Mid-term

Abhigya Koirala
akoirala@lakeheadu.ca
Lakehead University
Thunder Bay, Ontario, Canada

1 INTRODUCTION

This task deals with Analysis and Prediction of 'Closed Questions' on Stack Overflow [1]. The researchers in this paper have analyzed and characterize the complete set of 0.1 Million 'closed' questions. They then machine learning framework and build a predictive model to identify a 'closed' question at the time of question creation. A detailed analysis has been done and number of graphs and results have been produced at the end of this paper citing different problems and findings.

For this task assigned of Analysis and Prediction of 'Closed Questions' on Stack Overflow, The dataset given is analysed witth the help of Jupyter notebook invloving different sets of libraries. The detailed process has been explained in the further sections.

2 TASKS

There are seven tasks assigned in total for this assignment. The methodology for each of the tasks is further explained in the section below.

2.1 Task1: Categorizing closed questions by year.

In this task we were assigned to show a table with different percentage of closed questions each year. Since were only provided with close questions in our dataset for this task we also needed number of open-questions as well. For this task I used stack exchange data explorer to get the total number of questions as shown in Figure 1.

```
select COUNT(Id) As TotalPost, datepart(yyyy, [CreationDate])
from Posts
where PostTypeId = 1
group by datepart(yyyy, [CreationDate])
order by datepart(yyyy, [CreationDate])
```

Figure 1: SDEE Querry for taking out all question

This returned me with the total questions count for each year. From the dataset given closed question count for each year was then calculated and then the percentage of closed questions each year was found as shown in the Figure 2

From the table obtained we can find that in 2014 most questions were closed (5.59%) where as in 2008 least questions were closed (0.32%).

For the second task, to categorize percentage of the questions by the category the closed questions were grouped by the respective category The categories were Duplicate,Off-topic,Subjective,Unclear and broad. The count for all the closed questions in different categories were calculated and the result was shown in form of a Pie-chart as shown in the Figure 3

	year	TotalPost	PostClosed	Percentage
0	2008	57873	184	0.32%
1	2009	342338	1171	0.34%
2	2010	691891	5722	0.83%
3	2011	1191876	25449	2.14%
4	2012	1632381	67745	4.15%
5	2013	2043402	111560	5.46%
6	2014	2148512	120158	5.59%
7	2015	2204256	97416	4.42%
8	2016	2206825	107733	4.88%
9	2017	2122430	109673	5.17%
10	2018	1899349	86035	4.53%

Figure 2: Percentage of closed questions by cattegory

From the pie chart obtained we can find that the duplicate category involved most closed question 54.5% where as subjective category involved least closed questions 6.2%

2.2 Task 2: Temporal distribution analysis of 'closed' questions on Stack Overflow over the 120 months time window between August 2008 to August 2018)

For this task we were asked to do temporal distribution analysis of closed questions on Stack Overflow over the 120 months time window between August 2008 to August 2018. The closed questions were grouped into different 5 categories. The count of closed question in each category by their closed dates was obtained. The date were divided in span of 6 months from 2008 to 2018. Time window was taken for the grouping is shown in the Figure 4.

After obtaining the count for each category of questions in different time window, the ratio of closed questions for that particular category to total question closed was obtained. A line graph was made out after obtaining the ratio as shown in the Figure 5. On the x-axis is time slots and respective time slots can be figured out from the Figure 4. Overall, we find an decreasing trend of the percentage of 'closed' questions in each category i.e. we find that

