Question 3 Part B

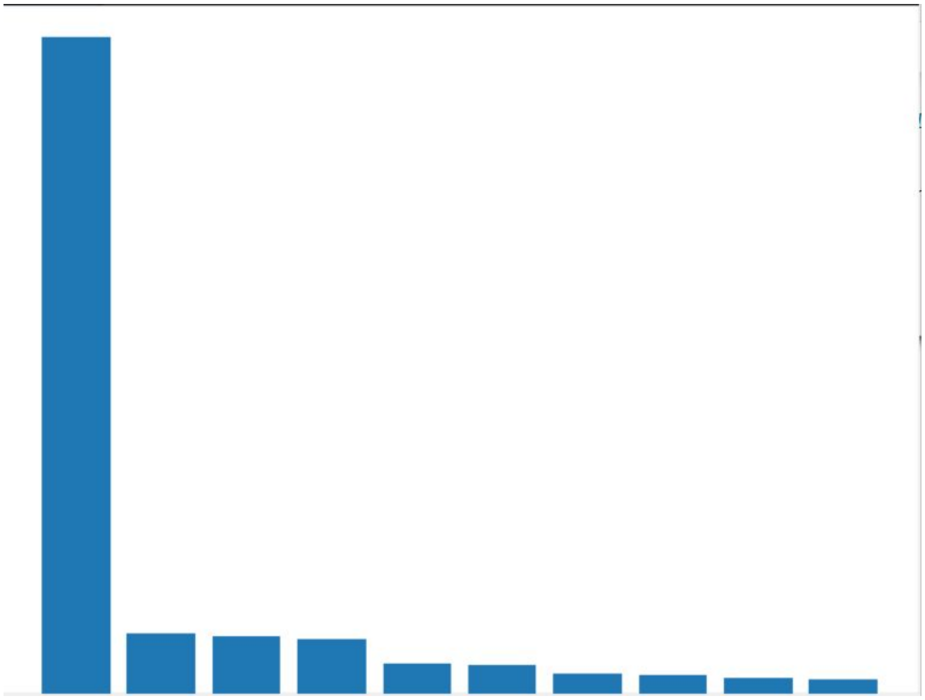
MongoDB Collection of the parsed XML Files

Link to the File Collection

To parse the xml files we used a python module: xmldictparse that converts xml files to json strings and then the json strings are stored in the mongodb database.

