ANALYSIS OF TOPICS AND RELATIONS IN CROSS VALIDATED R

Venkata Guru Sai Vemulakonda
vvemulak@asu.edu

Jaswitha Vankineni
jvankine@asu.edu

Karthik Lakshmi Narayana Sarma
klakshm6@asu.edu

Sagar Burra
sburra@asu.edu

I Abstract

This project deals with visual Analytics on R data in cross validated website. Today, Q&A sites like stack overflow, cross validated provides a platform for many users to get their queries clarified but none provides the statistics that help both developers and community in knowing about the trends. Analyzing the content helps community to better understand the needs of the developers. In this project, we have used text analysis technique which is term frequency to analyze the post content and provide the users with most trending topics in R programming language. We have also analyzed the related topics for the most frequently used topics and their related posts by performing topic modeling. Projecting most related content to the user when two specific topics are selected is a key aspect of our project. Our project besides providing the trending topics and related posts, it also provides the user analysis between the users who has posted data for a particular post which provides user a privilege of selecting the best answer based on the proficiency of the user.

Keywords: *Q&A sites, text analysis, topic modeling, user analysis, proficiency computation.*

II Introduction

Cross Validated, which is similar to stack overflow contains thousands of threads on various topics. These topics can belong to statistics, machine learning, data analysis & mining. We did an analysis on posts related to R programming and built an interactive analytics forum so that a user can find solutions to problems through topics which are interrelated.

Frequently, a user finds solution to a programming question through stack overflow by selecting answers which might have received most number of votes. For some languages, problems can be very specific, causing less number of user votes, views and comments. We categorized posts based on top 30 most frequently occurring topics in R programming language. An analysis on the topic – topic similarity helped identify posts which are specific to certain domain and enabled users find solutions through post – topic similarity.

Our dashboard allows users to visualize the top 30 frequently discussed topics from R programming language. Our analysis involves studying the interaction of 5 most related topics for each of the top 30 topics that we extracted. Analyzing post data which is linked to 2 topics help understand the relationship between those two topics. The number of posts linked to the 2 respective topics shows probability of co-occurrence of 2 topics. We portrayed the similarity by taking Euclidean distance as a measure.

Similarity values were assigned between 0 to 1. 0 being least similar and 1 being most similar between any two topics. Posts with similarity values higher than or equal to 0.3 towards each of the 2 selected topics are labelled significant. This project consists of the following modules:

- Identifying top 30 most frequently used topics
- Identifying post content which is similar to any two selected topics from the topics specified above.
- Identifying solutions based on votes
- Identifying solution based on user statistics
- User analysis for each post and identifying solution based on user reputation, up votes, down votes and views.

To illustrate our method, let’s consider the two topics ‘probability’ and ‘distributions’. These topics co-occur frequently since probability is an important metric in identifying distributions. The plot for these topics show post content having questions related to these topics. Some of them can even be classified as significant since these two are closely related.

III Dataset

Cross Validated is an Online Question & Answers site that provides a platform for all developers to get their technical problems clarified. For each post made by the user, other users can up vote, down vote etc. which helps user to determine the best answer. Each user also has reputation and badges which helps to determine user proficiency in those topics.

We have collected 14,109 posts of type questions that belong to programming language R. For each post their respective answers and comments are retrieved in order to show answer analysis which are 13,781 in number.

For all posts of type questions and answers, user data is collected to portray user proficiency in terms of reputation, views, up votes and down votes.

All the required information is collected from query stack exchange (https://data.stackexchange.com/stats/query/new) which helps to query the database of cross validated website.

All the data is queried using simple SQL and results are stored in csv format which helped in projecting the data easily.

IV Motivation

Analyzing content of ‘Cross Validated’ forum which are specific to R programming language, discovering the most frequently used topics, their relationship and analysis of post content belonging to respective topics is the objective of this project. Analyzing semantic structures which are latent in post data enables engineering community to cater the needs of developers better. We often want to know what developers are mostly talking about. An analysis of data from stack exchange will give a clear and general idea about finding constructive answers to post data. We also want to know why developers are talking about a topic in a thread, when are they talking about topics which are interrelated. Based on this data, we draw different conclusions on how different topics are interrelated and find methods to distinguish sensible answers from the forum.

