Stack Overflow Java Data Visualization

Anto Oswin Nihal
1213186167
CSE 578 Data Visualization
antooswin@asu.edu
Arizona State University

Abstract

Stack Overflow is the place to go for programmers when they run into trouble. It is backboned by a community of experienced, intermediate and novice programmers who either post questions or answer questions that have already been posted. For this project, we chose to collect the data pertaining to all such posts which relate to the programming language – JAVA. By analyzing and visualizing this dataset, we present a platform that provides insights from the data through visualizations techniques. Data visualization plays a key role in helping to analyze and understand data. The human brain is known to process and pay attention to visual information more than to text information. Because of this reason, visuals can help us see patterns and connections in data much more efficiently by hacking into the attention span of our brain. In the forthcoming sections, we will cover how we processed the data, and visualized it to perform text analysis using bubble chart, time series analysis using bar graphs, sunburst and network analysis using a concept map. We designed the visualizations to be as interactive as possible by and did our best to follow the proper coloring, sizing and positioning schemes for text, legends, graphs, axis and all other elements.

1 OVERVIEW

1.1 Introduction

This project was aimed to gain insights from the Stack Overflow JAVA data. JAVA is one of the most popular topics on Stack Overflow and analyzing this subset will give us an idea about the whole dataset. Our main motivations were to:

- Perform User Analysis: We took the top users of the dataset and analyzed the trend in the number and quality of the posts posted throughout the year.
- Perform Topic Analysis: Similar to the previous one, we analyzed the top topics that were mentioned in the posts of the dataset.
- Perform Time Series Analysis: This was done so as to provide answers like what would be the best time to ask a question? At what time Stack Overflow is actually Stack Overflowed?
- Generate Resume: We gathered the Meta data of the user from the Stack Overflow API and generated the resume of the user based on the user's activity on the website.

1.2 Implementation

1.2.1 Data Analysis

We performed Data Analysis as given below for different visualizations:

1.2.1.1 Interactive Bubble Chart - Text Analysis: To visualize the top packages discussed, we filtered the data and fit it in json in the following format:
items: [{text: "topic", count: "#number"}

1.2.1.2 Concept Map - Data Analysis:
We converted our data into the following format for Concept/ Network map:.
[[userid:1,[associated tags]],......]
The data we used contains all tags associated with the respective user. We converted it by writing a MATLAB script to automate it.

1.2.1.3 Resume Visualization: We refered the user profile data from Stack Overflow API at :
https://api.stackexchange.com/2.2/users/<userid>?order=desc&sort=reputation&site=stackoverflow

1.2.1.4 Time Series Analysis: We all rows of the Stack Overflow data and combined it into one CSV file using a MATLAB script.

1.2.1.5 Sunburst - Data Analysis: We converted the data in the following format:
Day->Day_Of_Week->Type->Tag using a Python script.

1.2.1.6 Stacked Bar Graph: We filtered data based on the top 50 packages used and then converted into JSON using a Python script.
Month = [{ "Package": "Android", "AcceptedAnswers": <count>, "Answers": <count>, "Questions": <count> }

1.2.1.7 SVM Classification:

We preprocessed the data to have post count for each day of the week throughout the year 2014. We labeled days of the week as High or Low based on the premise that if number of posts on that day is less than the median it is labeled as Low active day, else labeled as a High active day.

2 VISUALIZATION DESIGN (IMPLEMENTATION)

2.1 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 controlling the size, color of the bubble, 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.

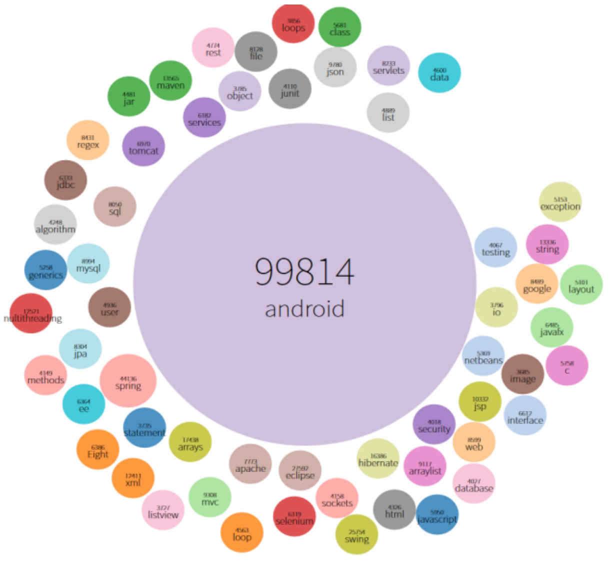


Fig 1- Bubble Chart to display the top topics discussed.