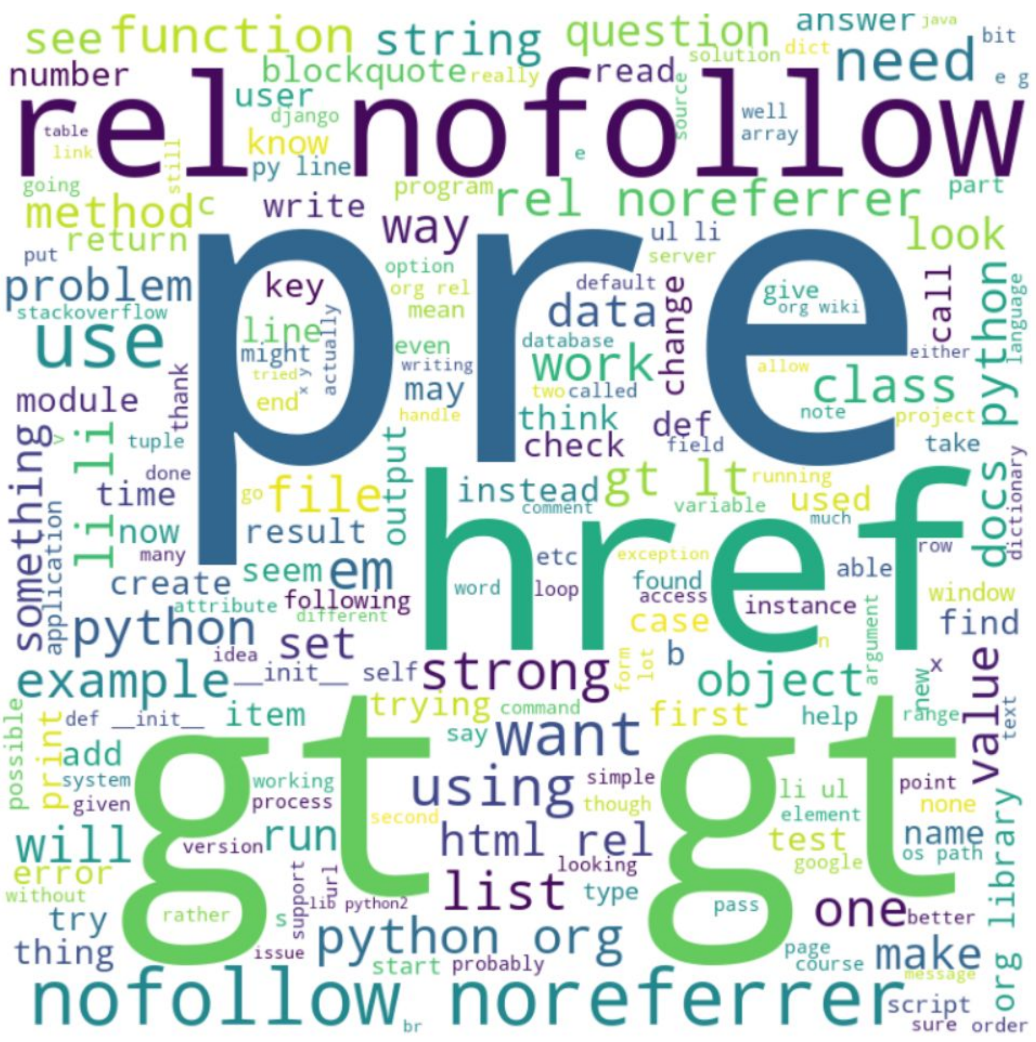
The analysis done below for the five XML files indicates that for subsampling, posts, users and tags have been chosen on the basis of if they are related to computer science(mainly application development and data science).

1. Data of top ten tags(sorted in a descending order): ('python', 1358860), ('python-3.x', 125609), ('pandas', 119220), ('django', 111953), ('numpy', 62820), ('python-2.7', 60590), ('list', 42883), ('matplotlib', 38246), ('dataframe', 33927), ('dictionary', 29821)



The bar plot shows that the number of python posts are much higher than the number of python-3 pots and on observing carefully, all the top ten occurring tags are related to python libraries. Django is a python framework used for web development, list and dictionaries are python structures. We also see the terms matplotlib and data frame with a high frequency and these are related to data science.

