browses the question he needs the answer to and does not worry about answering other questions in the community.

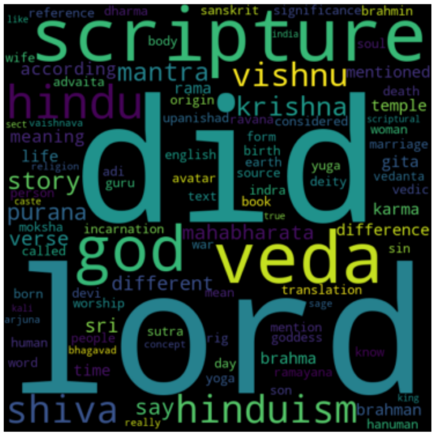


8.4 Posts

Stack Exchange questions allow an answer to be attributed as an 'Accepted answer'. This is the answer which answers the requirements stated in post in a satisfying manner. 90 percent of the questions do not have accepted answers, i.e. they are open.

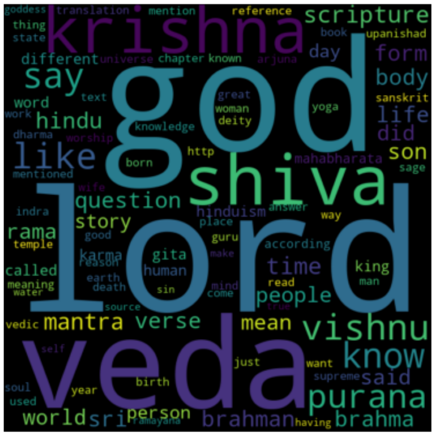
8.4.1 Word Cloud of Posts' Title

The word cloud is one of the most attractive analysis tools. Here, it shows lord, scripture and veda to be one of the most popular words in the titles.



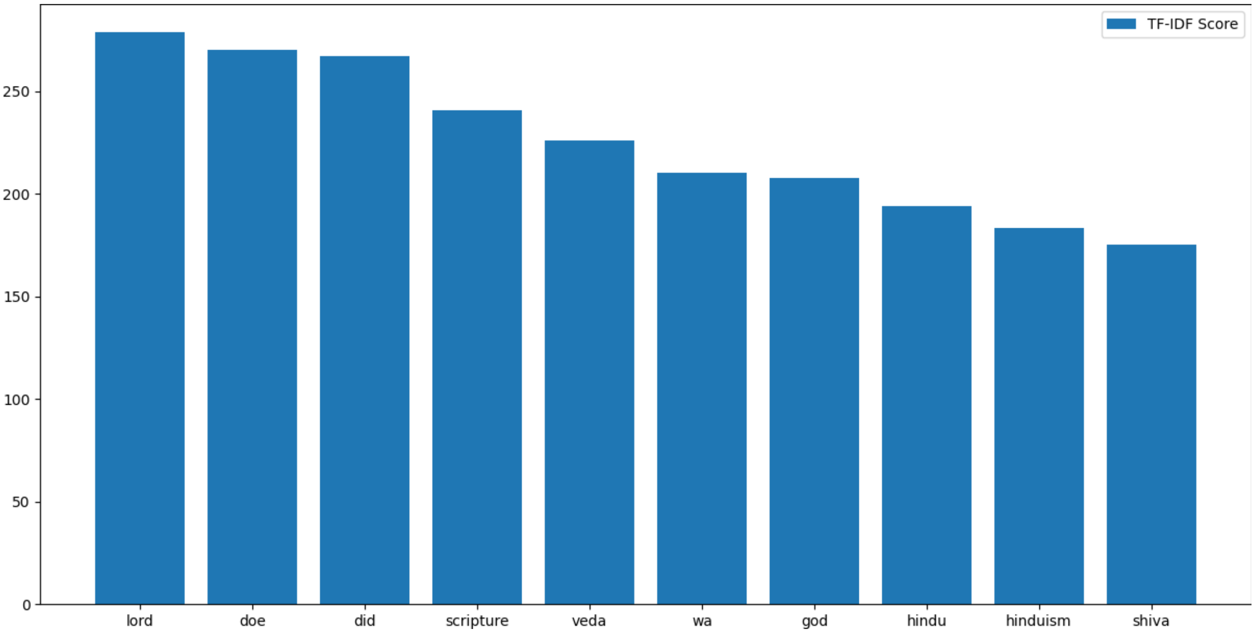
8.4.2 Word Cloud of Posts' Body

In this word cloud we see that 'lord', 'god', 'veda' and 'shiva' are very common. Since, this is hinduism stackexchange, this trend is expected.



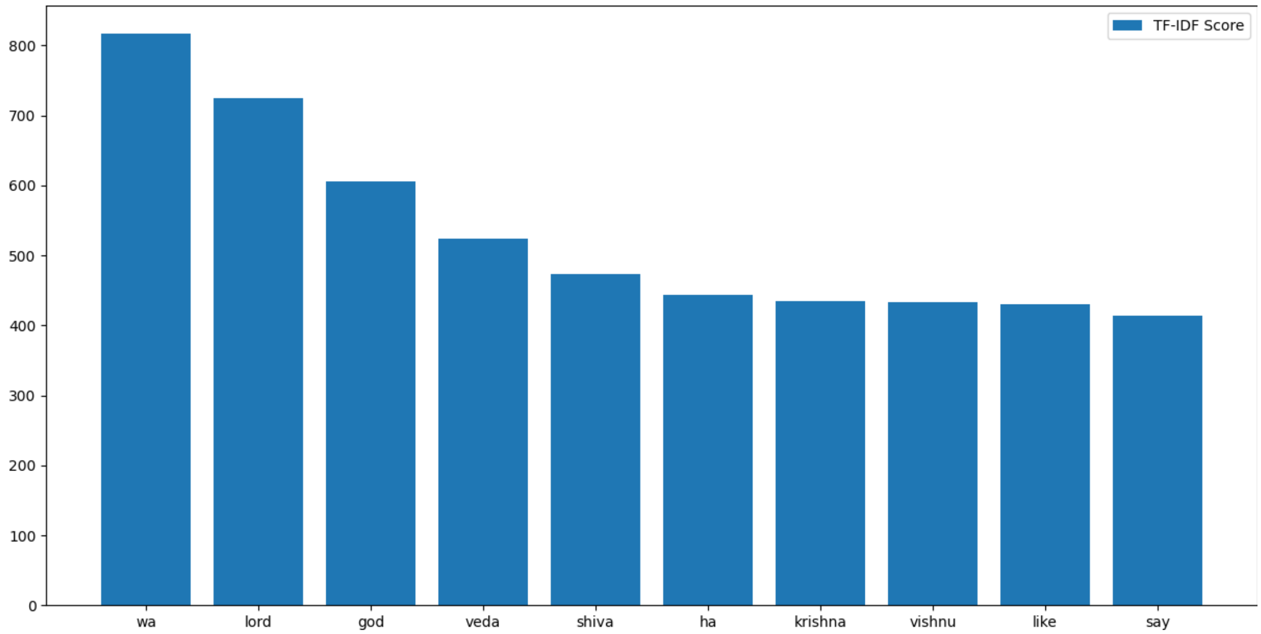
8.4.3 MapReduce of Posts' Title

The MapReduce is calculated as mentioned for the Users' About me section, but with the document being the Post title here.