These questions arise because of various reasons. Researchers want to know what are the new trends in a language that the communities/developers are discussing. This helps them understand the need of the developers or the people who are participating in the discussions.

In this project, we will discuss about how we have extracted and identified most important topics in R programming language and how we identified the most closely related topics among these top topics.

V Visualization Design

Our dashboard consists of a word cloud having most frequently used 30 topics as circles. The size of the circle signifies the frequency with which the topic is being used in the forum. The snapshot of the homepage is as shown in Figure 1. The frequency order was identified based on the word frequency vector extracted from the post content.

The top 5 most related topics pertaining to each topic gets highlighted once the user clicks on any of the displayed topics on the word cloud. This is shown in Figure 2. Our visualization strategy was to steer the user to subsequent visualizations upon clicking. This helps highlight only required visuals and abstracts those which are unimportant. The dynamic nature of our dashboard allows flexibility of changing data for any domain of choice. Layering front end helps updating information in future.

Upon clicking the highlighted topic, a graph having topic 1 similarity on x axis and topic 2 similarity on Y axis appears.

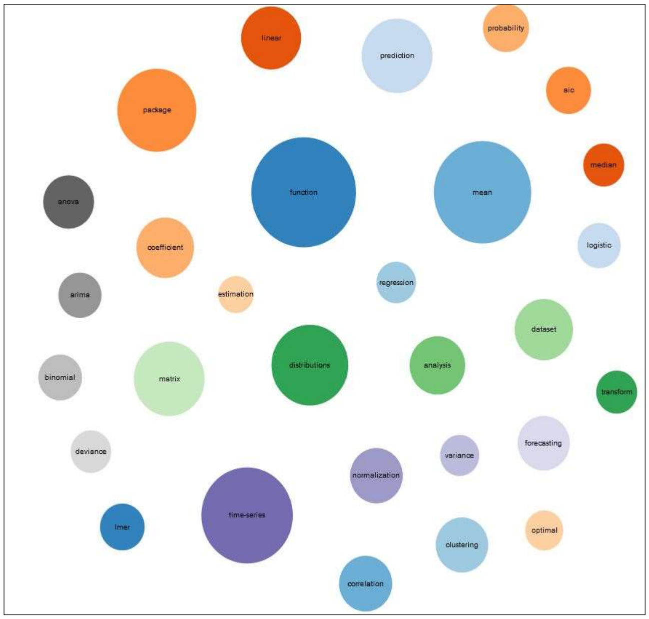


Figure 1: Word cloud showing most frequently occurring topics

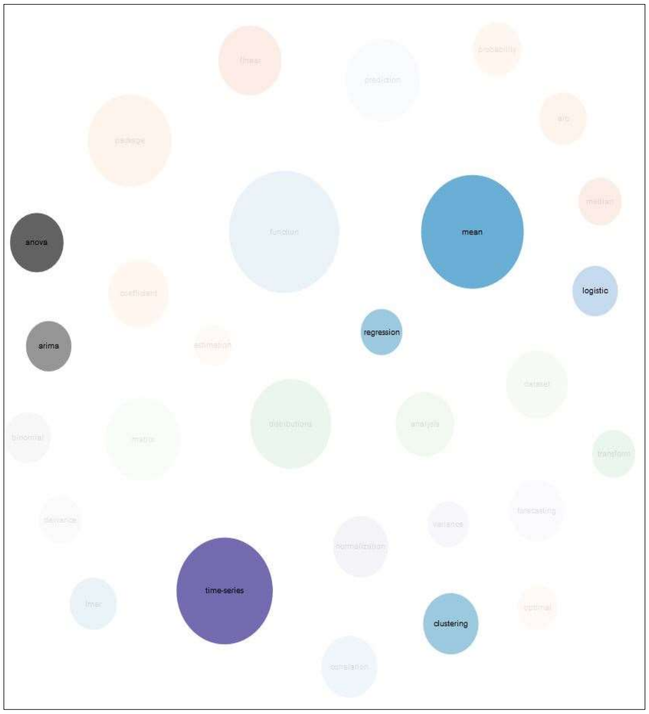
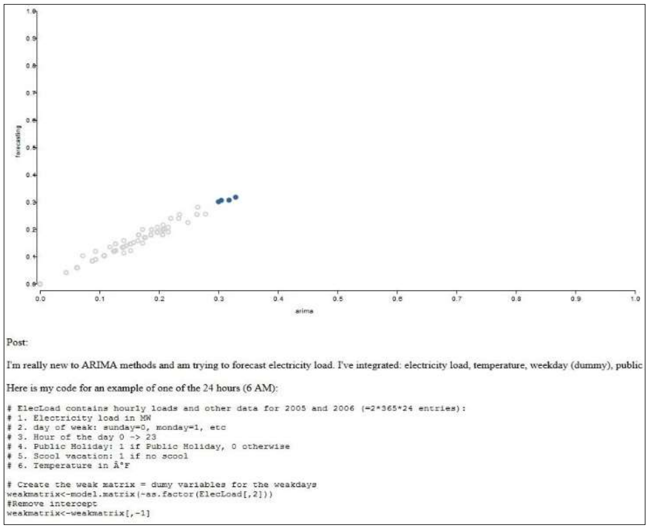