2. Drawing the word cloud of the text in all the posts we get:

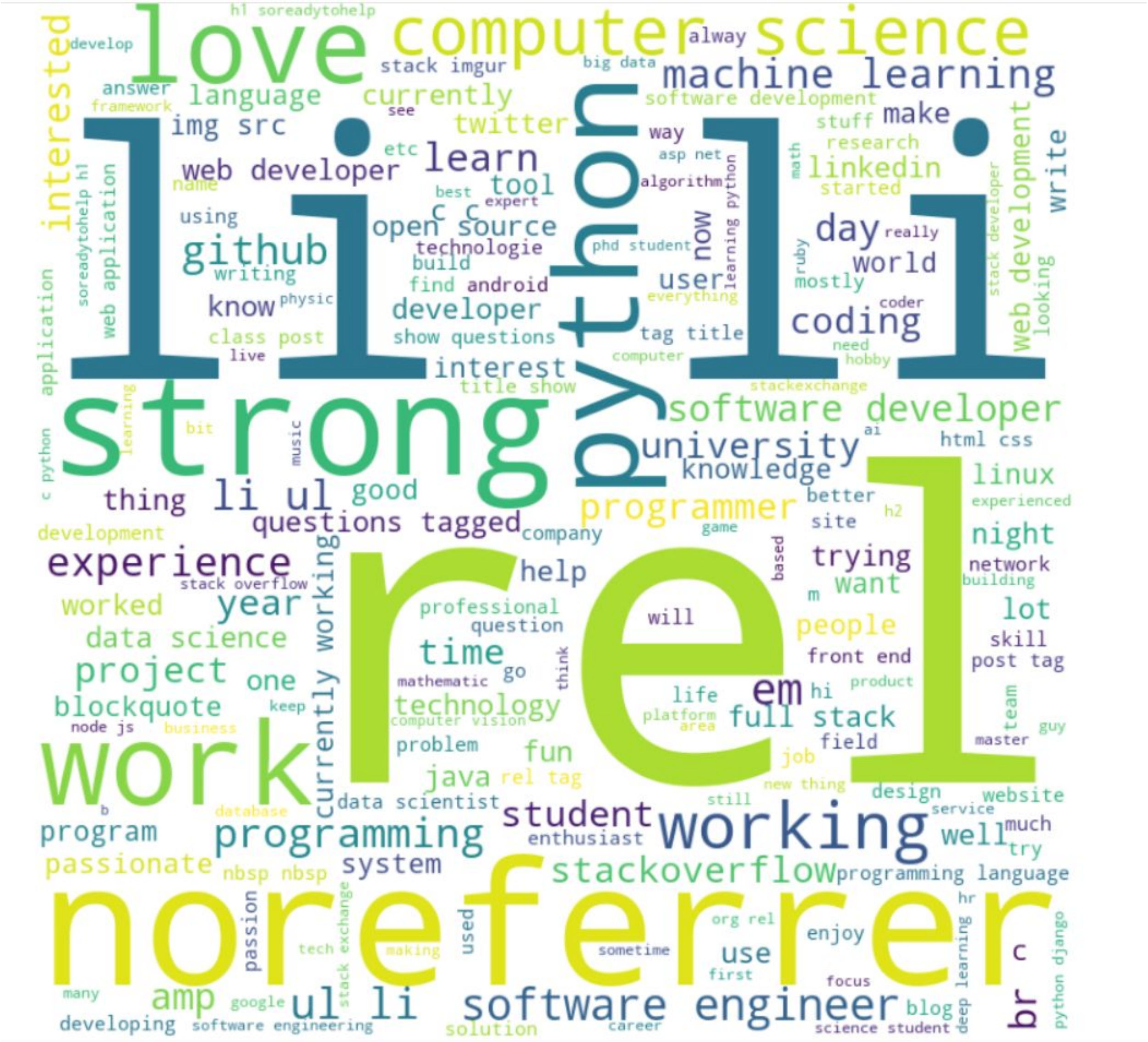


Two main points that one can get from this word cloud:

1. Python is the most frequently used word. Most of the keywords used here (try, __init__, list, dict, print, module etc) are python related words. Other terms are programming-related that are used frequently in the development of applications and code.

2. HTML terms are used frequently which indicates the development of web applications being popular.

Now let’s draw a word cloud for user descriptions.

