Stack Overflow Java Data Visualization

Ankush Kanungo
Arizona State University
Tempe AZ
United States
akanungo1@asu.edu

Anto Oswin Nihal
Arizona State University
Tempe AZ
United States
antooswin@asu.edu

Malkiyat Singh
Arizona State University
Tempe AZ
United States
msingh55@asu.edu

Pranshu Varshney
Arizona State University
Tempe AZ
United States
pvarshn1@asu.edu

ABSTRACT

This report presents visualizations developed for the Stackoverflow dataset for the Java tag. The dataset used is for the year 2014 and we have analyzed it by doing some text analysis, Network analysis, Time series analysis [7] on weekly basis or monthly basis. A couple of visualizations were developed in order to display some useful insights or trends from the dataset. The whole visualization is hosted online and is accessible at the link *https://antooswindvproject.github.io/* and the demonstration of the visualization is uploaded to link https://youtu.be/OTZ17OMP2vM. We have added all the interactivity and coloring schemes based on some standard formats[3] to get better insights about the data like when to use similar shades, light shades or dark shades. At the end a classification model was also trained based on SVM classifier.

KEYWORDS AND PHRASES

Sunburst, Concept Map, SVM, D3 [8], JSON [5].

1 INTRODUCTION

Data Visualization is important to better analyze and understand the data. Human brain process the visual information in the form of graphs, charts more easily as compared to text data which helps business leaders, organizations or other individuals to take rapid actions. It is helpful in finding the correlation between business functions and market performance by introducing some machine learning techniques. The coming sections covers the data pre processing, data collection and visualization techniques used like Text Analysis using Bubble chart, Time Series analysis [7] using Bar graphs, Sunburst and Network analysis using Concept map. All the visualizations are designed interactive to best possible extent by choosing the proper coloring schemes for text, legends, graphs, axis etc.

1.1 Data Collection:

We have used dataset from following resources-

- Stack Exchange:
  (https://api.stackexchange.com/2.2/users/<userid>
  ?order=desc&sort=reputation&site=stackoverflow)
- (Sampled year 2014) of forum posts in topic Java from StackOverflow.com {215,968 questions, 183,477 answers, 95,617 accepted answers }

1.1 Data Analysis:

We performed Data Analysis as given below for different visualisations.:

1.1.1 Interactive Bubble Chart - Text Analysis:

To visualize the top packages discussed, the dataset was filtered to have following json [5] format and then visualize in the form of a bubble chart:

items: [

{text: "android", count: "99814"}.../ other packages */]*

1.1.2 Concept Map - Data Analysis:

Used json [5] format for Concept/ Network map. The data used contains all tags associated with respective user. This data is created using *Matlab* script which preprocessed data into required json [5] format. Data in following json [5] format is used for interactive visualization using D3 [8]:

[[userid:1,[associated tags]],......]

User Profile data is taken from stackoverflow API i.e.
(https://api.stackexchange.com/2.2/users/ <userid>
?order=desc&sort=reputation&site=stackoverflow)
you can specify a user id in the above link.

For Time series analysis [7] of a user's activity, all the stackoverflow data with features like *type, user_id, time* and *vote* was used. All the data is combined into one CSV file with Matlab script and then visualized in the form of time series [7] graph using D3 [8].

1.1.3 Sunburst - Data Analysis:

Sunburst used a hierarchical data [3]. We fed the javascript a json [5] file which had data in the following format

Day->DayOfWeek->Type->Tag.

The format for the json [5] *file is given below:*

```
{
  "name": "Day",
  "color": "<color value in Hex>",
  "percent": "",
  "children": [
   {
     "name": "Sunday",
     "color": "<color value in Hex>",
     "percent": "<percentage>",
     "children": [
      {
        "name": "Accepted-Answer",
        "color": "<color value in Hex>",
        "percent": "<percentage>",
        "size": <size>,
        "children": [
         {
           "name": "android",
           "color": "<color value in Hex>",
           "percent": "<percentage>",
           "size": <size>
         }, ... // for spring,swing,  ... Others.
       }... // for Question & Answers
    }... // for Monday ... Saturday
  }
```