Figure 2: Top 5 most related topic for each topic

Taking one topic on the X-axis and the other on the Y-axis allows posts to be evaluated on a 2D space. Posts which lie in between and farther away represents higher similarity measures for each of the 2 topics. Hovering post dots show post content below the graph. Figure 3 shows the similarity between the topics 'Arima' and 'Forecasting'. As we can see, the insignificant posts are represented in grey color. These posts fall below the threshold for being significant. Those represented in blue color represents significant posts. The number of blue dots represent the frequency with which interrelated topics are discussed in the forum. Hovering dots on the graph shows post content below the graph. Wrapping information in data visualization gets difficult when the volume of data is huge. The two methods to simplify the process is to display information upon zooming and to display specific information at a time and to provide a layer of abstraction for each phase. We decided to follow the second approach.



Post:

I'm really new to ARIMA methods and am trying to forecast electricity load. I've integrated: electricity load, temperature, weekday (dummy), public

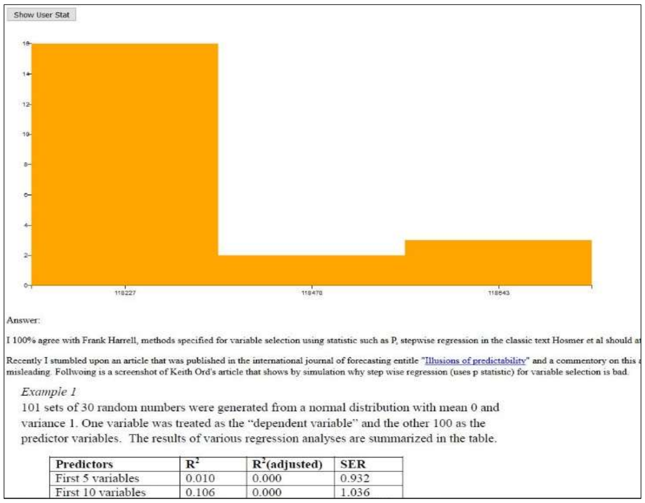
Here is my code for an example of one of the 24 hours (6 AM):

```
# ElecLoad contains hourly loads and other data for 2005 and 2006 (=2*365*24 entries):
# 1. Electricity load in MW
# 2. day of weak: sunday=0, monday=1, etc
# 3. Hour of the day 0 -> 23
# 4. Public Holiday: 1 if Public Holiday, 0 otherwise
# 5. Scool vacation: 1 if no scool
# 6. Temperature in Â°F

# Create the weak matrix = dumy variables for the weakdays
weakmatrix<-model.matrix(~as.factor(ElecLoad[,2]))
#Remove intercept
weakmatrix<-weakmatrix[,-1]
```

Figure 3: Similarity graph for topics 'Arima' and 'Forecasting'

Finding solutions to questions becomes difficult when the topics are not discussed often. This leads to unpredictability of right answers. The solutions might have less number of up votes, less down votes, low view count and it solely depends on the reputation of the user. Hence, we needed a mechanism to visualize answers to post data. Our visualization allows users to select answers based on up votes if it is available. We provided a bar graph so that users could select answers based on height difference for answer bar's.