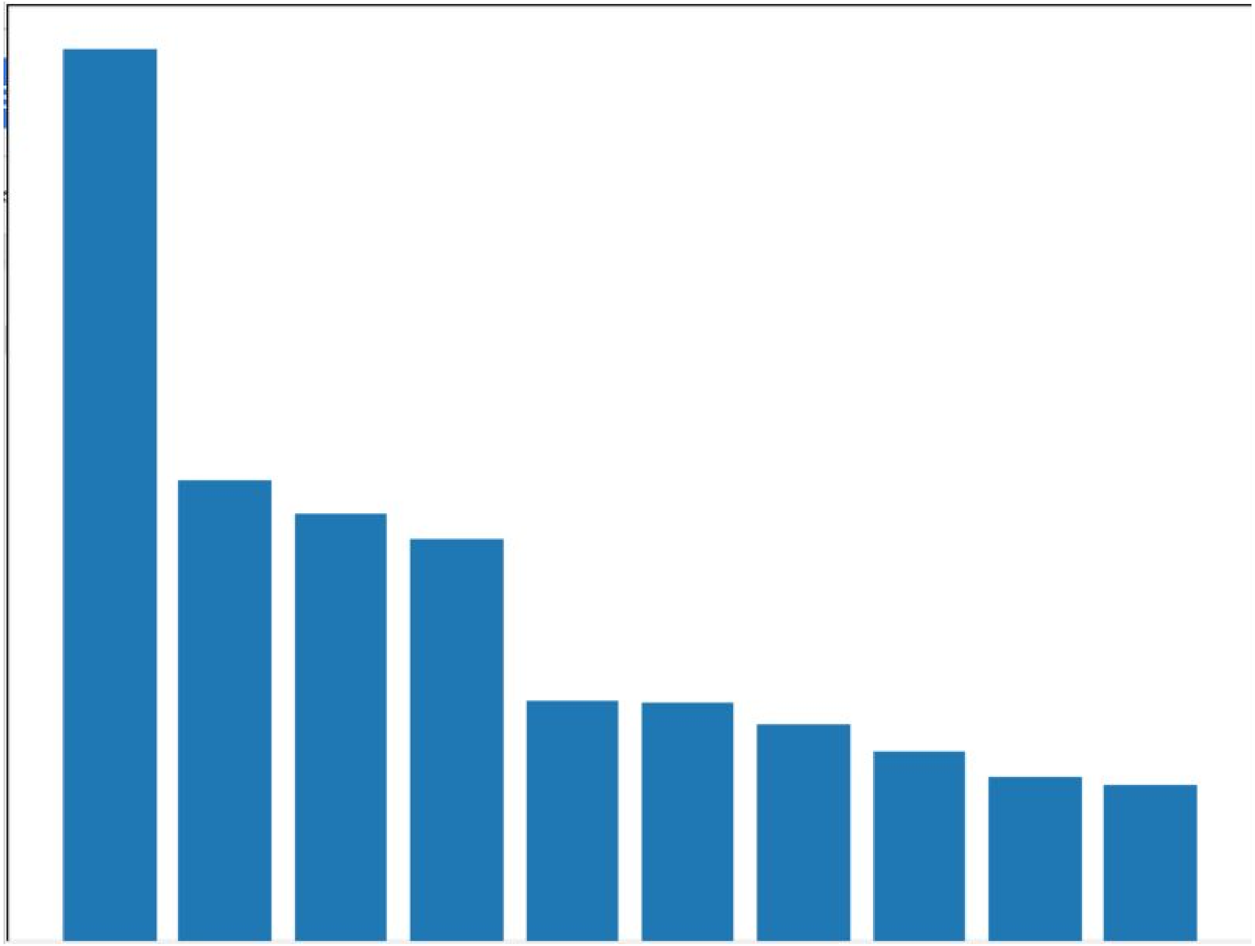


We can draw the following insights from this word Cloud:

- Most of the users have described themselves as a student/engineer/programmer/working, and used adjectives like love/passionate/interest and this shows that the user base is passionate about learning and is working in a science-related field
- One can notice heavy use of HTML tags, terms like python/data science/java/android, and other tech terms which implies that most of the users are working in the field of web development/app development/data science/software engineering.