1.1.4 Stacked Bar Graph:

The dataset for the year 2014 was filtered based on the top packages *Android, Spring, Swing, Hibernate, Maven, JSP, JSON* [5] and then used in JSON [5] format as shown below to have interactive stacked bar graph in D3 [8]:

```
var Jan = [
{
  "Package": "Android",
  "AcceptedAnswers": <count>,
  "Answers": <count>,
  "Questions": <count>
 } {.../* for each package */} ..../* for each month */ ]
```

1.1.5 SVM Classification:

Data was preprocessed to have post count for each day of the week throughout the year 2014. Days of the week were labelled as High or Low based on the assumption i.e. if the number of posts on that day is less than the median it is labelled as Low active day, else labelled as a High active day.

2 MOTIVATION

We wanted to gather insights from the stackoverflow data. We have considered the main two factors in our analysis the dataset for following insights/trends:

We wanted to check for user activity by leveraging the count of different type of queries. We can use the data to perform. Our other intention was to check for the ongoing trends in tern of the topics in discussions.

2.1 User Analysis

We identified top users and their activities on popular topics, which is helpful to have users with the maximum contributions and the topics on which they have expertise. Also to visualize the number of votes and variation in votes on monthly basis for the top users.This would give us an insight into the top users and their contribution. Stackoverflow has a lot of users and it is not effective to track most of them as most of them are not very active. So, we took the most active users in consideration which could help us in determining the current topics which are trending as the more number of answers are provided by these users.

2.2 Topic Analysis

Useful analysis to get trending topics, determining corresponding trending technologies to that topic. We can determine the direction in which discussions are heading or if a topic is getting obsolete or getting more popular with time. We did analysis of different topics as it could be used in determining the hot topic and where the industry is heading. We know that on stack overflow newer the topics and newer the technology, the more questions people have and this could help us in determining next technology trends. This information along with top users contributions can determine which is the technology adopted by most people in the industry and which technologies are emerging or trending.

2.3 Time Series Analysis [7]

Time series analysis to determine which topic is popular on a monthly basis, weekly basis and the user activities on those topics.This could help us in determining trends as well and how it has changed over time. We can see the topic getting changed and again determine the important technologies in the market. Time analysis are always useful and they are helpful in predicting the future for the technology with respect to time. We can also check the user activity using the time series graphs and data. We can point out certain trends and

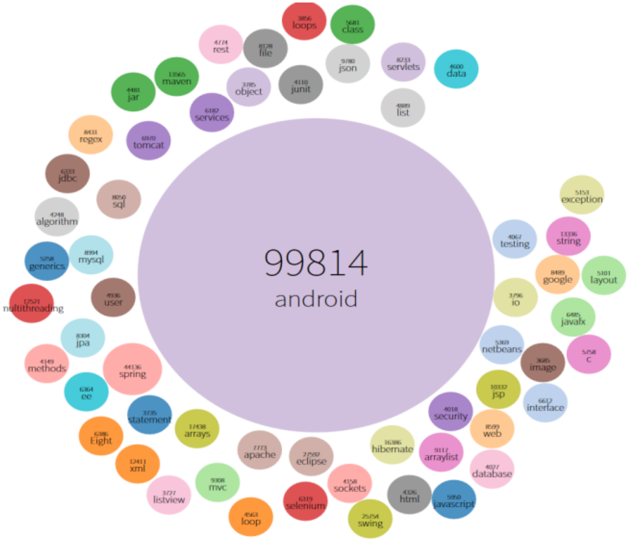


Fig 1- Bubble Chart to display the top topics discussed.

determine certain time which would be beneficial for posting questions or answers.