Answer:

I 100% agree with Frank Harrell, methods specified for variable selection using statistic such as P, stepwise regression in the classic text Hosmer et al should a

Recently I stumbled upon an article that was published in the international journal of forecasting entitle "Illusions of predictability" and a commentory on this misleading. Follwoing is a screenshot of Keith Ord's article that shows by simulation why step wise regression (uses p statistic) for variable selection is bad.

Example 1

101 sets of 30 random numbers were generated from a normal distribution with mean 0 and variance 1. One variable was treated as the "dependent variable" and the other 100 as the predictor variables. The results of various regression analyses are summarized in the table.

Predictors	R^2	R^2(adjusted)	SER
First 5 variables	0.010	0.000	0.932
First 10 variables	0.106	0.000	1.036

Figure 4: Bar chart showing answers based on scores

Hovering bar chart allows users to visualize answers for the posts. X axis represents post Id's and Y axis represents 'Scores'. This graph doesn't appear for posts which are unanswered. Bar chart enables users to select answers easily rather than scrolling through the entire thread to select the right answer. There can be multiple ways to approach the same problem. In that case, choosing the right answer for a problem becomes easier with bar chart. When the vote count is low for answers which belong to post topics which are not so frequently discussed, choosing the right answer relies entirely on user statistics. The following metrics were used to analyze the users who posted answer to posts in the forum:

- Reputation
- Views
- Up Votes
- Down Votes

This strategy allows viewers to select users having less down votes and more for reputation, up votes and views. A radar chart is plotted for each of the users which takes into consideration, the four-metrics discussed above. An illustration of such graph is shown below:

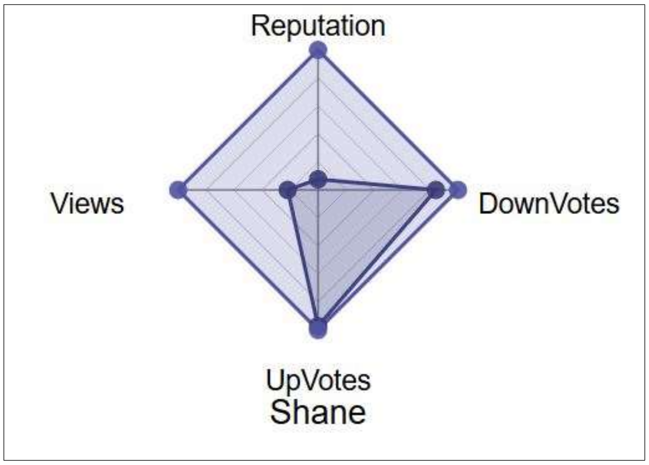


Figure 5: Radar graph showing user statistics

Authentic user statistic has plots lying on the upper left and bottom left quadrants. This means, the more the number of down votes the user has received, the less authentic his answers are. The metrics were evaluated on a scale of 0 - 5. 0 being least preferred and 5 being most preferred for each metric. This visualization helps user choose answer when scores are low/not evident.

The benefits of plotting a user spider chart is as follows:

- Easier to identify user strengths and weaknesses.
- Better when data consists of multiple measures and that require different quantitative measures.
- Objective of our user analysis is to evaluate the symmetry of user metrics rather than to evaluate their magnitude.

This way, one can easily identify answers posted by professional users and verify their authentic source. This methodology would greatly help users in question – answer forums. Figure 6 shows the comparison of four users who posted answers to the topics 'Function' and 'Arima'. The second answer received the most number of scores. This is analogous to the user statistic. Though the user doesn't have reputation, he has been keen on receiving the most number of up votes.

User analysis may not always prove the basis for choosing answers. It may also depend on the domain expertize of the user who is posting answer to inexperienced domain. That's why down votes play a major role in deciding answers to questions.