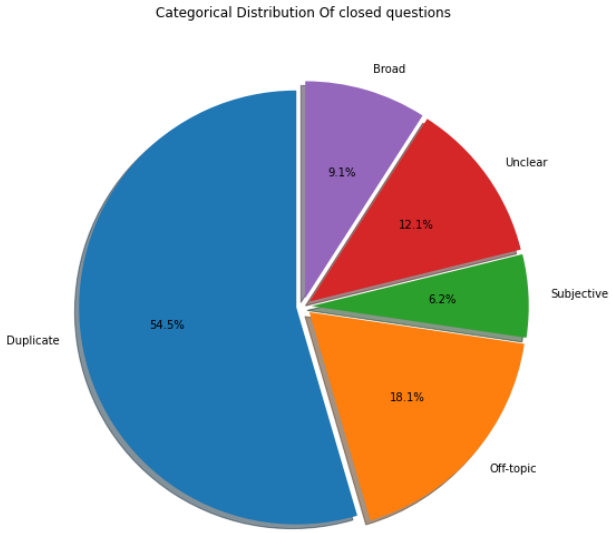

Figure 3: Percentage of closed questions by category

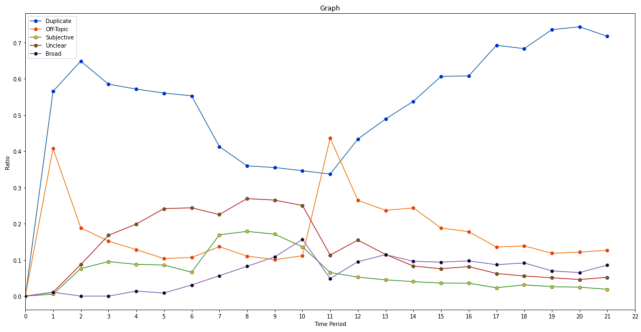

Figure 4: Time window taken for grouping

the number of questions 'closed' over time has an downward curve after time-period 12 which is 2014-01-01 to 2014-07-01. We also see that the most common categories of 'closed' questions over 2008 to 2018 are Exact Duplicate and Off-topic category.

2.3 Task 3: Effect of newly registered users)

For this task we were asked whether closed questions are asked by newly registered users or not. For this purpose a time duration as shown in Figure 4 is taken into consideration. The ratio of closed questions asked by newly registered users over all closed questions asked in that time duration is obtained. To determine if the user of the closed question is a newly registered user, the criteria was that if the user of the closed question created the account on the same month the question was asked, consider the user was taken as newly registered user. After finding out the ratio the line graph was obtained as shown in Figure 6.

	StartDate	EndDate
0	2008-01-01 00:00:00.000000000	2008-07-01 00:00:00.000000000
1	2008-07-01 00:00:00.000000000	2009-01-01 00:00:00.000000000
2	2009-01-01 00:00:00.000000000	2009-07-01 00:00:00.000000000
3	2009-07-01 00:00:00.000000000	2010-01-01 00:00:00.000000000
4	2010-01-01 00:00:00.000000000	2010-07-01 00:00:00.000000000
5	2010-07-01 00:00:00.000000000	2011-01-01 00:00:00.000000000
6	2011-01-01 00:00:00.000000000	2011-07-01 00:00:00.000000000
7	2011-07-01 00:00:00.000000000	2012-01-01 00:00:00.000000000
8	2012-01-01 00:00:00.000000000	2012-07-01 00:00:00.000000000
9	2012-07-01 00:00:00.000000000	2013-01-01 00:00:00.000000000
10	2013-01-01 00:00:00.000000000	2013-07-01 00:00:00.000000000
11	2013-07-01 00:00:00.000000000	2014-01-01 00:00:00.000000000
12	2014-01-01 00:00:00.000000000	2014-07-01 00:00:00.000000000
13	2014-07-01 00:00:00.000000000	2015-01-01 00:00:00.000000000
14	2015-01-01 00:00:00.000000000	2015-07-01 00:00:00.000000000
15	2015-07-01 00:00:00.000000000	2016-01-01 00:00:00.000000000
16	2016-01-01 00:00:00.000000000	2016-07-01 00:00:00.000000000
17	2016-07-01 00:00:00.000000000	2017-01-01 00:00:00.000000000
18	2017-01-01 00:00:00.000000000	2017-07-01 00:00:00.000000000
19	2017-07-01 00:00:00.000000000	2018-01-01 00:00:00.000000000
20	2018-01-01 00:00:00.000000000	2018-07-01 00:00:00.000000000
21	2018-07-01 00:00:00.000000000	2019-01-01 00:00:00.000000000

Figure 5: Temporal distribution analysis of 'closed' questions on Stack Overflow over the 120 months time window between August 2008 to August 2018

From the graph obtained we can find that there is steep increase in the percentage of question being asked by the user between time period 0-6 and a slight decrease after that time period. The highest percentage of closed question being asked by the new user being almost 30% for the time period 6 and the lowest percentage of closed question being asked by the new user being almost 20% for the time period 1 (Time period 0 data is not available) . The corresponding time period can be obtained from Figure 4 .

2.4 Task 4: Community participation analysis

For this task we were asked to show the temporal distribution of close votes over the studied time period. The closed questions were grouped into different 5 categories. The count votes of closed question in each category by their closed dates was obtained. The date were divided in span of 6 months from 2008 to 2018. Time window was taken for the grouping is shown in the Figure 4. After obtaining the count of votes the line graph was obtained as shown in Figure 7

A significant percentage of almost 40% votes are closed by single votes.