3 VISUALIZATION DESIGN (IMPLEMENTATION)

3.1 Interactive Bubble Chart - Text Analysis

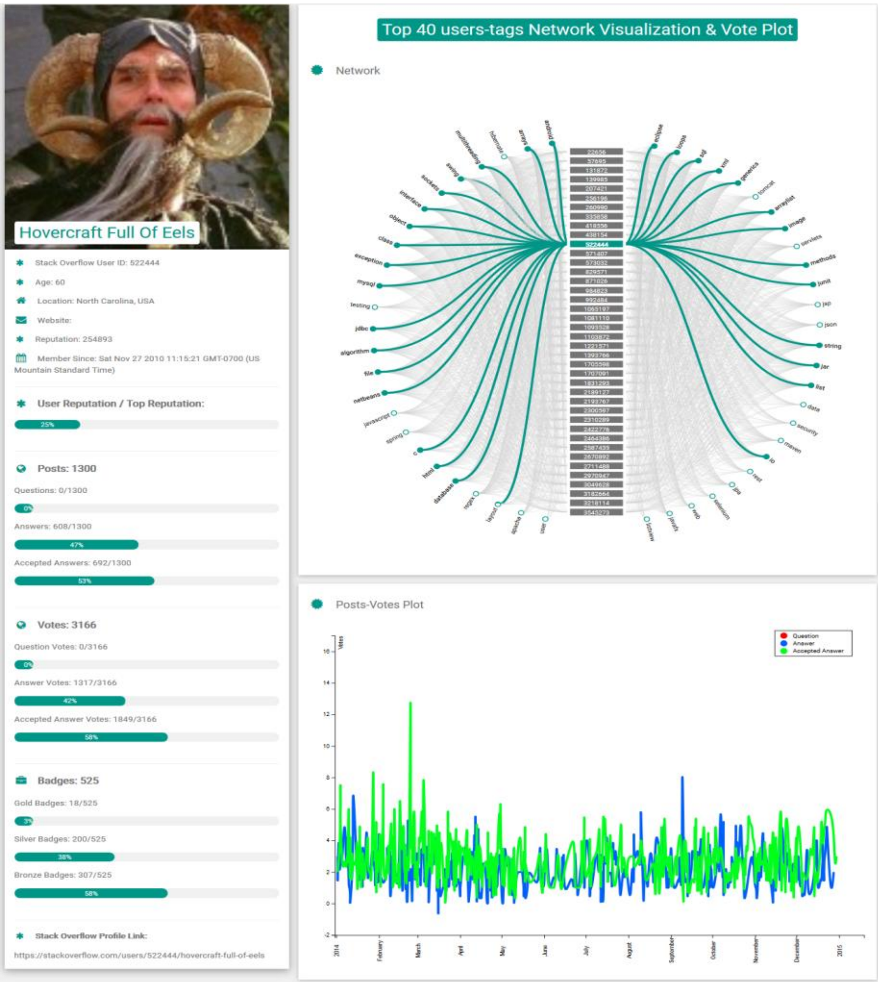
We started with the text analysis of *tags* column from the dataset to identify the top packages/topics discussed for the year 2014. Top 50 packages were filtered out and visualized in the form of interactive bubble chart using D3 [8], controlling the size, color of the bubble as mentioned in [2], clicking on a bubble will give the count i.e. number of times that topic was involved in the discussion for the year 2014. The size of the bubble is based on the number of times that topic was discussed. It can be seen in the Fig 1 that Android is the hottest topic followed by Spring, Swing, Hibernate etc.

3.2 Concept Map[10]

The relationship between top tags or topics like android, spring, sql etc and top 40 users were visualized. Top users were identified based on their activities in terms of number of posts they have on Stackoverflow for the year 2014. The idea is to see how top tags on whole dataset are related to top users. We can see top users are associated with top topics most of the time. Moreover, there is an option to see user's profile once you click on a specific user id. User's profile is displayed using StackOverflow API[11] which shows the location, activity of that user. Time series analysis [7] on monthly basis is also designed for the votes received in three types of activities i.e. Questions, Answers and Accepted-answer for the year 2014. The reason for Time series analysis [7] is to determine which topic is popular on a monthly basis and how user activities/votes varies over time.