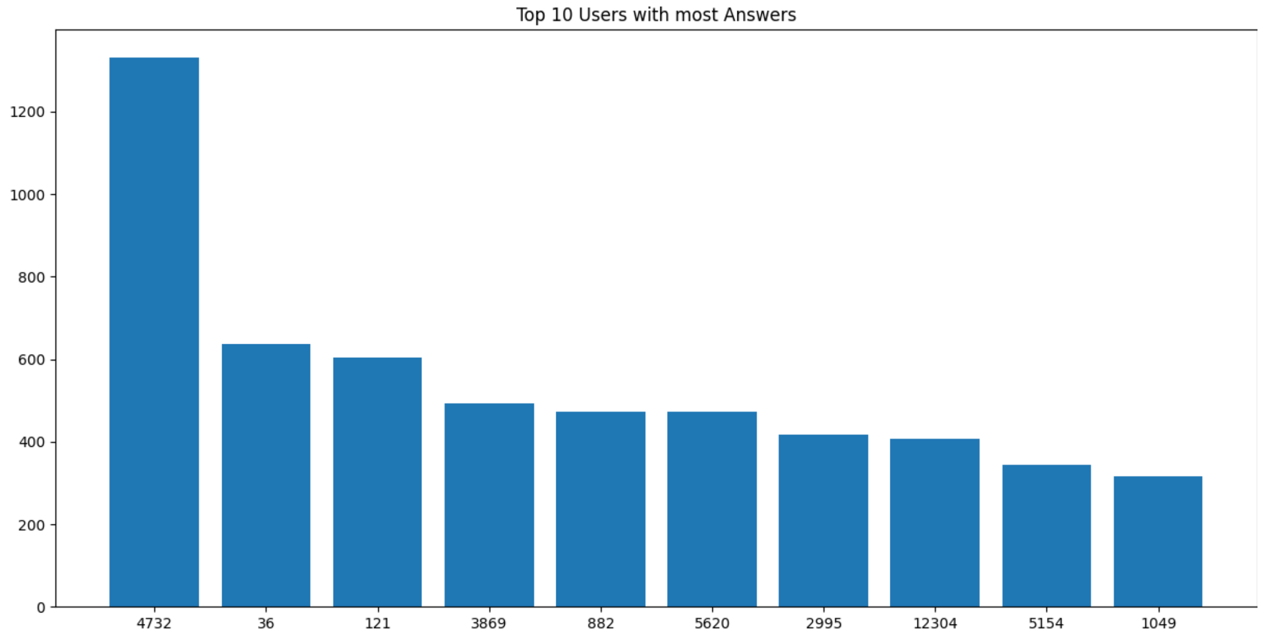
8.4.4 MapReduce of Posts' Body

The MapReduce is calculated as mentioned for the Users' About me section, but with the document being the Post body here.



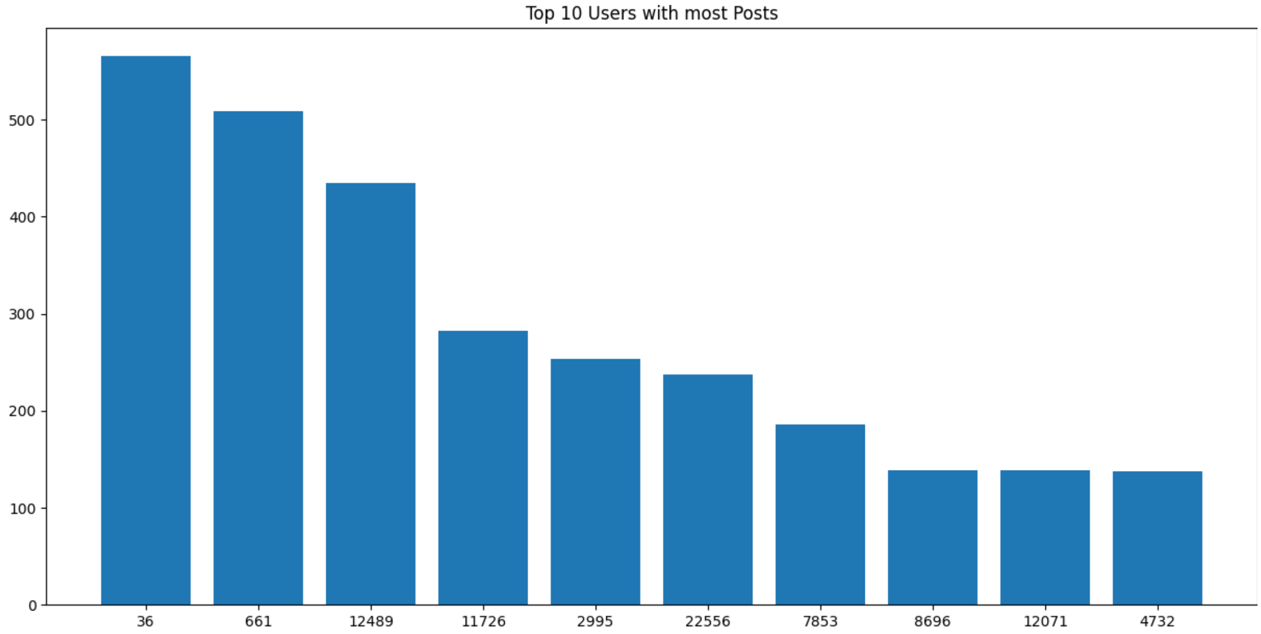
8.4.5 Top 10 users by number of answers

We see that there is 1 user which is very active (in terms of answers authored) and after that the trend becomes rather continuous.