The sequence of visuals for analyzing the top 30 most frequent topics are as follows:

- Word cloud to display frequent topics. (Circle size corresponds to topic frequency).
- Word cloud highlighting 5 most related topics for each of the top 30 topics.
- Bar chart showing answers (X-axis) and Scores (Y-axis)
- Radar chart showing user statistics.

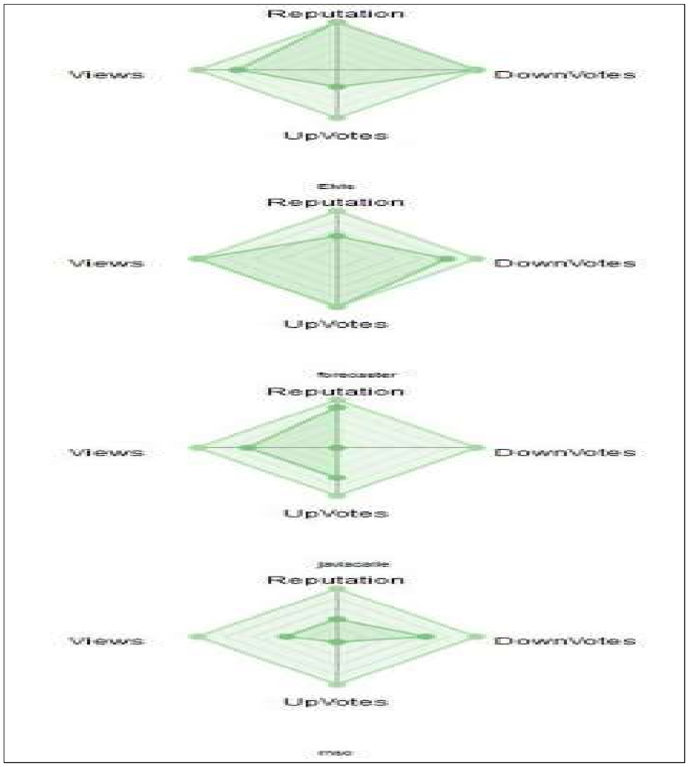


Figure 6: User Analysis

VI Data Processing

- **Posts Extraction:** In this project, we have queried cross validated database and retrieved 14,000 questions and their corresponding answers and user data.
- **Preprocessing:** We have collected the content of all the posts, removed all the stop words, performed stemming using snowball stemmer.

- **Topic and their Relations:** Term frequency is used to retrieve most frequently used topics. For each topic closely related topics are also retrieved based on the frequency with which it has occurred along with the respective topic.
- **Post Data:** Each post is plotted with respective to the selected topics by finding the cosine similarity with respective to each topic.
- **Answers Data:** All the answers data is collected for all the questions and those are plotted based on the score of the answer post which is done by considering up votes, down votes etc.
- **User Data:** For all the posts of type answer user data is collected and analyzed using radar chart in terms of reputation points, up votes, down votes and views for those answers.

VII Methodology

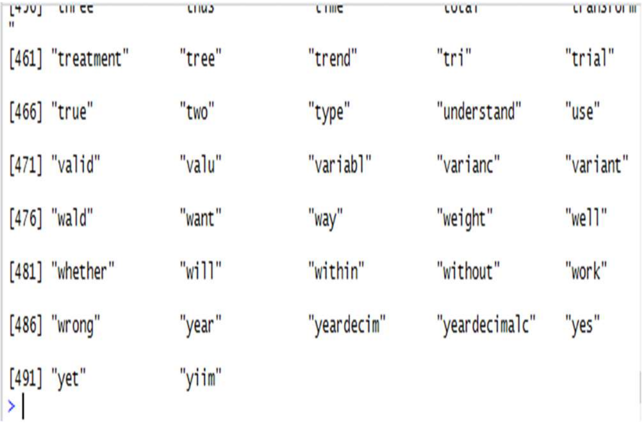
- **Calculating the most frequently used topic and their relations:**

Analyzing the content of the posts finds promising results in determining the constructive post data, latent semantic relationships of post content, user interactions and helps in identifying relevant information of programming community. In our project, we have collected the post content and cleaned the data by removing the stop words, performed stemming using Snowball algorithm and performed term frequency technique which helped in determining the most frequently used topics by the developers. For each most popular topics discussed on the website we have retrieved their relations with other topics. In this process we have determined how frequently each topic appears with other topics in the post content. In this process, we have also used similar methodology which is term frequency in order to calculated how frequently each topic appears with others.