Fig 2 shows the one example of the visuzaltion after a user id is clicked. Following are the observations:

1. The network graph[10] in fig 2 shows one user id is associated with 60% of total top topics. Moreover, after you hover on user id, the associated tag connections/links and tags gets highlighted. The unassociated or inactive links/connections are displayed with light grey color and is not highlighted.
2. The left panel of the visualization shows the user's profile from stackoverflow API[11]. It shows user's profile picture, user's basic information like age, location, website, reputation, joining date, badges information. The user's reputation is used to rank the user considering all user's reputation in the dataset. The Posts section shows number and percentage of activity in three categories Questions, Answers and Accepted-answer for the selected user. Similarly, Votes section shows the number of votes and vote percentage in three types of activities received by the selected user.
3. The Time series analysis[7] at the bottom shows the user journey over time in three categories i.e Questions, Answers and Accepted-answer represented by Red, Blue, Green color respectively. We can see from the below multi-series line graph[9] that the user is active whole year and gets votes on its answers and accepted answers with some high votes in accepted answers. Also it was observed that a highly reputed user generally asks very less or no questions and more involved in answering the questions posted by others as shown in Fig 2 the selected user has asked 0 questions but answered more than 600 questions on the stackoverflow. These insights could be helpful for the fact that if someone asks questions on stackoverflow, than with which answer he should proceed based on the accepted answer percentage of the person answering his question.



Hovercraft Full Of Eels

✱ Stack Overflow User ID: 522444

✱ Age: 60

Location: North Carolina, USA

Website:

✱ Reputation: 254893

Member Since: Sat Nov 27 2010 11:15:21 GMT-0700 (US Mountain Standard Time)

✱ User Reputation / Top Reputation:

25%

Posts: 1300

Questions: 0/1300

0%

Answers: 608/1300

47%

Accepted Answers: 692/1300

53%

Votes: 3166

Question Votes: 0/3166

0%

Answer Votes: 1317/3166

42%

Accepted Answer Votes: 1849/3166

58%

Badges: 525

Gold Badges: 18/525

3%

Silver Badges: 200/525

38%

Bronze Badges: 307/525

58%

✱ Stack Overflow Profile Link:

https://stackoverflow.com/users/522444/hovercraft-full-of-eels

Top 40 users-tags Network Visualization & Vote Plot

Fig 2- User Profile and User's association with top tags is displayed.

3.3 Sunburst [4]

The idea for using the sunburst [4] was to show the hierarchical relation [3] among different fields. In this case, we wanted to show the relationship between the day of the week, the count of 'type' of query (questions , answers and accepted answer) for each day of the week and the topics covered (like android, spring, swing etc.) for each 'type' of query. Sunbursts [4] are one of the basic tools to cover this kind of data but we used an interactive modified version of sunburst to show our analysis [6]. The sunburst [4] has three different levels:

1. *Day of Week*: The Day of the week formed the first level of hierarchy [3] for the sunburst. The day were named from *Sunday - Saturday* in the order of their count i.e. the fields were in order the of the count of entries on that day of week.
2. *Type of Entry*: The type of entry specified for each day of the week, the count of the type of entry that would be the count of number of question, answers or accepted answer for that day of week.
3. Count of each popular topic for that entry: The count of topic will let us know the percentage of the entries belonging to a specific topic. In our case we have taken the seven major topics separately and clubbed the rest as others. The seven most popular topics used here are android, swing, spring, hibernate, json, maven and jsp.

In our version of sunburst [4] we can click on any area and it would show us its children and adjust the percentages accordingly. So, in case we click on say Thursday, we get the sunburst transformed into a newer version of sunburst which would contains the Question, Answers and accepted answers distribution for thursday along with the distribution of popular topic for each type of query. Clicking on the circle in the centre will bring up it's (Thursday's) parent view [6],

we can hover over each slice which will give us the relative percentage of that slice for that view. Suppose in the first view with all days included, we can have the percentage of each field with respect to the size of the slice it takes but, i f we click on any day say thursday, we can see an increase in the percentage for that slice as now it will be shown relative to the new view and where the slice fits in the new view.