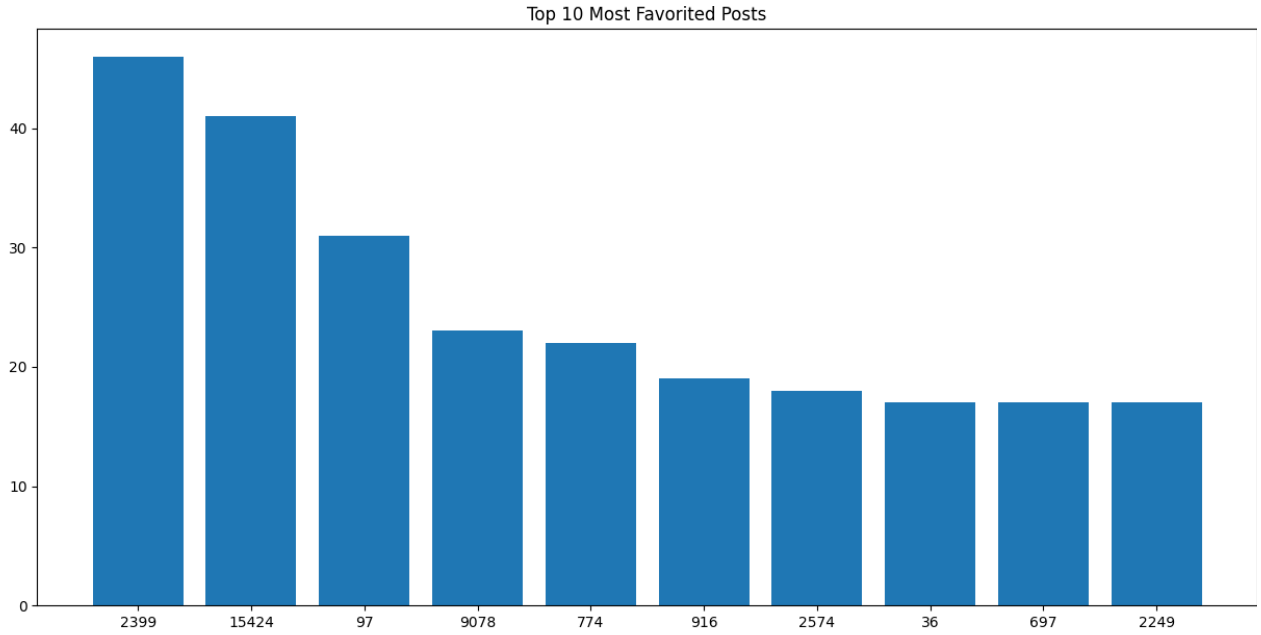
8.4.6 Top 10 users by number of posts

We see that there are 3-4 users which are very active (in terms of posts created) and after that the trend becomes rather continuous.



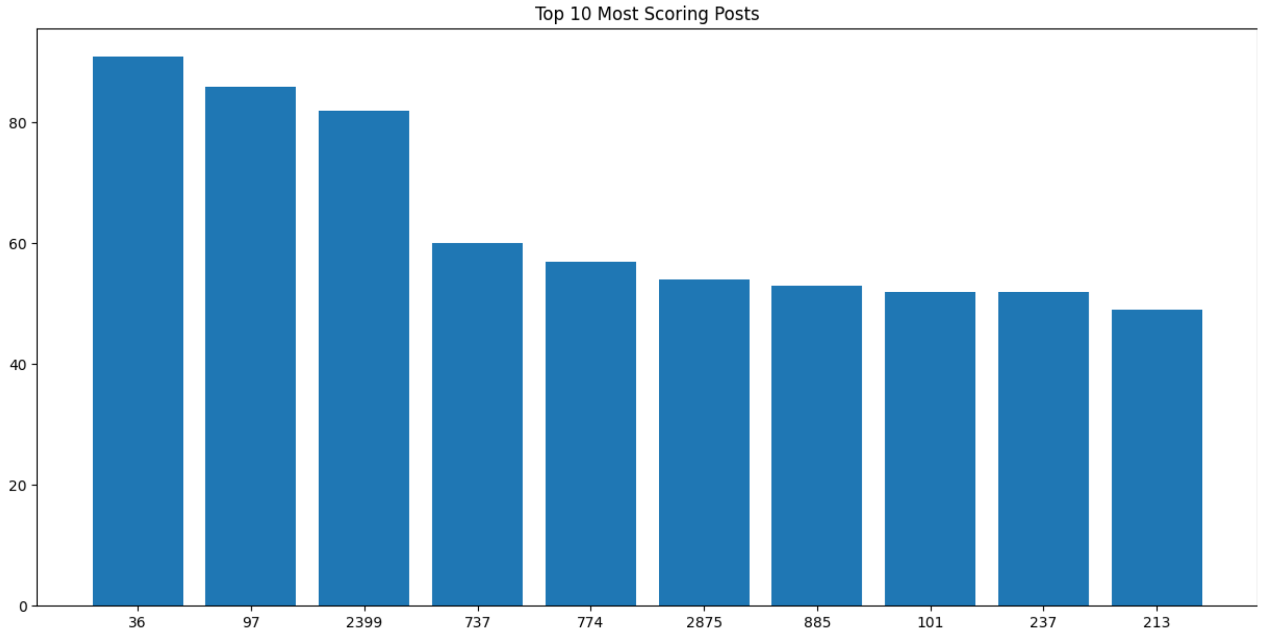
8.4.7 Top 10 most favorited posts

The most favorited post is titled ”Is our destiny predetermined? If yes, then why do our actions affect our karma?”.