3. Now we will filter the top ten badges. Note that the dictionary is in a sorted order.

```
[('Popular Question', 1619728), ('Notable Question', 835164), ('Yearling', 774306), ('Nice
        Answer', 729652), ('Editor', 437171), ('Student', 431185), ('Scholar', 392565),
        ('Supporter', 344190), ('Teacher', 299054), ('Custodian', 284517)]
```



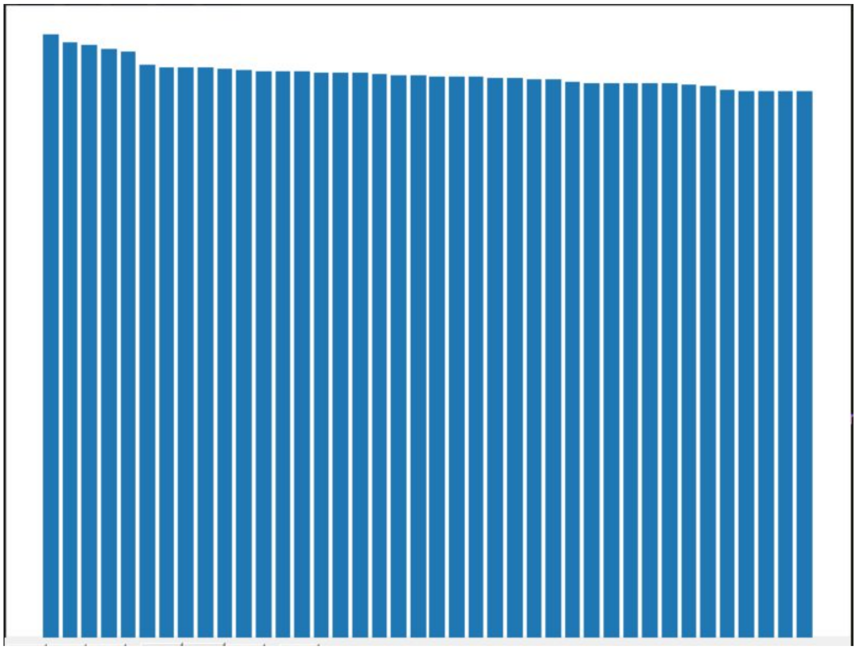
This is the corresponding graph plotted.

We can infer that all the badges are related to positive encouragement and contribute to creating a healthy and helpful community, which is a positive motivation for us :).

4. For votes:

Dictionary of the creation date and corresponding number of votes (in a sorted order):