Also, While labelling we had to make sure that all the labels for the slices of the sunburst [4] with small slice size won't be superimposed. So, for each view we ensured that if the slice size is small, we don't label and the same slice gets labelled if we click on any of its parent resulting in a different view with the label visible if the slice size has increased for that particular field. We can zoom in the sunburst while clicking on a slice and zoom out by clicking in the centre of the sunburst. The colors in the sunburst have been set using the Martijn Tennekes and Edwin de Jonge 's paper based on the tree coloring for hierarchical data visualisations titled "Tree Colors: Color Schemes for Tree-Structured" [16] . We have used hierarchical colors such that the children have the same color in a different shade and all the related nodes have similar colors. This behavior might be a bit off in some case but it is only because the colors shades required were far greater and finding unique (24x7 + 3x7 + 7) shades. It was a challenge getting all the different colors while keeping similar shades for related nodes.

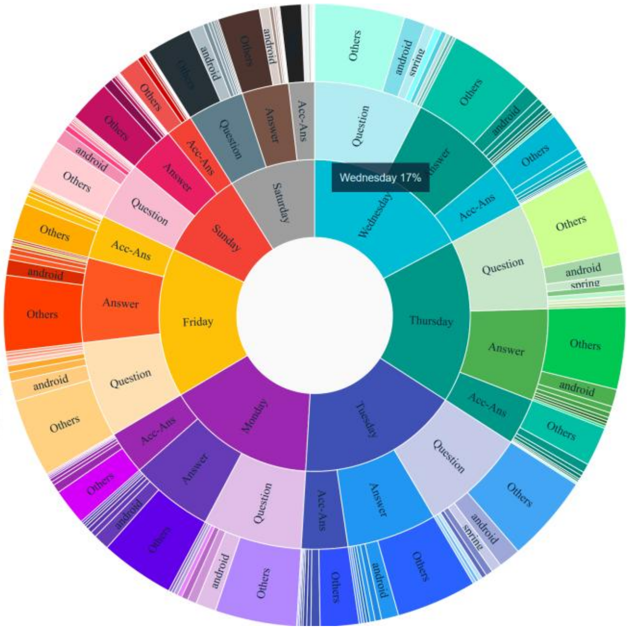


Fig 3- Sunburst - Highest level - All days of week.

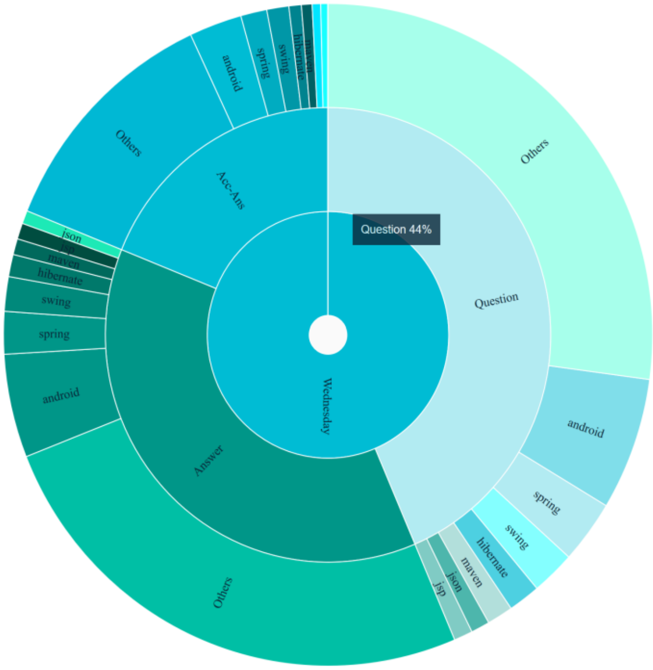


Fig 4- Sunburst - A specific day - Clicked Wednesday.