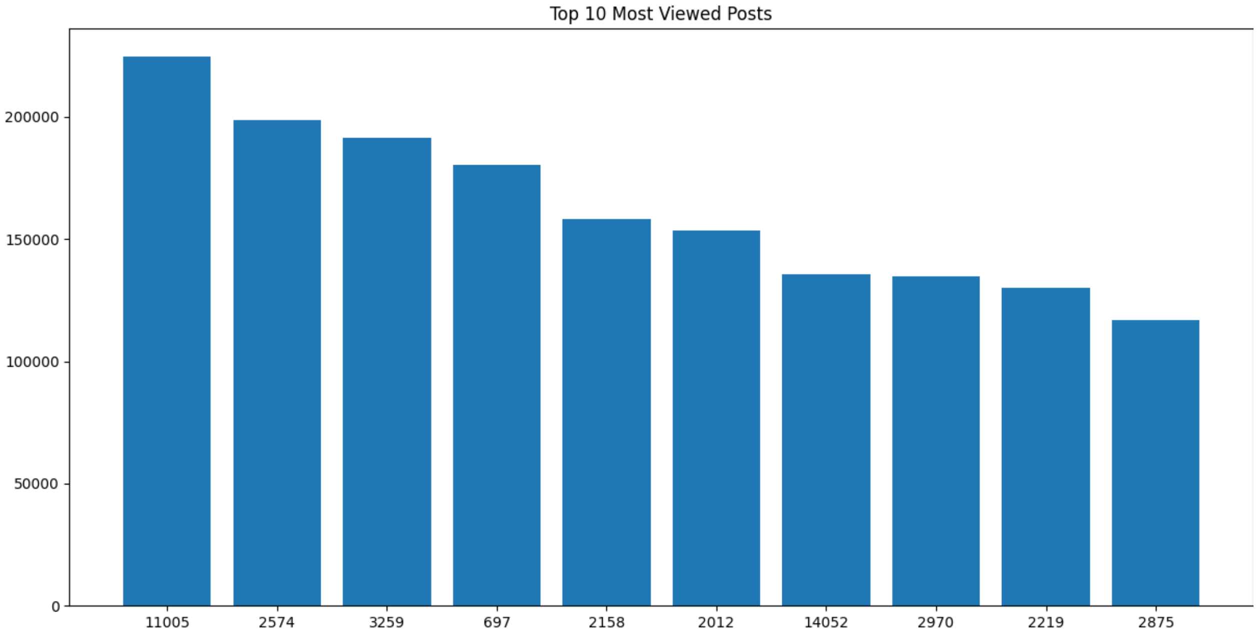
8.4.8 Top 10 most scoring posts

Score is a metric calculated by the subtracting number of downvotes from total upvotes for a given post. So, the most scoring ones are the which where the upvotes clearly outnumber the downvotes. The highest scoring post in this case is titled Why do Hindus believe in cremation instead of burial?



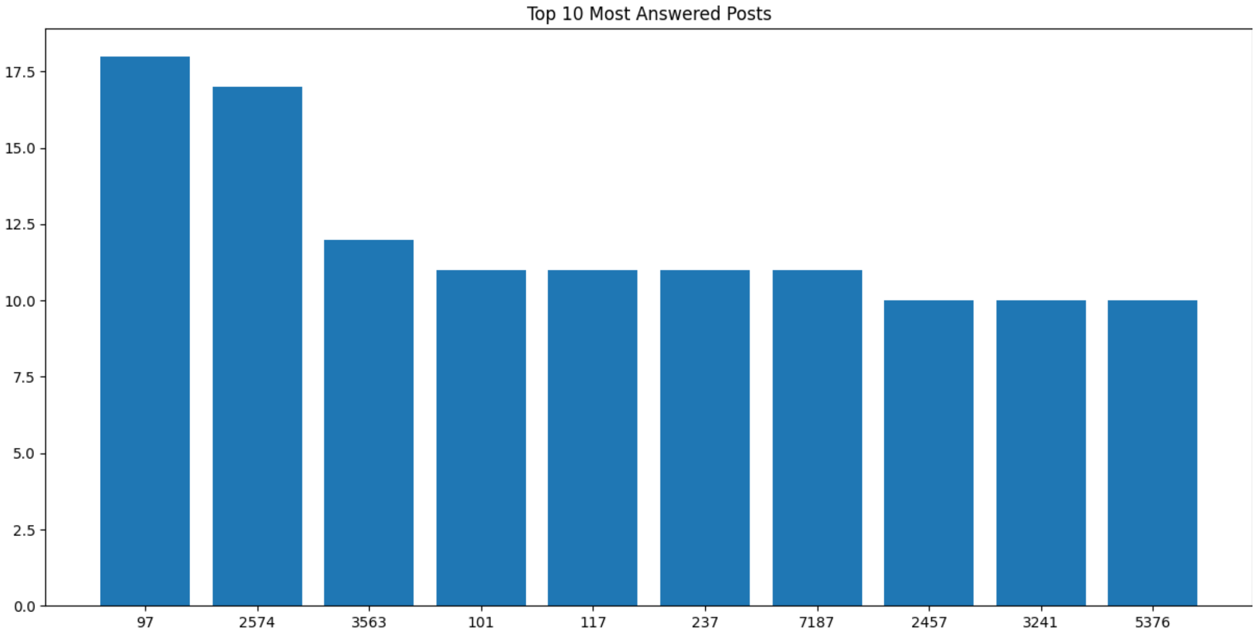
8.4.9 Top 10 most viewed posts

The most viewed posts are the ones that answer the most common questions. Hence, a lot of people have visited these posts. The most viewed post is titled ”Is there any significace of twitching of eyes?” (Note: The question title still has a grammatical error as seen).



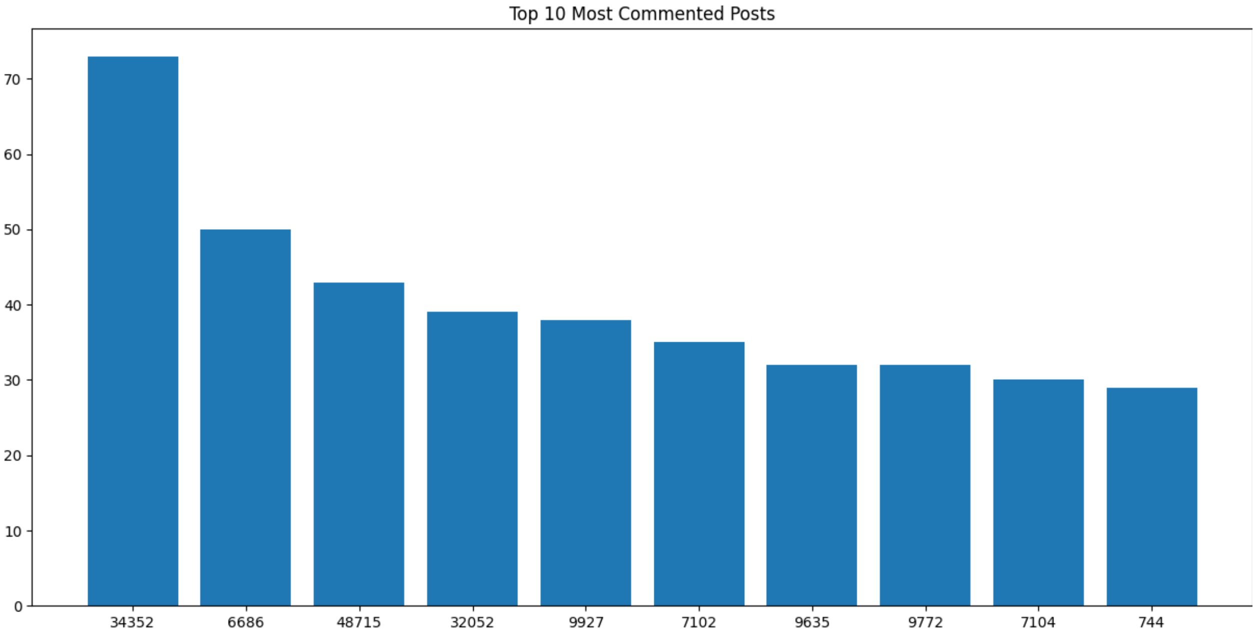
8.4.10 Top 10 most answered posts

The most answered post is titled "Why can we eat 'living' plants but not 'living' animals?".