2.2 Concept Map

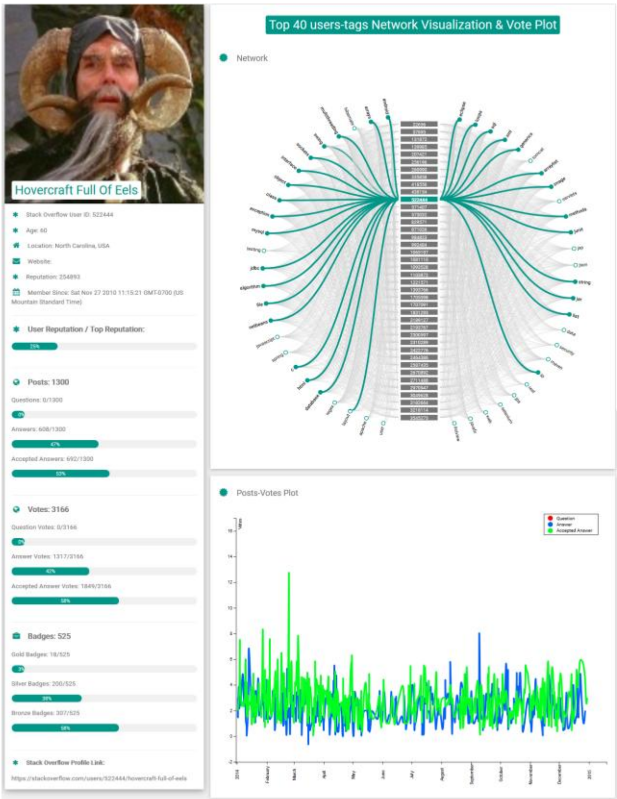
The relationship between top 52 tags and top 40 users were visualized. Top users were identified based on their activities in terms of number of posts they posted on Stack Overflow during the year 2014. The idea is to see how top tags on whole dataset are related to top users. We can see top users are associated with top topics most of the time.

2.3 Resume Visualization

In the same page as the concept map, there is an option to see user's profile once you click on a specific user id. User's profile is fetched using the Stack Overflow API and is displayed so that viewers can judge a user based on the user's Stack Overflow statistics.

2.4 Time series analysis

We designed a visualization to analyze the activity and productivity of a user by plotting the user's plot vs votes graph on monthly. The votes received can be for three types of activities i.e. Questions, Answers and Accepted- The intuition behind time series analysis is to determine which topic is popular on a monthly basis and how user activities/votes varies over time.


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

Fig 2 - Concept Map(top right), Resume (left), Time Series Analysis (Bottom Right)

2.5 Sunburst Data Analysis

The intuition behind using the sunburst is to show the hierarchical relation between different features of the data. In our case, we wanted to show the relationship between the day of the week, the count of each 'type' of post (questions , answers and accepted answer) for each day of the week and the topics covered (e.g. android, spring, swing etc.) for each 'type' of post. Our sunburst visualization has three different levels:

Day of Week: Monday - Sunday

Type of Entry: Question/Answer/Accepted Answer

Topics: All top topics

Along with the above, the hierarchical count of the topics is also displayed under the heading. The count of topic will let us know the percentage of the entries belonging to a specific topic. All the headings are clickable and will zoom in for easy viewing. We also used hierarchical colors for easy differentiation.

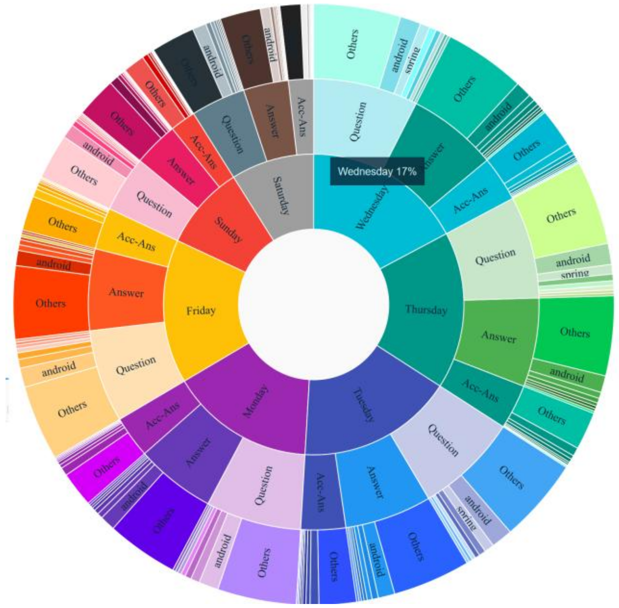


Fig 3 - Sunburst with all the tags

2.6 Stacked Bar Graph

We analyzed the top topics on monthly basis by plotting a stacked bar graph choosing color, size as based on the frequency of the occurrence of topics. After analyzing the trend we found out that people asking questions or answering questions are less active towards the end of the year which might be due to the holiday season towards the end of the year. 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).