3.4 Stacked Bar Graph

The top topics are analyzed on monthly basis as well i.e. activity corresponding to each popular topic is visualized in the form of *stacked bar graph* choosing color, size as mentioned in [1], where top part of the stack shows the number of questions asked for a topic in the selected month, middle part of the stack shows number of answers and bottom part shows the accepted answers. A dropdown menu is provided to choose a specific month and analyze the trend for that month. It was observed that the activity for most of the topics is high at the beginning of the year or say first quarter of the year as compared to the end of the year. Most of the topics follow the same trend except for few like *Swing* which gets more active than Spring for some months like in March, April, although in text analysis it was observed that Swing was next to Spring. After analyzing the trend it seems that people asking questions or answering questions are less active towards the end of the year which might be because of the holiday season towards the end of the year. Fig 5 shows the activity for month of February selected from the dropdown menu. Hovering over a particular slice of the stacked bar graph should give the count for that package for the selected month (i.e. number questions or number of answers or number of accepted answers).

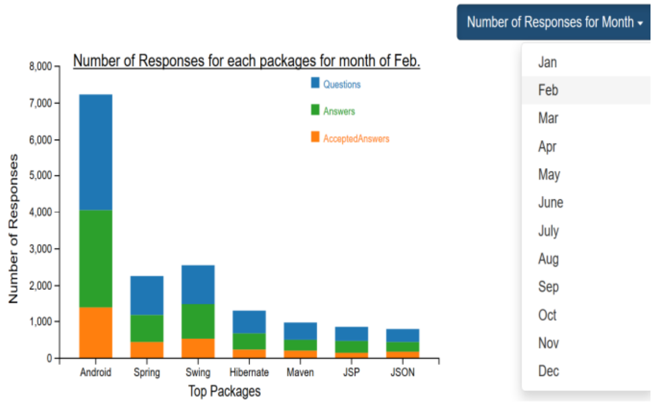


Fig 5- Stacked bar graph for monthly analysis of top topics

3.5 SVM Classification model

We trained a classification model in order to classify a day of the week as a *High* active or *Low* active day based on the number of posts on that day. The data was preprocessed based on the number of posts for each day of the week and any day having posts above the median (number of posts) is labelled as High and below the median is labelled as Low. The SVM classification model with polynomial kernel (degree=2) is trained on the preprocessed data in R. The trained model is plotted against the data points as shown in Fig 6. The symbol 'x' in the Fig 6 shows the support vectors for the decision boundary and 'o' are the data points. Numbers 0 - 6 represent the days of week from Sunday (0) - Saturday (6). It could be observed from the Fig 6 that most of the data points for day 0 (Sunday) or day 6 (Saturday) are on the Low side of the decision boundary i.e. they are less active days as compared to the other days like mid of the week. The model takes day of the week, number of posts on that day as inputs and classify that day as High Active or Low active day.

R libraries like *Shiny, ggplot2, taRifx* are used to train the model and plot the data points against the model. The R script along with the dataset created for training the model are also submitted along with the final submission.

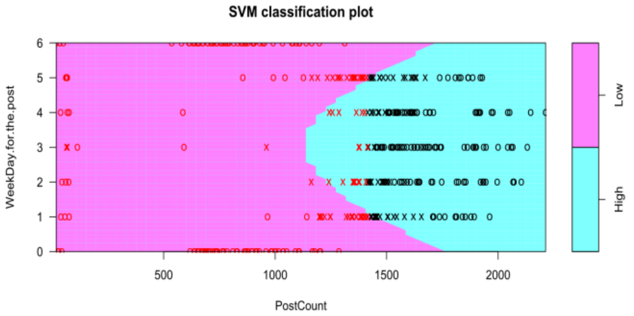


Fig 6- SVM classification model for day of the week and number of posts.

4 METHODOLOGY

- We have used the stacked bar chart to show the distribution of different type of activity for each topic for each month of the year. we have aggregated the data for a topic for each month and then added them in the stacked bar chart.
- We used *concept map* for identifying the relation among users and the topics of their expertise in the form Network Visualization. The bars in between would be for the top users and it showed the links to the popular topics.
- We used *sunburst* to visualize hierarchical relationship among day of week, type of activity for each day of week and popular topics frequency for each type of activity. We arranged the hierarchical data in the different levels of the sunburst.It was a zoomable sunburst [4] .
- We used tree color scheme [16] to determine coloring for parent and children in the sunburst visualization. This scheme was used to determine the children for each node of the sunburst.
- We used stacked bar chart for analysis of type of activities for popular topics for every month in 2014 and visualized each activity in different color.
- We used SVM classification model to classify the day of the week as high or low active day based on number of posts on that day.

