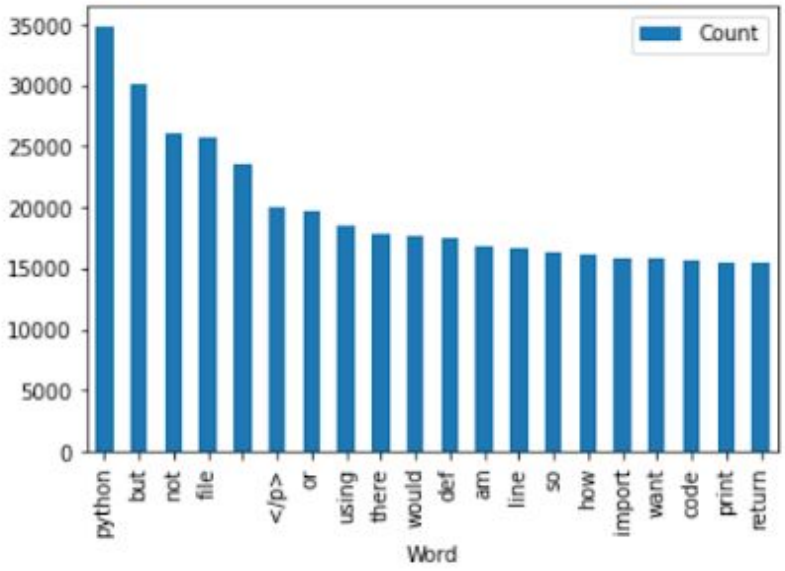
Task-III

1. The subsampling method involves the extraction of data from the bigger dataset. In this the subsampling criteria is extracting important information from a bigger set of information. Most of the data is either regarding user posts on StackOverflow and its related fields or tags which stack overflow has data for. For eg., The criteria for subsampling in Badges.xml seems to be the Badge filter ‘Teacher’. Some other common attributes are class:3 and date and time. For tags.xml, the file has been made by including all the tags of common topics of concern for the users. For Posts.xml, all the posts are from 2008.

2. Following is the word cloud on Tags.xml. It shows how javascript is the most occurred tag. All the sizes are according to the frequencies of the tags

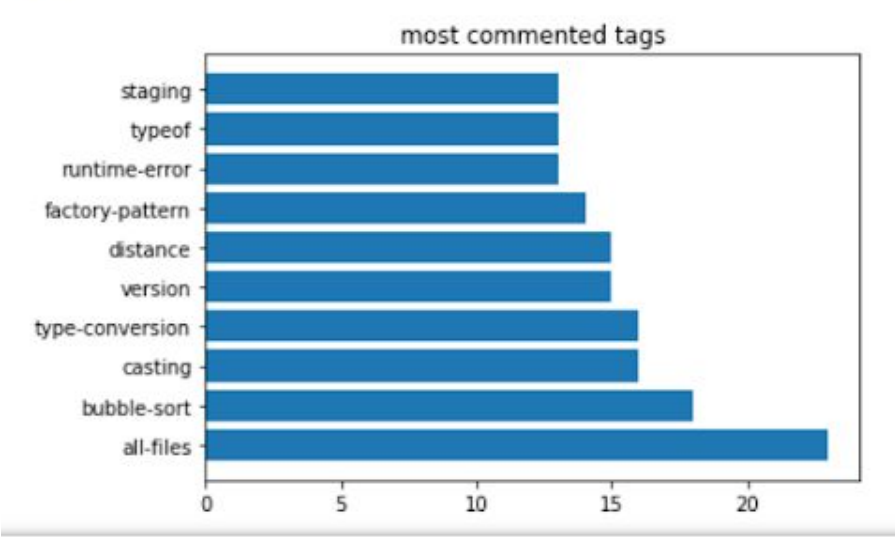
3. Following is the data analysis carried out on Posts.xml

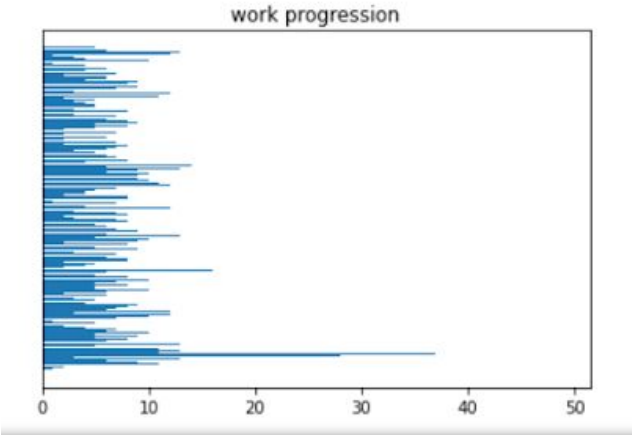
This is a Bar Graph on the top 10 most viewed tags



This is a bar graph of the most occurred word in all the posts.

- This shows what the main concerns of the users were. Like python is the most occurring word, hence, most users seemed to be concerned by python related queries, this also goes well with the previous graphical output.