- **Topic modeling**:

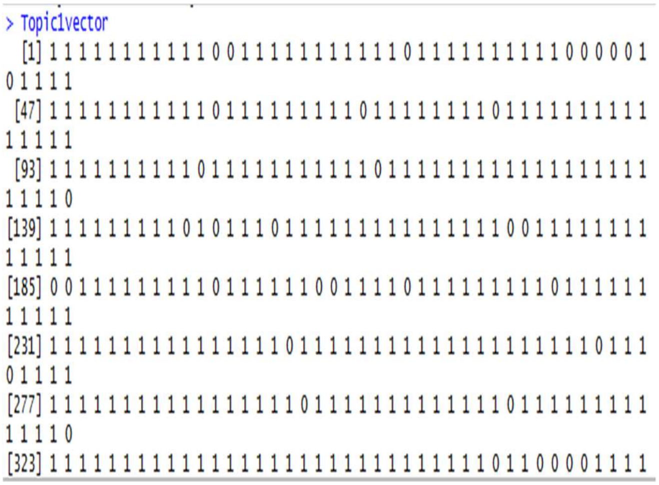
For the two topics selected by the user, graph is plotted with each topic on x and y axis. The post that is most related to both with lie in between and farther away. These posts are plotted based on the cosine similarity values calculated for each post with respective to each topic on x and y axis. For finding cosine similarity, word vector is constructed using common words that appear for both topics by analyzing the posts that has both topic keywords in the content. This word vector acts as basis for determining the binary vectors for each topic and the posts as the post vector and topic vectors should be of same length to determine the similarity. Binary vector for topic1 is determined by analyzing the content of the posts that has only topic1 keywords but not topic2 keywords in the content. Similarly, topic2 vector is determined by considering the posts that has only topic2 keywords but not topic1 keywords. After which, each post is considered as a document, performed data cleaning by removing the stop words, and binary vector is determined by finding the frequently repeated words and looping through the common word vector and determines the presence of each frequently repeated word. In this manner, binary vector for each topic and posts are determined after which cosine similarity value of each post with respective to both topics are obtained. These values acts as x and y coordinate to plot on the graph.

This is done by writing an R script, that performs cleaning of the data, finding the common word vectors, finding binary vector for each topic with respective to common word vector and calculating cosine similarity of each post.



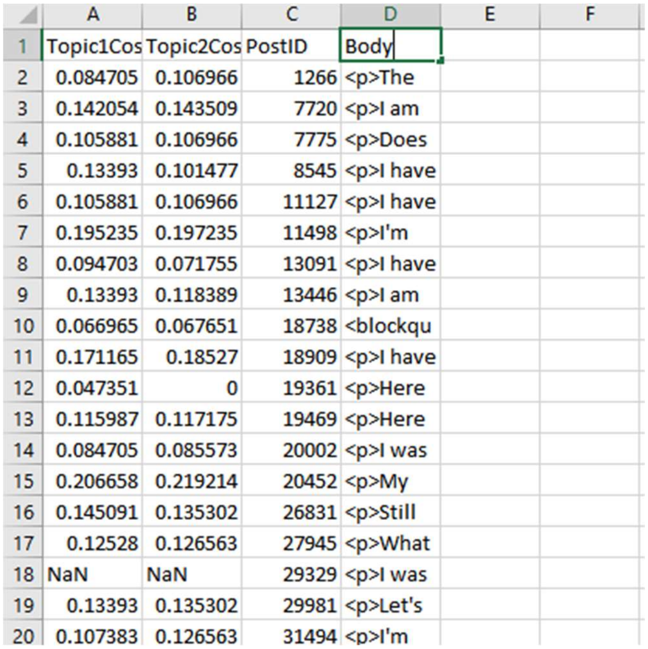
```
[461] "treatment"   "tree"        "trend"       "tri"          "trial"
[466] "true"        "two"         "type"        "understand"   "use"
[471] "valid"       "valu"        "variabl"     "varianc"      "variant"
[476] "wald"        "want"        "way"         "weight"       "well"
[481] "whether"     "will"        "within"      "without"      "work"
[486] "wrong"       "year"        "yeardecim"   "yeardecimalc" "yes"
[491] "yet"         "yiim"
>
```