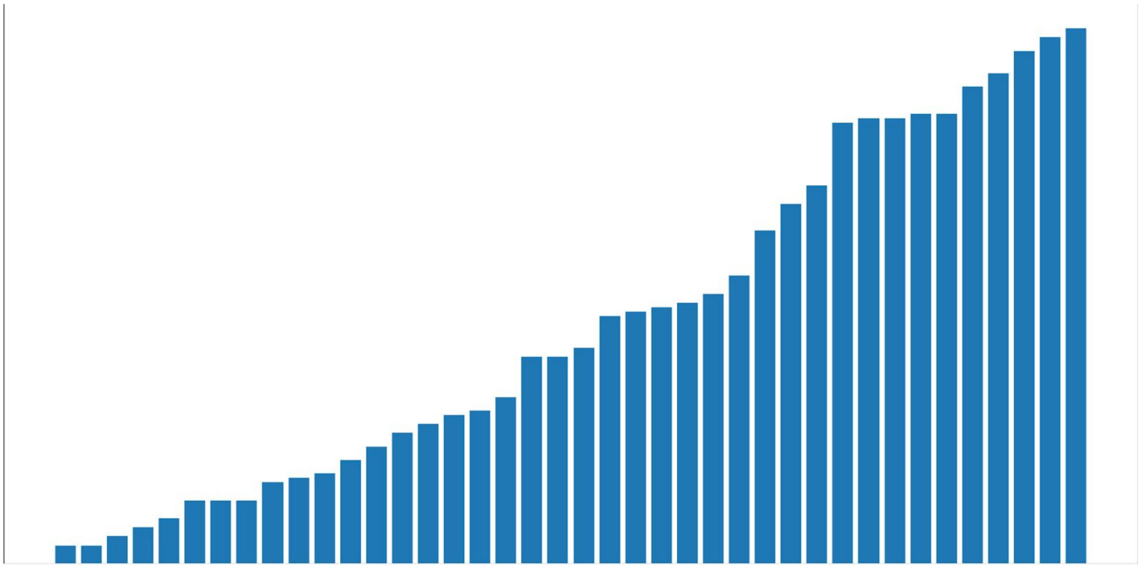
```
('2019-11-20T00:00:00.000', 9540), ('2020-02-26T00:00:00.000', 9412),
          ('2020-02-19T00:00:00.000', 9365), ('2019-11-21T00:00:00.000', 9313),
          ('2020-02-20T00:00:00.000', 9258), ('2020-02-25T00:00:00.000', 9066),
          ('2019-03-06T00:00:00.000', 9012), ('2019-12-03T00:00:00.000', 9010),
          ('2019-11-26T00:00:00.000', 9005), ('2019-11-14T00:00:00.000', 9003),
          ('2020-02-18T00:00:00.000', 8964), ('2019-11-19T00:00:00.000', 8951),
          ('2020-01-22T00:00:00.000', 8948), ('2020-02-12T00:00:00.000', 8945),
          ('2020-02-04T00:00:00.000', 8925), ('2019-03-12T00:00:00.000', 8924),
          ('2020-02-05T00:00:00.000', 8924), ('2020-02-27T00:00:00.000', 8915),
          ('2019-02-20T00:00:00.000', 8891), ('2019-11-18T00:00:00.000', 8883),
          ('2019-12-11T00:00:00.000', 8873), ('2020-01-21T00:00:00.000', 8863),
          ('2020-02-06T00:00:00.000', 8860), ('2019-02-13T00:00:00.000', 8853),
          ('2019-03-19T00:00:00.000', 8844), ('2019-03-26T00:00:00.000', 8834),
          ('2020-02-13T00:00:00.000', 8834), ('2019-02-12T00:00:00.000', 8796),
          ('2019-02-19T00:00:00.000', 8768), ('2019-03-14T00:00:00.000', 8768),
          ('2019-02-21T00:00:00.000', 8764), ('2019-04-16T00:00:00.000', 8760),
          ('2019-02-27T00:00:00.000', 8757), ('2019-05-14T00:00:00.000', 8748),
          ('2019-12-04T00:00:00.000', 8716), ('2019-12-12T00:00:00.000', 8654),
          ('2019-12-17T00:00:00.000', 8648), ('2019-03-28T00:00:00.000', 8647),
          ('2019-03-27T00:00:00.000', 8646), ('2020-01-16T00:00:00.000', 8642)
```



So , all the highest number of posts are all created in the year 2019-2020, about 9000 votes created (with a difference of 500 votes more or less) per period. Since there is low variation, to know if this is the case with the whole sample, we decided to plot for the days where the activity was least.

So we find that this is the activity was very low in the year of 2008

```
[('2008-08-10T00:00:00.000', 4), ('2008-08-14T00:00:00.000', 4), ('2008-08-02T00:00:00.000',
      6), ('2008-08-09T00:00:00.000', 8), ('2008-08-16T00:00:00.000', 10),
      ('2008-08-11T00:00:00.000', 14), ('2008-08-15T00:00:00.000', 14),
      ('2008-08-23T00:00:00.000', 14), ('2008-08-08T00:00:00.000', 18),
      ('2008-08-24T00:00:00.000', 19), ('2008-08-03T00:00:00.000', 20),
      ('2008-08-12T00:00:00.000', 23), ('2008-08-04T00:00:00.000', 26),
      ('2008-08-13T00:00:00.000', 29), ('2008-08-07T00:00:00.000', 31),
      ('2008-08-20T00:00:00.000', 33), ('2008-08-19T00:00:00.000', 34),
      ('2008-08-17T00:00:00.000', 37), ('2008-08-05T00:00:00.000', 46),
      ('2008-08-27T00:00:00.000', 46), ('2008-09-07T00:00:00.000', 48),
      ('2008-08-26T00:00:00.000', 55), ('2008-08-18T00:00:00.000', 56),
      ('2008-09-06T00:00:00.000', 57), ('2008-08-22T00:00:00.000', 58),
      ('2008-09-14T00:00:00.000', 60), ('2008-08-06T00:00:00.000', 64),
      ('2008-08-25T00:00:00.000', 74), ('2008-11-02T00:00:00.000', 80),
      ('2008-08-21T00:00:00.000', 84), ('2008-08-28T00:00:00.000', 98),
      ('2008-09-05T00:00:00.000', 99), ('2008-12-26T00:00:00.000', 99),
      ('2008-08-31T00:00:00.000', 100), ('2008-11-15T00:00:00.000', 100),
      ('2008-12-25T00:00:00.000', 106), ('2008-09-13T00:00:00.000', 109),
      ('2008-10-12T00:00:00.000', 114), ('2008-12-27T00:00:00.000', 117),
      ('2008-09-03T00:00:00.000', 119)]
```

