Topic modeling and Sentiment analysis on User posts

Spoorthi Karnati
Arizona State University
Tempe AZ
United States
skarnat1@asu.edu

Manasa Pola
Arizona State University
Tempe AZ
United States
mpola@asu.edu

Pooja Kosala
Arizona State University
Tempe AZ
United States
pkosala@asu.edu

Shuddhatam Jain
Arizona State University
Tempe AZ
United States
shuddhatm@asu.edu

ABSTRACT

The aim of this project is to analyze and visualize Stack Overflow data and represent them interactively. This report describes the visualizations we developed to analyze Stackoverflow posts. We developed a range of charts like time series, dual axis bar charts, word cloud, scatter plots. We applied sentiment analysis to visualize the positive and negative posts which helps in understanding what is the sentiment of user's towards Java over time during the year 2018. We applied topic modelling on the posts to identify hidden topics in the posts and analyzed each topic separately. Topic modeling helps in identifying hidden clusters in the posts and keywords in the clusters which can be used to auto-generate tags for the posts in a group. Taking into account the major technologies, we analyzed posts from 2013 to create a network of technologies used together.

The whole visualization is hosted online and is accessible at the link https://shuddh.github.io/, Project source code is available at https://github.com/Shuddh/Shuddh.github.io and the demonstration of the analysis is uploaded https://youtu.be/lwCMKfzCr_M. We added all the interactivity and color coding based on standard formats[1] for better understanding and visual appeal.

KEYWORDS AND PHRASES

Sentiment Analysis, Topic Modeling, D3, JSON, Clustering Chart, Principal Component Analysis, Vader Sentiment Analyzer.

1 INTRODUCTION

Data Visualization is the presentation of data in a graphical format. It involves analysing, processing data and using information graphics , useful patterns and trends are visualized. Data Visualization is the best way to communicate insights from data through visual representation. According to the World Economic Forum, the world produces 2.5 quintillion bytes of data every day, and 90% of all data has been created in the last two years. It becomes increasingly difficult to manage and find distinct patterns or observations from such a huge data. Hence visualizations has become important these days.

We used several visualizations to portray few patterns in StackOverflow data. They are Text Modeling using Word Cloud, Stacked Bar Chart, Sentiment Analysis of user posts using multi-labelled Bar Chart and Line and Network analysis using Network chart developed using Gephi tool. All the visualizations are designed interactive to the best possible extent by choosing the proper coloring schemes for text, legends, graphs, axis etc.

2 MOTIVATION

We chose stackoverflow dataset to understand the behavior of programmers. Stackoverflow is a public website where programmers and developers present their technical challenges in the form of posts. These posts could be related to any technology. The three areas of study are: understanding the user sentiment while posting questions, answering or commenting on stackoverflow, finding hidden topics in a group of stackoverflow posts, to study how effective topic modelling is in clustering posts together, see whether topic modelling can be used as a method to automatically identify tags for a user query. The motivation behind network analysis was to have a deeper knowledge on what technologies or languages user use in their project

together, helping new programmers to choose any technology in their future projects.

2.1 User Sentiment Analysis

We identified that JAVA has been the primary language of choices for masses and decided to explore the sentiment of java programmers. This helps in tracking the sentiment of users towards particular technology and notify the technology creators to help the programmers. This model can be easily extended for all the technologies. Further identifying sentiment of a user towards a technology can help identify the experts and the beginners and better serve the users of the community. For example, if many users complain about a feature is say sequelize.js then stackoverflow can recommend sequelize.js to add more documentation to the feature. With this motivation to understand user's emotion over time we sought to implement the dual axis bar and line chart to plot positive and negative sentiment over time.

2.2 Topic Analysis

We learnt about Topic Modelling in class and wanted to apply in the context of stack overflow as each post has more than one topic associated with it. Although tags can be used to group posts together into a cluster, they are very high level groups. But programmers face similar issues and grouping posts together based on similarity has many use cases like if posts are grouped based on the scenario it describes. For example, RabbitMQ and kafka are both messaging queue services, the same query which is answered in RabbitMQ could be of reference when dealing with same query in Kafka. Topic modelling can identify such hidden topics and cluster posts irrespective of their tags. With these two motivations we performed extensive topic modelling on stackoverflow posts and performed detailed analysis of each topic which are described in section 3.1

2.3 Network Analysis

Network analysis is one of the major analysis done with big data. Clusters of top technologies were configured and visualized. The analysis was done on the top programming languages in the year 2019. We analysed 250,000 questions from the year 2013-2019 and picked the tags attached to the post. The motivation behind network graph was to analyse how strongly and weakly two nodes are connected implying the use of two languages together. This tells us that these languages are preferred by users, the more the use of languages/technologies together, the more compatible they are. This visual can also help other users decide which pair to choose. For example Mongodb and DynmoDB both are noSQL databases, but Python has more support for Mongodb and thus more users have prefered Mongodb over DynamoDB.