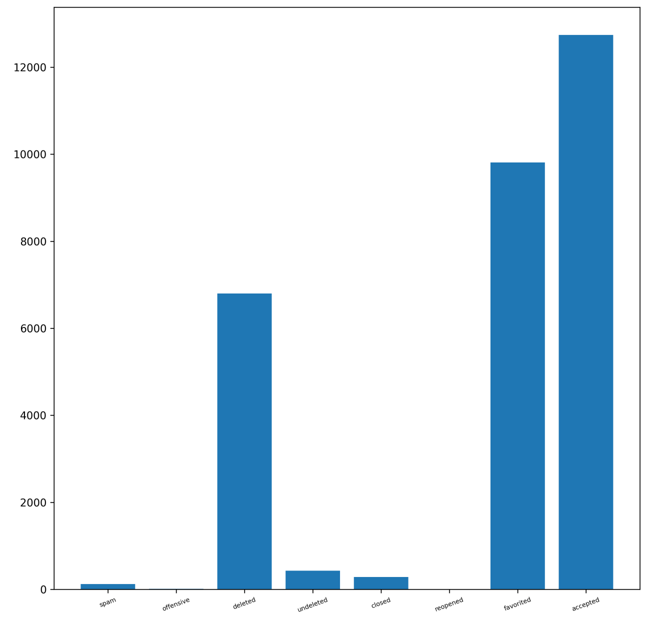
8.4.11 Top 10 most commented posts

The most commented post is titled "Is science any different from religion? Especially when it comes to falsifiability of claims?".



8.4.12 Special types of posts

These posts are marked as spam or offensive by the users. False information, and cyber harassment are one of the major issues in today's digital world. And this analysis just highlights that.

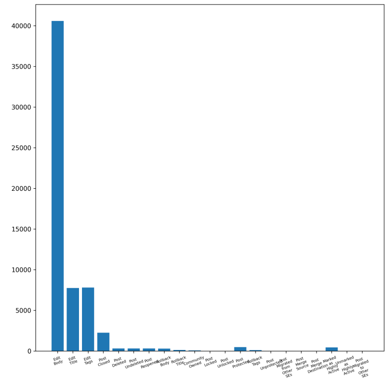


8.5 Post History

Post history includes the changes that are made to a post after it is created. Post history gives us details like what was the initial title, body and then what was the updated title and body. The possible reasons why a post closes can be since it was marked one of Duplicate, Unclear, Off-topic, Too broad, Subjective, Pointless, Localized, etc. among which

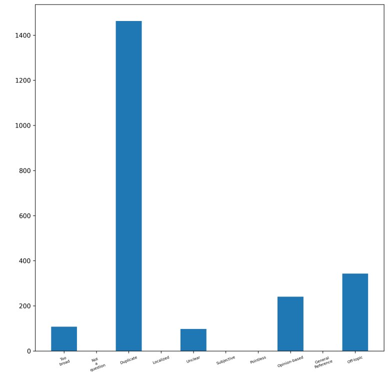
8.5.1 Event category type

We infer from the plot below that only Posts' body, title and tags are changed. Other things are generally not edited.



8.5.2 Post close category type

The most common close reason is "Duplicate", and "Off-topic" or "Opinion based" follow it, as expected from a forum voicing questions and discussion about religion.



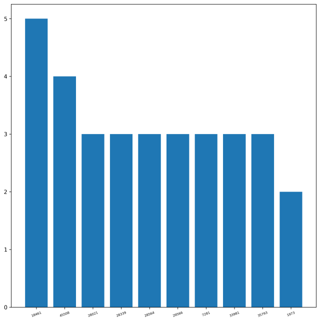
8.6 Post Links

Post links describe the relation between two posts. Like if one post is duplicate of another or two posts are related to each other. We have also created a graph and hosted it at the link: static_graph.html

(P.S. a similar graph is hosted for each of 6 aforemention analysed Stack Exchanges analysed at the following url, replacing `stackexchange-name` appropriately:
`https://stackexchange-miner.web.app/stackexchange-name /static_graph.html`)

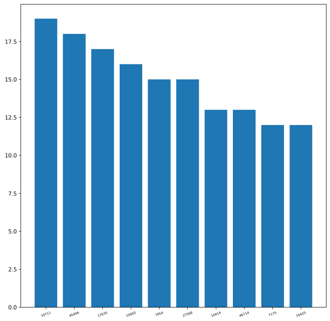
8.6.1 Top 10 post which have the most duplicates

The top 10 posts which have the most number of duplicates. This is a common problem, that people do not look at the existing database and ask questions blindly, that leads to repeating of questions.



8.6.2 Top 10 posts which are related to other posts

One post is related to many posts generally when that post is large. Because it contains a lot of matter and has a lot of tags associated with it, hence this is the case.

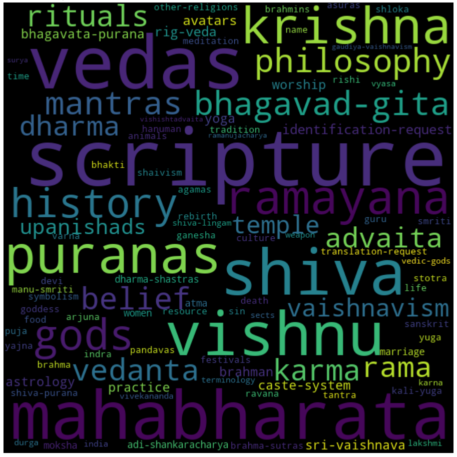


8.7 Tags

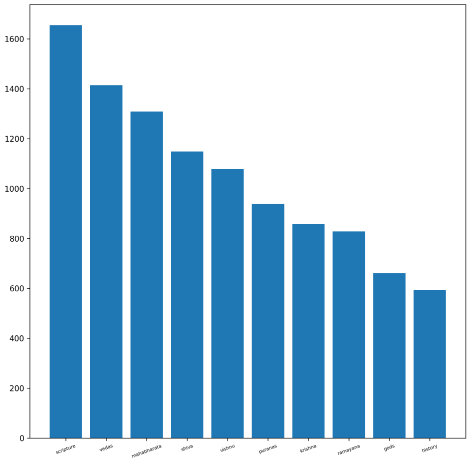
Tags are one-two word attributes of a post. They are used for many purposes like to search for a question, or to give user the idea if he is on the correct question or not.

8.7.1 WordCloud of top tags

Expectedly, the most common tags are about scriptures, the Vedas and the Mahabharata.