5 EVALUATION PLAN

As discussed above, our major focus was analysis of activity by users and activity over time along with the topics/packages discussed in that activity. We had all our visualisation focus on either one or both of these factors. After doing all the analysis of the data and designing the visualizations discussed in past sections helped to get insights/trends like:

- Interactive Bubble Chart: In the interactive bubble chart we found out the android was the the most popular topic as it has the biggest sized bubble. The size of the bubble was according to popularity so the output was straightforward.

- Concept Map: In the concept map, we found out that the ratio of answers and accepted answers was maximum for top users and these were the users with the highest reputation. We also noticed that the activity of the users kept on decreasing as the year went on. So, they had most activity towards the start of the year and least towards the end. We also noticed that if a top user has joined in the middle of the year, they have the maximum activity at the time of joining and keeps on decreasing as the year goes by.
- Sunburst: In the sunburst we came to the conclusion that the mid of the week is most active as opposed to weekends which are least active days.
- Stacked bar graph: We have noticed that the activity (number of posts) is minimum towards the end of the year as compared to the first quarter of the year. This behavior could be justified by the fact that there are holidays at the end of the year and people are not that active on Stack overflow.
- SVM Classification: Similar to Sunburst, we noticed from the SVM Classification model that mid week days are classified as High active day using SVM classification model as well verifying our belief.

5 DISCUSSIONS & FUTURE WORK

In the future, we can do a couple of enhancements to our existing systems.

We can have the visualisation automatically answers some questions.This implies that we can look for certain trends and if they occur we can have the visualisation point it out. For instance, for exceptional value of certain fields we can have a visualisation pop up.

Another change that we can make to the visualisation is that we can make certain adjustments to each visualization such that all the different visualisations are related and its is easy to transition from one visualisation to another. We can have the visualisation present a story rather than different visualisations in different silos. For instance, we can have the Text analysis point to the time series analysis [7] and user analysis presented using concept map

REFERENCES

[1] https://bl.ocks.org/mbostock/3886208

[2] https://bl.ocks.org/mbostock/4063269

[3] A New Way to Visualize Decision Trees - https://blog.bigml.com/2013/04/19/a-new-way-to-visualize-decision-trees/ - Last Accessed April 26th ‘2018

[4] Zoomable Sunburst w/ rotated labels http://bl.ocks.org/kaz-a/5c26993b5ee7096c8613e0a77bdd972b - Last Accessed April 26th ‘2018

[5] JSON - Introduction https://www.w3schools.com/js/js_json_intro.asp - Last Accessed April 26th ‘2018

[6] John Stasko, Richard Catrambone, Mark Guzdial and
Kevin McDonald, Georgia Institute of Technology, Atlanta, GA
An evaluation of space-filling information visualizations
for depicting hierarchical structures. (Accepted 31 May 2000)

[7] Muhammad Adnan, Mike Just, Lynne Baillie, University of Leeds, Heriot-Watt University - Investigating time series visualisations to improve the user experience - CHI'16, May 07-12, 2016, San Jose, CA, USA

[8] D3 - https://github.com/d3 - Last Accessed April 26th ‘2018

[9] https://bl.ocks.org/mbostock/3884955

[10] http://bl.ocks.org/virtuald/ea7438cb8c6913196d8e

[11] (https://api.stackexchange.com/2.2/users/ <userid> ?order=desc&sort=reputation&site=stackoverflow)

[12] https://www.w3schools.com/w3css/4/w3.css

[13]https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.min.css

[14] https://www.w3schools.com/lib/w3-theme-black.css

[15] https://rischanlab.github.io/SVM.html

[16] Martijn Tennekes and Edwin de Jonge - Tree Colors: Color Schemes for Tree-Structured Data - https://pdfs.semanticscholar.org/6f4e/96b5a487b556cccffc5f9e6b246bbbb33d63.pdf - Last Accessed April 26th ‘2018