Fig : Common word vector for given two topics



```
> Topic1vector
  [1] 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1
0 1 1 1 1
 [47] 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1
 [93] 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 0
[139] 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1
1 1 1 1 1
[185] 0 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1
1 1 1 1 1
[231] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1
0 1 1 1 1
[277] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1
1 1 1 1 0
[323] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1 1
```

Fig : Binary vector for each topic



	A	B	C	D	E	F
1	Topic1Cos	Topic2Cos	PostID	Body		
2	0.084705	0.106966	1266	<p>The		
3	0.142054	0.143509	7720	<p>I am		
4	0.105881	0.106966	7775	<p>Does		
5	0.13393	0.101477	8545	<p>I have		
6	0.105881	0.106966	11127	<p>I have		
7	0.195235	0.197235	11498	<p>I'm		
8	0.094703	0.071755	13091	<p>I have		
9	0.13393	0.118389	13446	<p>I am		
10	0.066965	0.067651	18738	<blockqu		
11	0.171165	0.18527	18909	<p>I have		
12	0.047351	0	19361	<p>Here		
13	0.115987	0.117175	19469	<p>Here		
14	0.084705	0.085573	20002	<p>I was		
15	0.206658	0.219214	20452	<p>My		
16	0.145091	0.135302	26831	<p>Still		
17	0.12528	0.126563	27945	<p>What		
18	NaN	NaN	29329	<p>I was		
19	0.13393	0.135302	29981	<p>Let's		
20	0.107383	0.126563	31494	<p>I'm		

Fig: Shows the cosine similarity of each post with respective to each topic.

Answer Analysis:

For all 14,109 questions, we have collected all the answers and projected the answers in the form of bar chart in which height of the bar determines the score count for each answer.

User Analysis:

We have also considered user data for the selected post. For the user to select the best answer possible, user analysis also plays a major role. User can select such an answer which is posted by the user who is good at that topics.

VIII Research Questions:

1. What are the most frequently discussed topics in Cross Validated in programming language R?

Present Q&A sites provides developers a platform to get their queries clarified. By finding the most frequently used topics shows the major areas of interest for developers. Based on this analysis developers can concentrate more on these topics to find the trends in the market. This also helps researches to find out the major changes in future.

2.What are the topics that are correlated with each other?

In this project, we have found topics that related to other topics in terms of questions and answers. This helps to find closely related topics which helps developers to know how each topic can lead answer related to another topic.

3.How to provide the best post that deals with the specific topic that user is interested?

This project provides most related post for given user selected topic. This is done by calculating similarity of each post with respective to each topic.

IX Results

- ➢ Identified significant posts through plots drawn based on cosine similarity between post.
- ➢ Probability values (p>0.5) for post-topic similarity shows valuable post.
- ➢ Identified constructive answers based on user-user comparison for a given post.
- ➢ Web plot shows which user contributions are important.

X CONCLUSION

In this paper, we have proposed a methodology to discover the most frequently used topics using term frequency technique on post content posted by millions of active users on cross validated website in R programming language. This analysis provides a brief idea of developer needs. Finding frequent topics also helps Cross Validated website to know on what topics the content is being generated and can help in moderating the flow of content by creating extra pages that deals with those topics.

Our methodology can be applied to other Q&A websites and trends can also be analyzed for more accurate information which can give developers clear idea on trending topics.

XI REFERENCES

[1] "Why, When, and What: Analyzing Stack Overflow Questions by Topic, Type, and Code" Miltiadis Allamanis, Charles Sutton School of Informatics, University of Edinburgh.

[2] "What are developers talking about? An analysis of topics and trends in Stack Overflow" Anton Barua ,Stephen W. Thomas, Ahmed E. Hassan

[3] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[4] http://www.cio.com.au/article/575209/5-tools-techniques-text-analytics/

[5] https://en.wikipedia.org/wiki/Stemming