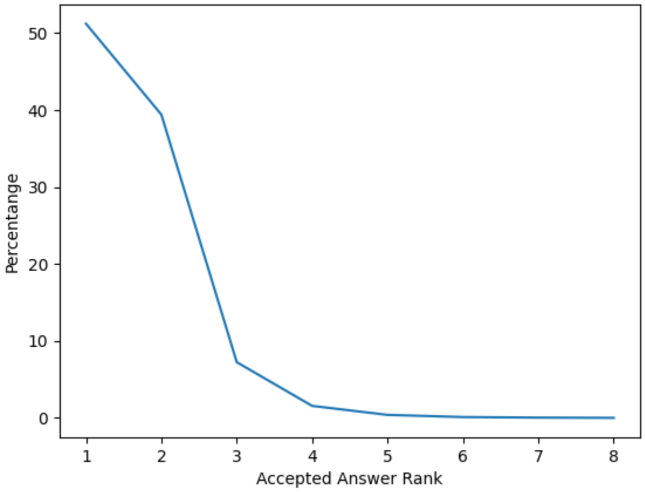


8.7.2 Top 10 tags



8.8 Fastest Gun In The West

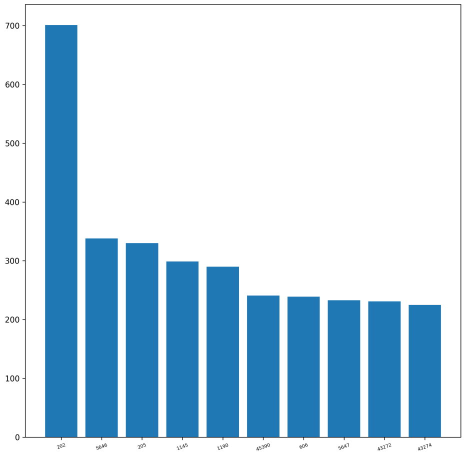
It is noticed that most accepted answers are generally the earlier answer for the respective questions. We observe the same the following plot of the percentage of accepted answers versus their rank (by time). It implies that indeed, earlier answers are more probable to be accepted.
For instance, the first answer is almost 50% probable to be accepted.



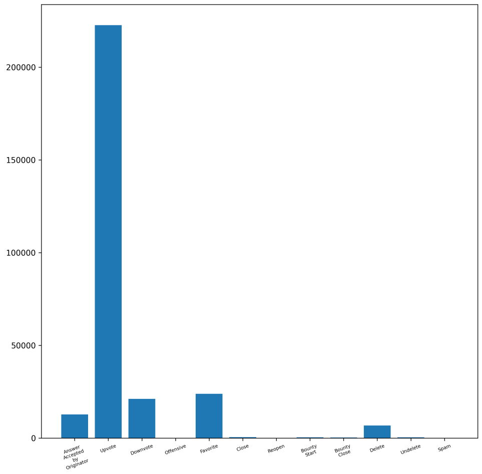
8.9 Votes

Stack Exchange depends on votes the most for credibility of the users' answers. There is no one who can monitor such an open knowledge-database, and there should be no one, other than the users themselves. The votes are exactly a measure of that.

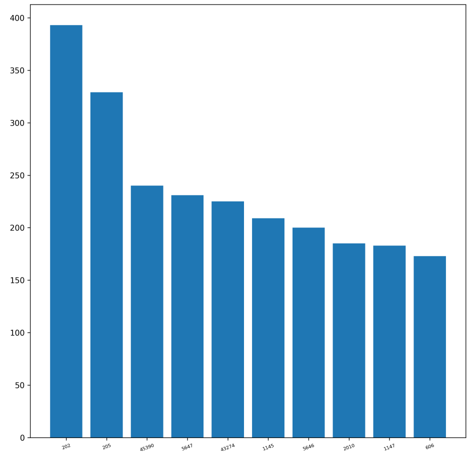
8.9.1 Questions with the max number of votes



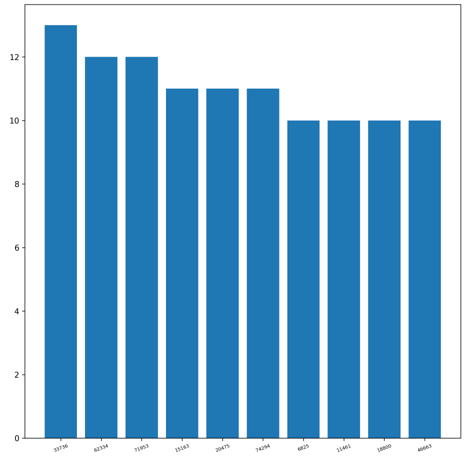
8.9.2 Votes numbered from different categories