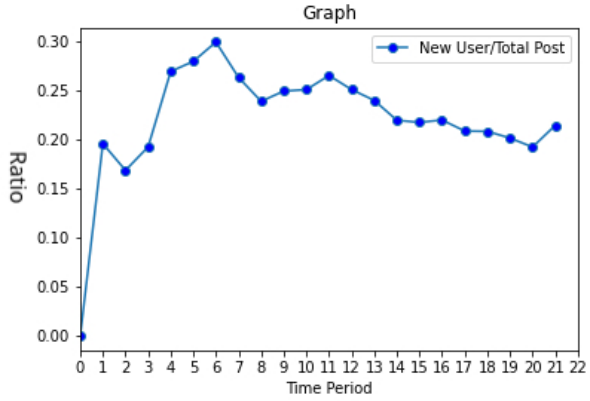

Figure 6: Ratio closed questions are asked by newly registered users to total closed question between August 2008 to August 2018

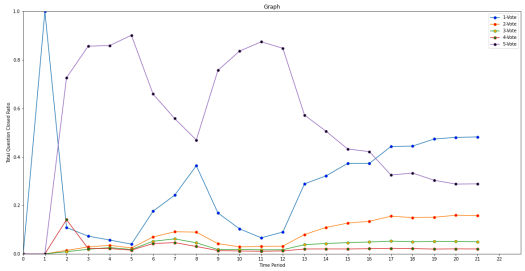

Figure 7: Temporal distribution analysis of votes of 'closed' questions on Stack Overflow over the 120 months time window between August 2008 to August 2018

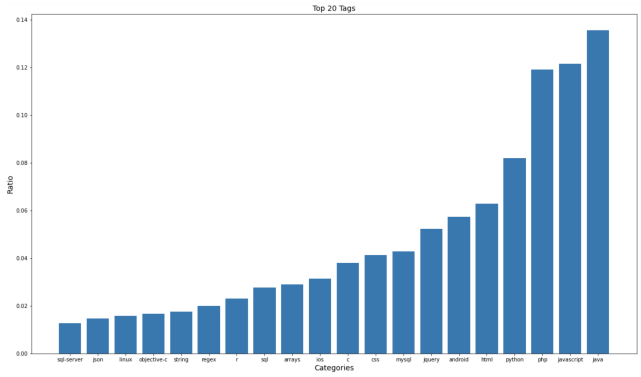

Figure 8: Percentage of tags over the questions

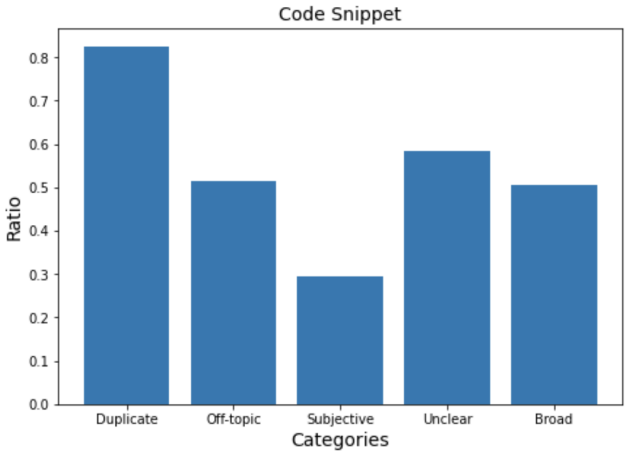

Figure 9: Ratio of code-snippets over the category

2.5 Task 5: Topic analysis

Each Stack Overflow question has some tags associated with it which is an identification of the topic of the question content. For this task ,for each tag in the given dataset, we had to determine the number of closed questions associated with the tag over the total number of closed questions. After that I determine the top-20 tags based on the calculated numbers and then a bar chart was obtained as shown in Figure 8.

Java was found to be the dominant tag in almost 14% pf the question where as sql-server was almost .01%. This infers that Java tagged questions has higher closure rate compared tot other tags.

2.6 Task 6: Content analysis

This task contained 4 sub task. For the first sub task we were asked to find the percentage of code-snippets containing question by different categories. The count of questions with each 5 categories was obtained and than plotted into a bar graph. The following bar graph was obtained as shown in Figure 9.

Duplicate category contained the most code snippet with almost 80% of the questions containing code-snippet whereas subjective category contained least snippet with less than 30% of the questions containing them.