3 METHODOLOGY (IMPLEMENTATION)

This section describes each of the visualization, data set used in the visualization, the approach and list of technologies and packages used to develop the chart and an overview of how to interpret the chart.

3.1 Data Collection:

We have collected dataset from the following resources-

- Stack Exchange (https://api.stackexchange.com/2.2/users/<userid>?order=desc&sort=reputation&site=stackoverflow)
- https://data.stackexchange.com/stackoverflow/queries
- https://data.stackexchange.com/stackoverflow/query/1033618/post-for-tag-from-year-2013
- https://cloud.google.com/bigquery/public-data/

3.2 PRE-PROCESSING

Posts downloaded from above sources were in XML format. We extracted the title, description, comments and answers from the XML. Title and description were concatenated. We performed text preprocessing like stop word removal, lemmatization and stemming to clean and remove unnecessary words before applying sentiment analysis and topic modelling. We used NLTK package to perform preprocessing.

Charts:

1.1.1 Topic Modeling:

Large amounts of text data is generated and collected online everyday. As more information becomes available, it becomes difficult to search for the required text. There is a need to organize, search and understand the vast textual data. Topic modelling provides capabilities to organize, understand and summarize large collections of textual data. It helps in discovering hidden topical patterns within the text. Annotating documents according to the topics helps in organizing the content. Annotated information can be used as metadata for indexing and faster retrieval.

Topic modelling finds a groups of words from a collections of documents. Each group serves as topic. Here we used each post's title and description to describe a document. Topic modelling can be obtained using a variety of techniques like Latent Semantic Analysis(LSA), Probabilistic Latent Semantic Analysis(pLSA), Latent Dirichlet Allocation(LDA), LDA in Deep Learning (lda2vec). lda2vec is the most accurate topic modelling technique as it learns the representation of words, documents and topic vectors but requires a lot of training resources.

We implemented LDA based topic modelling using gensim package provided in python. In LDA model each document is viewed as a mixture of topics. Each word in the document is said to attribute to the document's topics. LDA model discovers different topics and how much of each topic is present in the document. We specifically chose LDA model because each stackoverflow posts can be tagged by more than one tags. For example, a query about a table design can be tagged as "Database modelling, MSSQL". Hence, LDA is ideal for this dataset as it can be used to classify documents in more than one topic.

We identified around 18000 questions posted on stackoverflow from 2008 to 2018 on following 15 technologies: *Matplotlib, PowerBI, Tableau, Selenium, json, elixir, apache, chartjs,talend, keras, sequelize.js, ubuntu, neo4j, joomla, tomcat.*

LDA topic modelling could identify 10 topics from above texts: *Matplotlib, PowerBI, Tableau, Selenium, json, elixir, apache, chartjs,talend, keras, sequelize.js, ubuntu, neo4j, joomla, tomcat.*

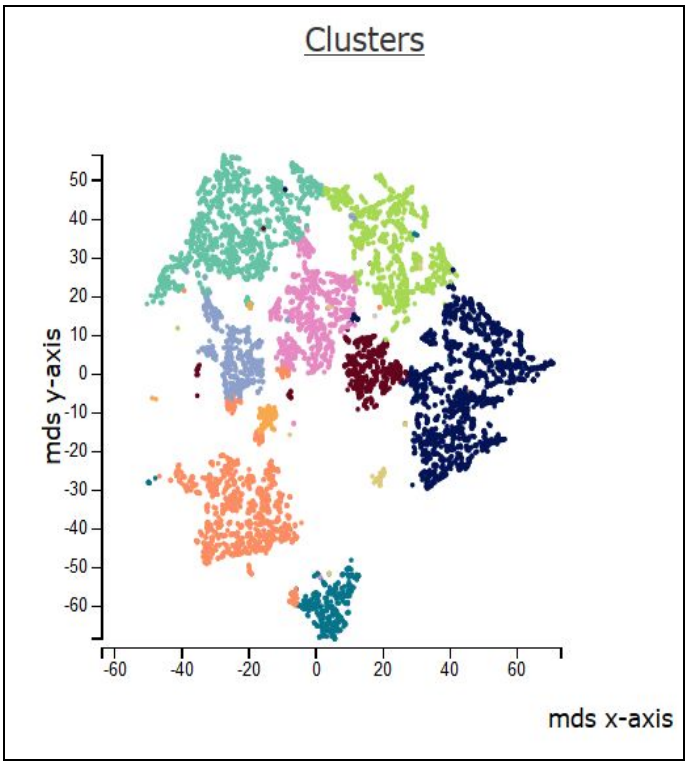
The other hidden topics LDA could identify from the posts are roughly categorized as : objects, errors, stings, tables. We believe these topics we learnt be LDA as these terms are frequently used in almost every stackoverflow post.