8.9.3 Most upvoted questions

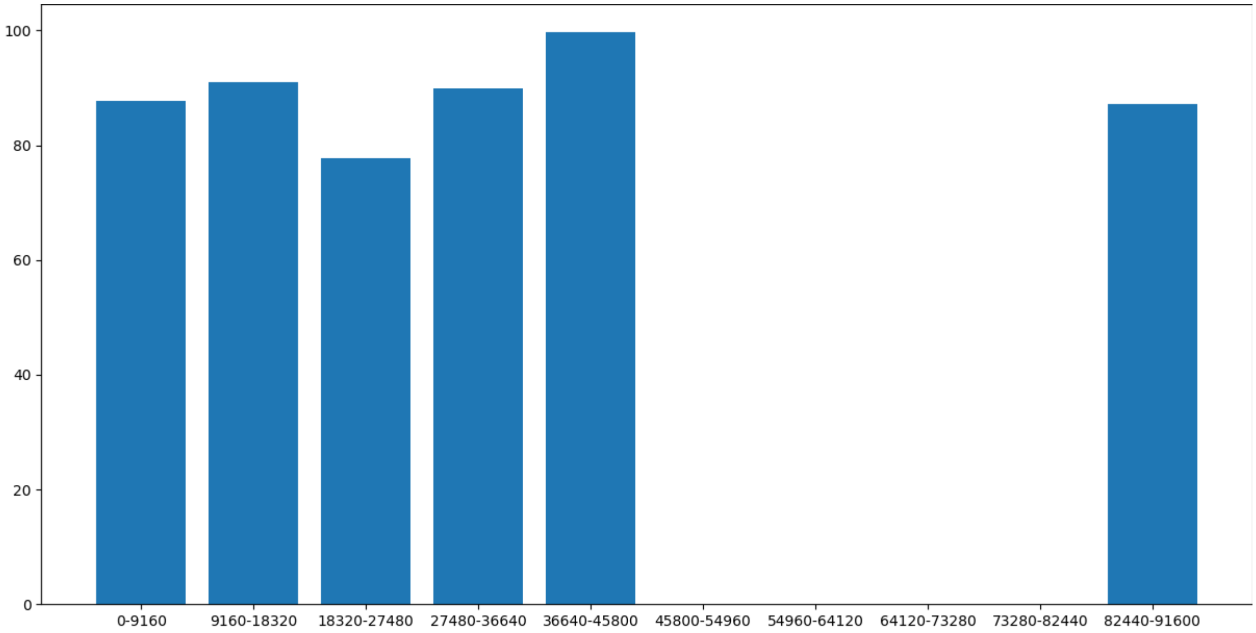


8.9.4 Most downvoted questions



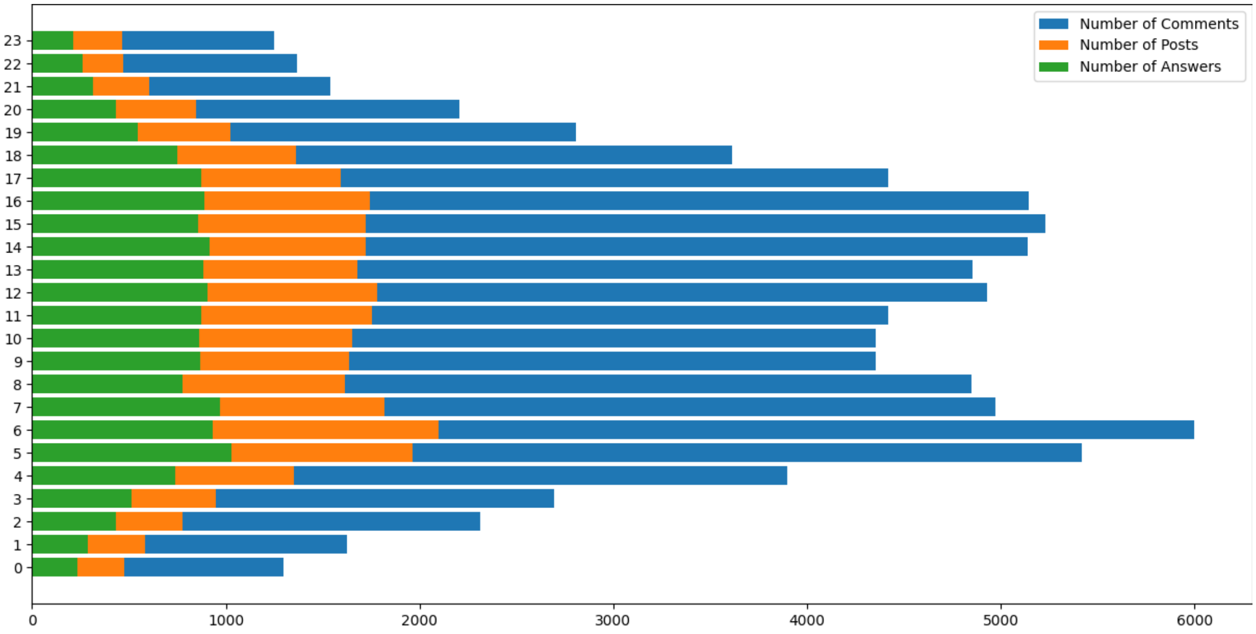
8.9.5 Upvote rate vs. Reputation buckets

Upvote rate is calculated as: upvotes/(upvotes + downvotes) * 100) for a user.
We plot a bar graph of upvote rate vs buckets of reputation.



8.10 Best Time to Ask a Question

The plot depicts the relative number of posts, comments and answers during different hours of the day (the hours are based on the GMT timezone).
We notice there is a high peak around 11:30 AM IST (6 AM GMT), and then there is a continuous active streak up until the night, around 10:30 PM IST (5 PM GMT).



8.11 Association Rule Mining

- Badges/Tags from each User/Post are encoded as a one-hot vector.
- We detect frequent itemsets to find the groups of badges/tags that frequently occur together.
- Association Rules are mined to see if presence of certain Badges/Tags imply the presence of others.

8.11.1 Frequent Itemsets

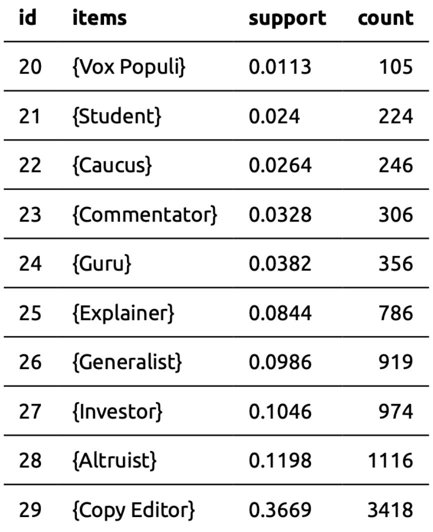
Tags

Here, are the last 10 itemsets. We have only shown 10 for brevity, the entire list can be viewed in the attached csv(ARM_tags_fits.csv).

id	items	support	count
379	{priests,devtas}	0.0018	56
380	{priests,kumarila-bhatta}	0.0025	78
381	{kumarila-bhatta,devtas}	0.0073	230
382	{evil,yoga-vashishtha,priests}	0.0015	46
383	{yama,nimbarkacharya,prisoners}	0.001	33
384	{samskara,lalita-sahasranama,exaltation}	0.0016	50
385	{non-attachment,title,exaltation}	0.0012	38
386	{teaching,nimbarkacharya,prisoners}	0.001	32
387	{appayya-dikshitar,kumarila-bhatta,devtas}	0.0013	42
388	{kumarila-bhatta,lalita-sahasranama,devtas}	0.001	33

Badges

Here, are the last 10 itemsets. Entire list can be viewed in the attached csv(ARM_badges_fits.csv).

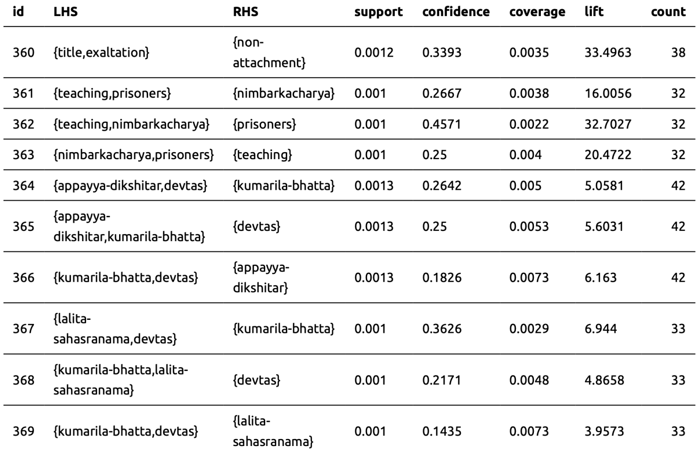


id	items	support	count
20	{Vox Populi}	0.0113	105
21	{Student}	0.024	224
22	{Caucus}	0.0264	246
23	{Commentator}	0.0328	306
24	{Guru}	0.0382	356
25	{Explainer}	0.0844	786
26	{Generalist}	0.0986	919
27	{Investor}	0.1046	974
28	{Altruist}	0.1198	1116
29	{Copy Editor}	0.3669	3418