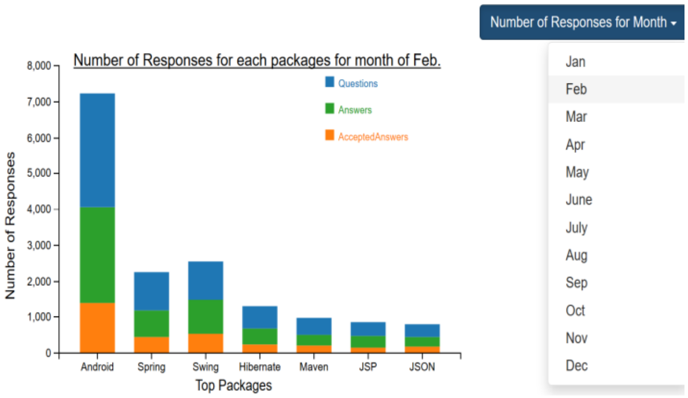


Fig 4- Stacked bar graph for monthly analysis of top topics

2.7 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 labeled as High and below the median is labeled 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 5 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 5 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. R libraries like Shiny, ggplot2, taRifx were used to train the model and plot the data points against the model.

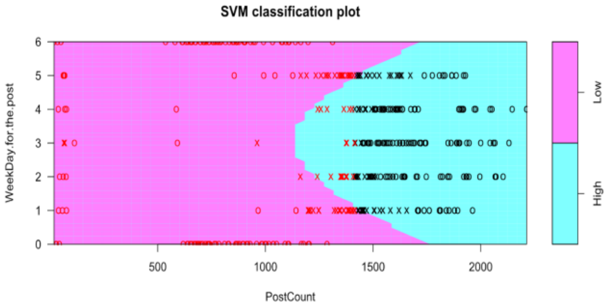


Fig 6- SVM classification model for day of the week and number of posts.

3 METHODOLOGY

- We used the stacked bar chart to show the distribution of different type of activity for each topic for each month of the year. We 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 shows the links to the popular topics.
- We used a zoomable 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 used tree color scheme to determine coloring for parent and children in the sunburst visualization.
- 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 a progress bar visualization to provide a comparison for viewers to judge Stack Overflow users on the Resume Visualization page.
- 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.

4 CONCLUSION

Our major focus was the analysis of activity by users and activity over time along with the topics/packages discussed in that activity. We had all our visualization focus on either one or both of these factors. After analyzing the data and designing the visualizations discussed in past sections we got insights/trends like:

- The top topics are always discussed by the top users frequently
- The ratio of answers and accepted answers was maximum for top users and these were the users with the highest reputation.
- The activity of the users keeps on decreasing as the year proceeds
- If a user has joined in the middle of the year, the user has maximum activity at the time of joining.
- The middle of the week is the most active period as opposed to weekends which are least active days.
- Activity (number of posts) is minimum towards the end of the
- Top users of Stack Overflow need not be having a sparkling resume!

5 PERSONAL CONTRIBUTION

This project was a team effort and would not be possible without the contribution of each and every member. In the initial phase, I was involved in doing research and picking the best visualizations for our ideas. I wrote the MATLAB and R scripts to prepare the data. During the implementation phase, I was involved in the main idea behind the resume visualization, the concept map and the time series analysis. Finally, I was instrumental in integrating the website together, putting together a template for the website, hosting the website and recording the demo.

6 LEARNING

I learnt to apply the concepts learnt in class into the real world. I learnt how to handle huge data and how to select and discard useful/ non-useful data. The project also enhanced my creativity skills. Learning D3 has helped me understand how to visualize data. My R, MATLAB and Javascript skills were tested to the full as well. At multiple times, I was forced into thinking ahead and to plan, design and choose accordingly. It was a great learning experience all in all!

7 TEAM

- Anto Oswin Nihal
- Malkiyat Singh
- Pranshu Varshney
- Ankush Kanungo

8 ACKNOWLEDGEMENTS

This project wouldn’t have been possible without the guidance and help of several individuals who in one way or another contributed their valuable assistance in preparation and completion of this study. I would like to convey my utmost gratitude to Dr. Sharon Hsiao, whose sincerity and encouragement helped us to complete this work.

9 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