We developed 3 charts to visualize topic modelling of 10 topics discovered by LDA: 1) A scatter plot to visualize the clusters in a two dimensional posts. 2) A word cloud to visualize the importance words used in each topic 3) A dual axis bar chart to show the count and importance of keywords in each topic.

Python's gensim package is used to apply LDA topic modelling. Scatterplot is a static chart, where as the two remaining two are specific to each topic. In order to view these charts for a specific topic select a topic from the list of topics present above the charts. Analysis of "Tableau" topic is shown on default load.

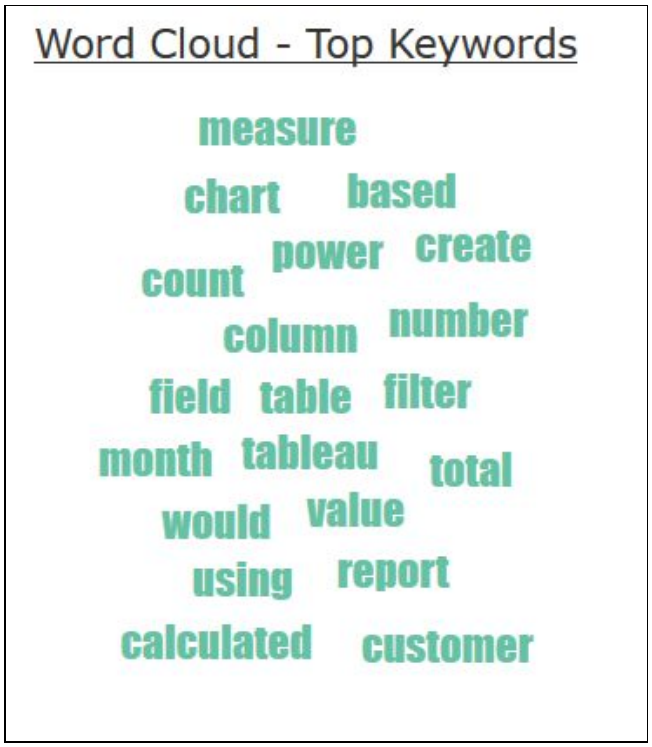
1.1.2 Scatter plot to visualize clusters:

This chart is used to visualize how distinct the topics are. The scatter plot visualizes the clusters of posts in a 2D space using t-SNE (t-distributed stochastic neighbor embedding) algorithm. Each point represents a stackoverflow post represent in word space model which has high dimensionality. We reduced the dimensionality using PCA (Principal Component Analysis). t-SNE converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. Cluster is distinguished by the color of the point. Each color represents a cluster in which the cluster is named. Same color is used to populate the other two charts to maintain consistency. I used TSNE method provided in sklearn's manifold package to convert high dimensional data to low dimensional data. The scatter plot chart is shown in figure below



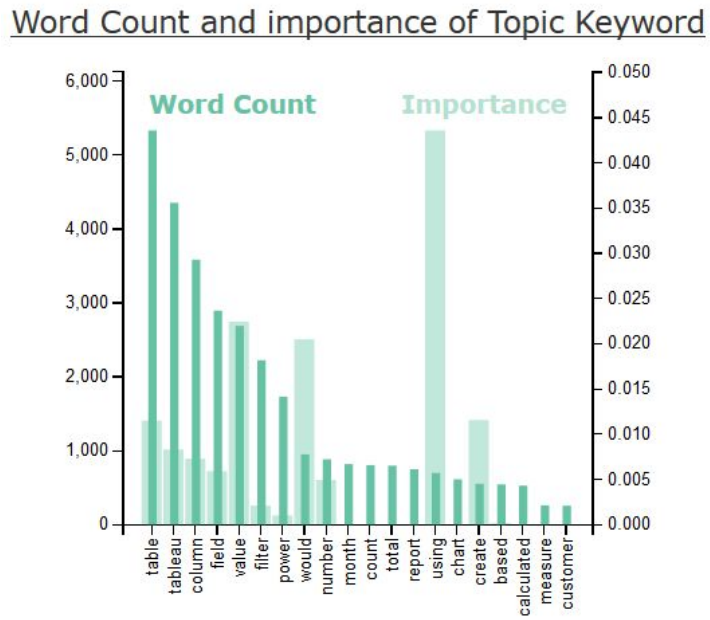
1.1.3 Word Cloud: Topic Modeling:

Each topic comprises of keywords which describe the topic. Infact LDA does not label the topics, we labelled the topics looking at the keywords it contains. Hence, we created a word cloud with the size proportional to the weight shows the important keywords in each topic and its importance.
As can be noted from Figure 2, the word “Table” is more important than the actual topic word in “Word Cloud” chart of Tableau topic.



1.1.4 Word Count and Importance of Keywords

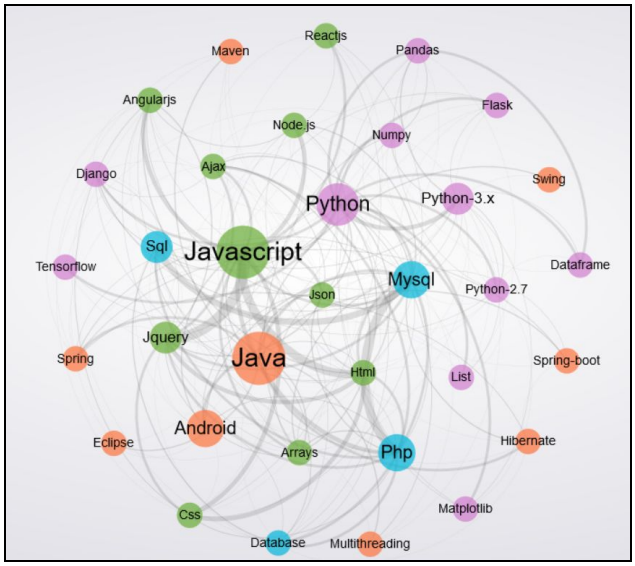
When it comes to the keywords in the topics, the importance (weights) of the keywords matters. Along with that, how frequently the words have appeared in the documents is also interesting to look. Some of the words occurred in multiple documents and are realized to be less important, we ignored such words by adding them to the stop words list.
In this plot x-axis represents the keywords occurring in the topic, left y-axis represents the number of times the word appeared in the cluster, and right y-axis represents the importance of the word in the cluster. Some of the keyword related to Tableau topic are: measure, table, calculate, value, filter. As can be seen in Figure 3,the word “using” even though is the most frequent word, it is one of the least important keywords.



1.1.5 Network Graph:

For network graph, the adjacency matrix was created. The raw data of all the questions with tags were queried from the Stack overflow datasource. Now with around 250,000 questions, we found 10,000-12,000 different tags clustered around “Java”,”Python”, “Php”, “Javascript” and “Mysql”. Now with a visualization having that many technologies together on a single network would overwhelm the graph making it very difficult to read. To overcome this we decided to select few top technologies (observed high frequency).

We generated an adjacency matrix of all the technologies used in these questions. The resultant matrix being very sparse, with 3000-5000 tags attached around top programming languages like Java, Python, PHP etc. To have the network clear to users we decided with top 10 tags with 4 most occured tags. With the cleaned and analysed data, the next step was to generate a clean network, to do that we used a tool “Gephi” to cluster and generate the graph files. We applied Fruchterman Reingold algorithm to create a force-directed layout for the graph. Now this graph file was visualized.



1.1.6 Time Series: Sentiment Analysis:

Sentiment analysis is the automated process of understanding an opinion about a given subject from written or spoken language. Sentiment analysis is a key tool for making sense of the data. Often times users who post a query on stackoverflow, post it with a sense of frustration. Answers include references to pain points. But sometimes a features is praised and lauded as well. We wanted to track the sentiment of user who posted queries related to JAVA in the year 2018. We collected around 500 thousand posts related to JAVA and applied sentiment analysis.