8.11.2 Association Rules

Tags

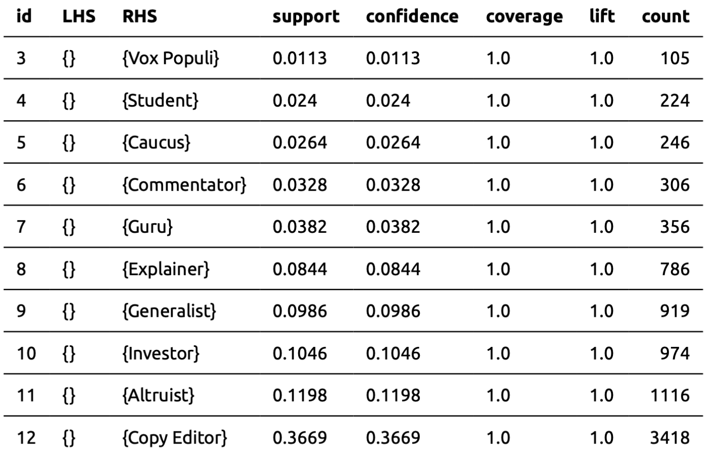
Here, are the last 10 rules. Entire list can be viewed in the attached csv(ARM_tags_mined.csv).



id	LHS	RHS	support	confidence	coverage	lift	count
360	{title,exaltation}	{non-attachment}	0.0012	0.3393	0.0035	33.4963	38
361	{teaching,prisoners}	{nimbarkacharya}	0.001	0.2667	0.0038	16.0056	32
362	{teaching,nimbarkacharya}	{prisoners}	0.001	0.4571	0.0022	32.7027	32
363	{nimbarkacharya,prisoners}	{teaching}	0.001	0.25	0.004	20.4722	32
364	{appayya-dikshitar,devtas}	{kumarila-bhatta}	0.0013	0.2642	0.005	5.0581	42
365	{appayya-dikshitar,kumarila-bhatta}	{devtas}	0.0013	0.25	0.0053	5.6031	42
366	{kumarila-bhatta,devtas}	{appayya-dikshitar}	0.0013	0.1826	0.0073	6.163	42
367	{lalita-sahasranama,devtas}	{kumarila-bhatta}	0.001	0.3626	0.0029	6.944	33
368	{kumarila-bhatta,lalita-sahasranama}	{devtas}	0.001	0.2171	0.0048	4.8658	33
369	{kumarila-bhatta,devtas}	{lalita-sahasranama}	0.001	0.1435	0.0073	3.9573	33

Badges

Here, are the last 10 rules. Entire list can be viewed in the attached csv(ARM_badges_mined.csv).

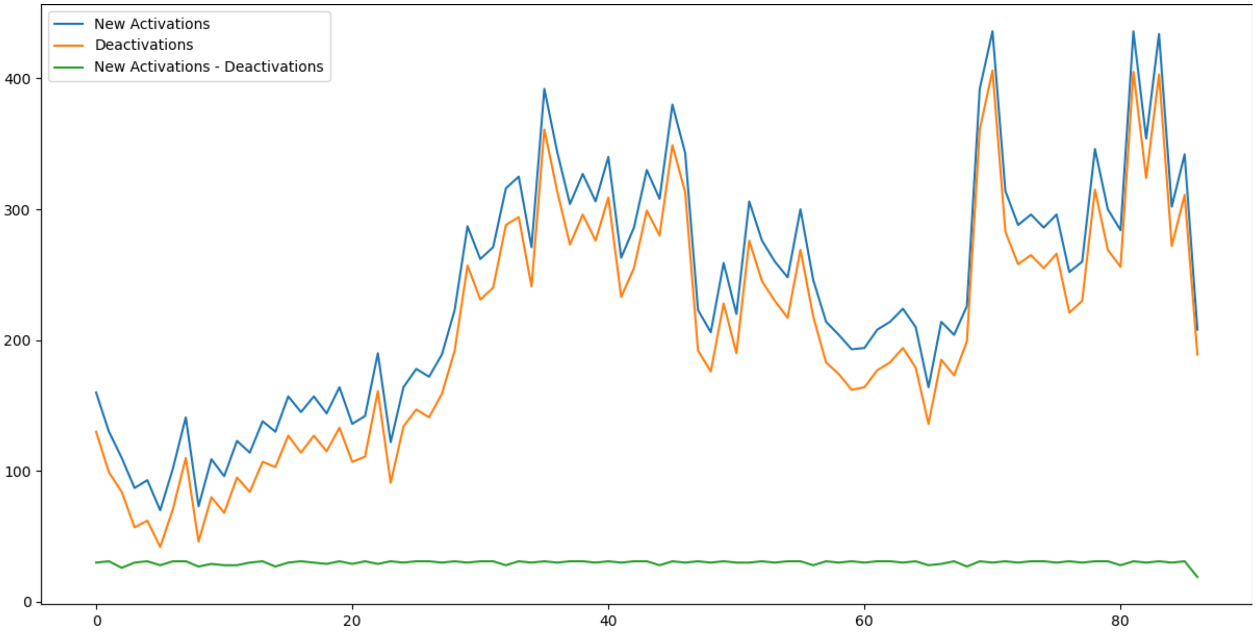


id	LHS	RHS	support	confidence	coverage	lift	count
3	{}	{Vox Populi}	0.0113	0.0113	1.0	1.0	105
4	{}	{Student}	0.024	0.024	1.0	1.0	224
5	{}	{Caucus}	0.0264	0.0264	1.0	1.0	246
6	{}	{Commentator}	0.0328	0.0328	1.0	1.0	306
7	{}	{Guru}	0.0382	0.0382	1.0	1.0	356
8	{}	{Explainer}	0.0844	0.0844	1.0	1.0	786
9	{}	{Generalist}	0.0986	0.0986	1.0	1.0	919
10	{}	{Investor}	0.1046	0.1046	1.0	1.0	974
11	{}	{Altruist}	0.1198	0.1198	1.0	1.0	1116
12	{}	{Copy Editor}	0.3669	0.3669	1.0	1.0	3418

8.12 Active users over time

Here, we analyse the number of active users over time on stack exchange. There is a good observation, that the number of users increasing ($activations - deactivations$) remains constant over the months, whereas the number of activations and deactivations is varying.

New users joining every month


New Activations
Deactivations
New Activations - Deactivations