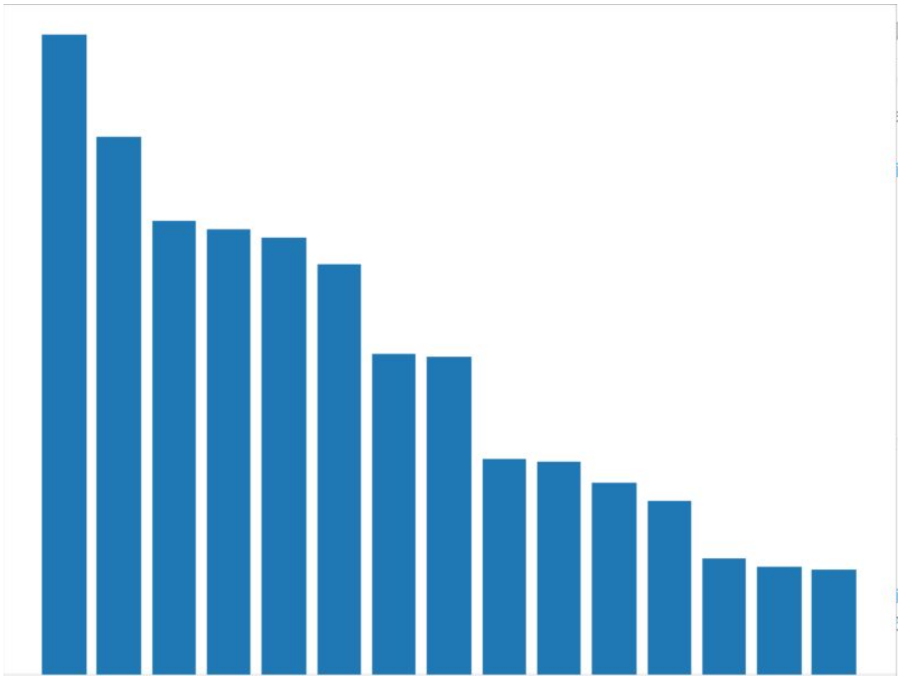


Note: All the bars in the graph correspond to the exact values in the dictionary

This shows that the platform has grown exponentially with time and in the recent years it has had stable amount of activity.

Now the Tags.XML file has a list of all possible tags. This will be very helpful for us to understand the criteria for subsampling. So we picked up the top 15 occurring tags (according to their frequency mentioned in the XML file).

```
[('javascript', 3911114), ('java', 3282204), ('c#', 2770440), ('python', 2718252), ('php',
2670100), ('android', 2508964), ('jquery', 1956824), ('html', 1941398), ('c++',
1313938), ('css', 1298872), ('mysql', 1175050), ('sql', 1060338), ('asp.net', 706738),
('r', 659140), ('c', 645200)]
```



Thus we see that the top 7 tags are the programming languages that are used mainly for app development and the last 8 tags are predominantly used by data scientists.