For the second sub task we were asked to calculate the distribution of number of tags in different category. The questions in different categories was first obtained and than tags for each questions were counted. The combined count for each category was than obtained and then box whisker plot was obtained for this task as shown in Figure 10

Duplicate and unclear categories have lesser tags associated with it

For the third sub task we were asked to calculate the distribution of length of title in different category. The questions in different categories was first obtained and than tags for each length of title were counted. The obtained length count for each category was than plotted into box whisker plot as shown in Figure 11

Off-topic and unclear categories have lesser length of title associated with it.

For the fourth sub task we were asked to calculate the distribution of length of body in different category. The questions in different categories was first obtained and than tags for each length of body

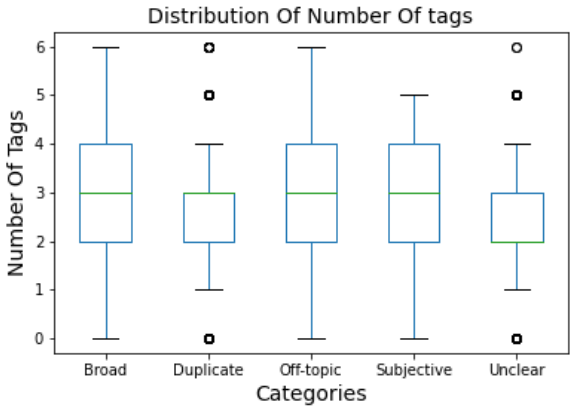

Figure 10: Distribution of tags over category

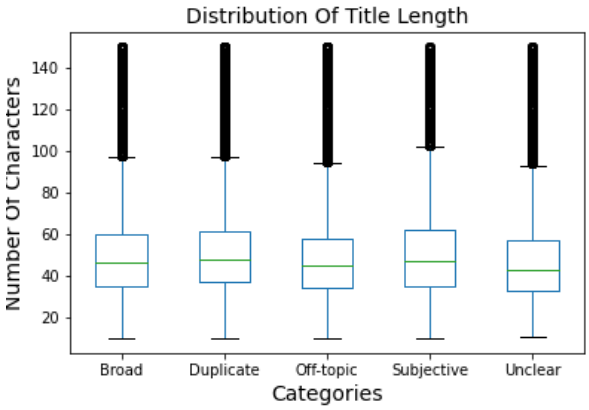

Figure 11: Distribution of length of title over category

were counted. The obtained length count for each category was than plotted into box whisker plot as shown in Figure 12

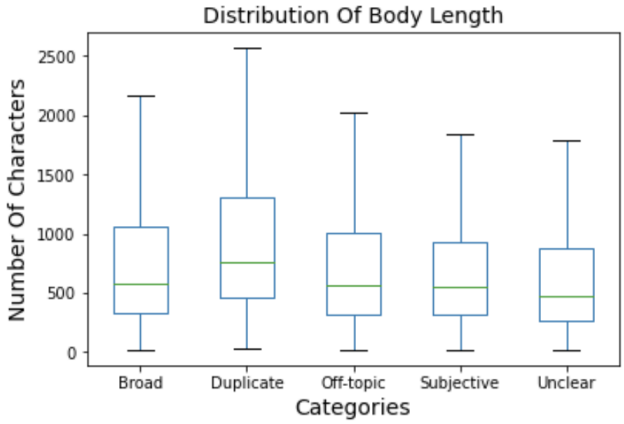

Figure 12: Distribution of length of body over category

Subjective and unclear categories have lesser length of body associated with it.

2.7 Task 7: Closure time

This task required us to determine the time in minutes required to close the question for each category. The difference of creation time and closure time was calculated and changed into minutes for all the question. It was grouped by category and a box whisker plot was plotted as shown in Figure 13.

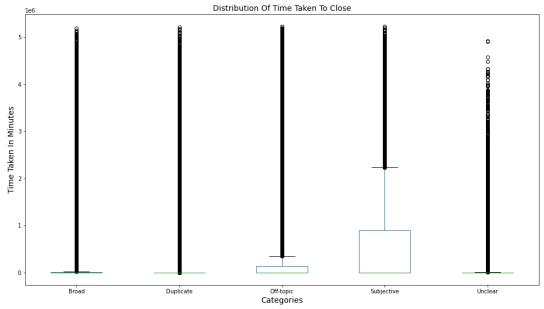

Figure 13: Time taken for closure over the category

REFERENCES

[1] Denzil Correa and Ashish Sureka. 2013. Fit or Unfit : Analysis and Prediction of 'Closed Questions' on Stack Overflow. *CoRR* abs/1307.7291 (2013). arXiv:1307.7291 http://arxiv.org/abs/1307.7291