The sentiment analysis tab of the website contain the dual axis bar and line chart. User can view the sentiment flow of different months by changing the dropdown. Figure 4 shows a snapshot of time-series sentiment chart. The bars represent the number of positive and negative posts created on the day. Line represents the positive and negative sentiment of the created posts. Since direction aggregation of posts may not convey the sentiment correctly we applied 4 statistical aggregation measures like mean, median, max and min. We noted that mean still accurately represented the sentiment and chose mean sentiment value for the day to plot the graph.

We used vader sentiment analyzer, VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of a sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative. We used vader sentiment analyzer because it not only tells whether a post is positive or negative, it tells whether it is neutral or has mixed sentiments.

From our analysis we noted that most of the posts had neutral sentiment, followed by composite sentiment, negative and positive sentiment.

4 METHODOLOGY

- We have used the combination bar and line chart to show sentiment over time as just presenting the sentiment is not enough. Lines represent the sentiment. Bars represent how many posts contributed to that sentiment score. We also did not combine positive and negative sentiments together because it might show the correct sentiment
- We used a scatter plot to visualize the clusters as they are very helpful in assessing the clusters. In our topic modelling cluster, points of a cluster are grouped together except for certain outliers which shows topic modelling has effectively clustered the topics even without the tag information
- We used word cloud to access each topic as word clouds represent the keyword which constitute the topic
- We used a dual axis multi-bar chart to visualize the importance of keywords with respect to its occurrence. As noted there is no direct correlation between count of words and their importance in the document. This helps in naming and grouping the topic correctly.
- In the Network graph, we gave more focus to the cluster roots and links to other nodes. The node size and edge width is varied with the correlation to their occurrence in the questions posted over the website. Having varied the width of edges, another panel is being added to give user the exact number of languages being used together.

5 EVALUATION PLAN

As discussed above, our major focus is to understand user sentiment over time and identify hidden topics in the posts when tag information is not taken into consideration. We focused our analysis along the same lines.

We had all our visualisation focus on either one or both of these areas. After doing all the analysis of the data and designing the visualizations discussed in past sections helped to get insights/trends like:

- Sentiment Time Series Chart: In this chart we noted that overall positive sentiment was still higher than the negative sentiment. This is because vader assigns positive sentiment score even to words which do not have a sentiment in technical sense. Hence we assessed that positive

sentiment is less effective in assessing the user's emotion.

Vader analyzer does not mark negative sentiments to neutral words, hence more importance should be given to negative score and count of negative posts.

- We see that the number of posts roughly follow a sine wave peaking during mid week (Tuesday-Thursday) and lowest during the weekends which is expected as programmers work majorly on weekdays.
- Both positive and negative sentiments fluctuate more during mid-week than on the weekends.
- Topic Modelling: We used posts from 15 technologies to apply topic modelling and LDA could correctly identify 10 such clusters. Remaining topics fell in different buckets like object, error. Given these are the major keywords of every technology we expect to experience this even in other scenarios. Overall LDA was able to correctly cluster **70%** of the posts.
- Word cloud and importance bar chart have enabled in naming the topics correctly and visually contrast of importance of a keyword.

5 DISCUSSIONS & FUTURE WORK

In the future, we can do a couple of enhancements to our existing systems. Sentiment over time can be extended to identify sentiment of user towards and entity to fully utilize the chart and make more sense of the visualization. We captured composite sentiment as well but did not incorporate in the visualization. We could extend the chart to show sentiment of other technologies.

We performed topic modelling on posts with single tag and used the tag as truth value. Since topic modelling is capable of cluster a post to multiple categories we can use posts with multiple tags to see how topic modelling performs in such posts. LDA uses bag of words to generate vector representation of each word. We can create a neural network based deep learning model which learns the word, post and topic representation.

For Network graph, we just used the data for a few languages and furthermore we can extend the network to handle more nodes and grow the network according to the node selected gaining spaces for more languages.

6 REFERENCES

1. https://github.com/cjhutto/vaderSentiment
2. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24
3. https://www.nltk.org/
4. https://radimrehurek.com/gensim/
5. https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.min.css
6. https://www.w3schools.com/w3css/4/w3.css Martijn Tennekes and Edwin de Jonge - Tree Colors: Color Schemes for Tree-Structured Data -
7. https://pdfs.semanticscholar.org/6f4e/96b5a487b556cccffc5f9e6b246bbbb33d63.pdf - Last Accessed April 26th '2018
8. https://cdnjs.cloudflare.com/ajax/libs/font-awesome/4.7.0/css/font-awesome.min.css
9. https://github.com/raphv/gexf-js
10. https://gephi.org/users/tutorial-visualization