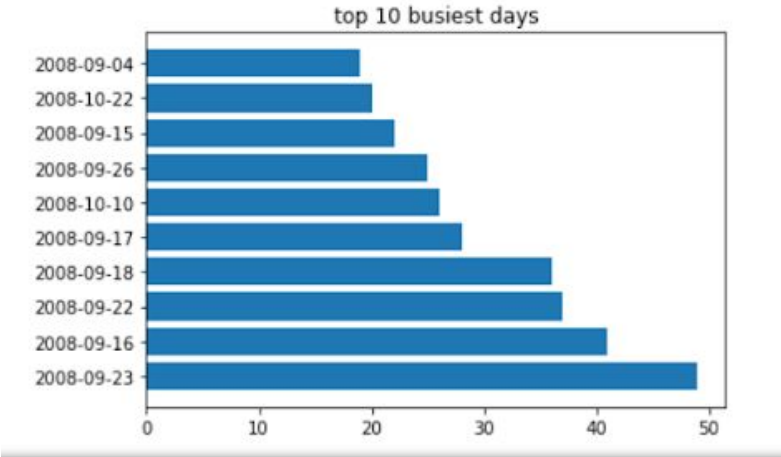




- This graph shows the progression of user activity. Continuous progression can be observed.

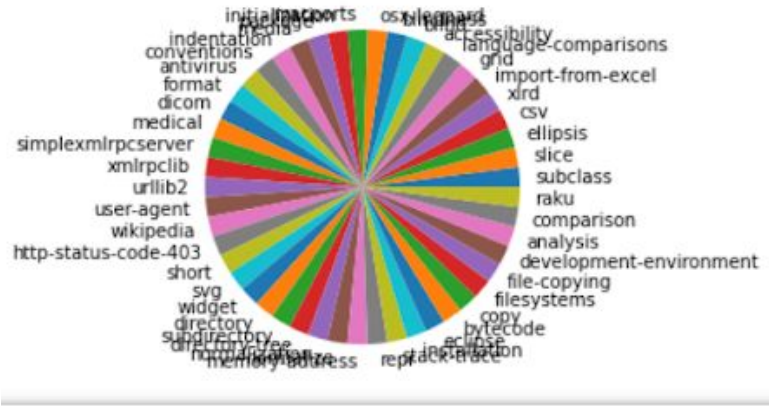
- This is a graph of the most commented tags. This shows that these
  Tags welcomed most people to share their thoughts on the post.

19

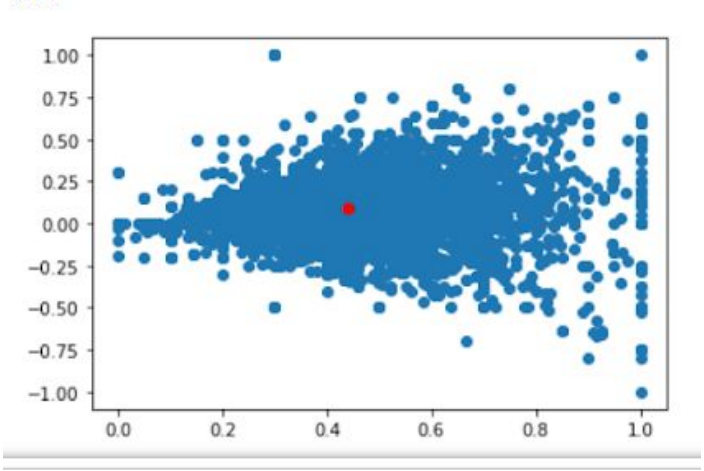


This shows the dates which gathered the most posts.

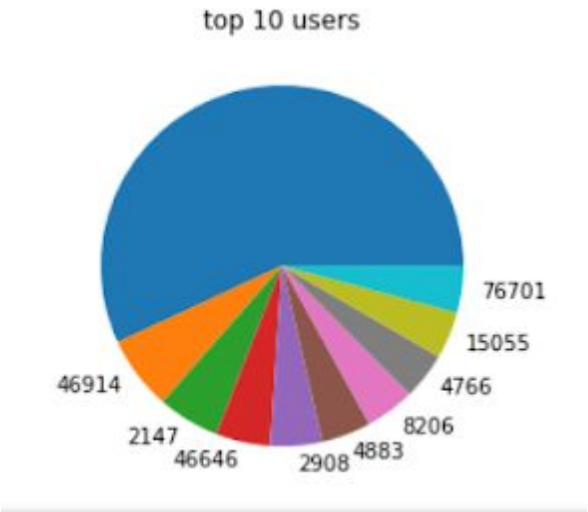
1



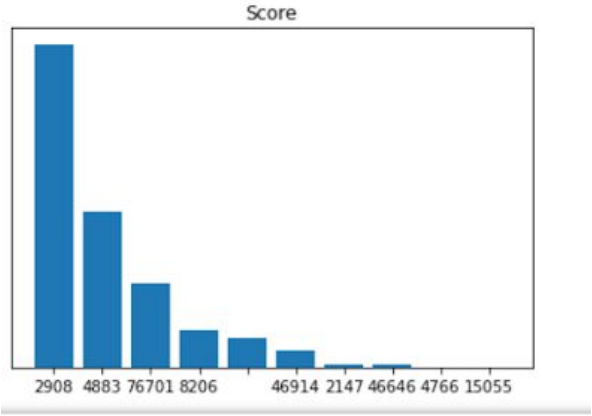
This is a pie chart of the tags questions were asked on on the busiest day



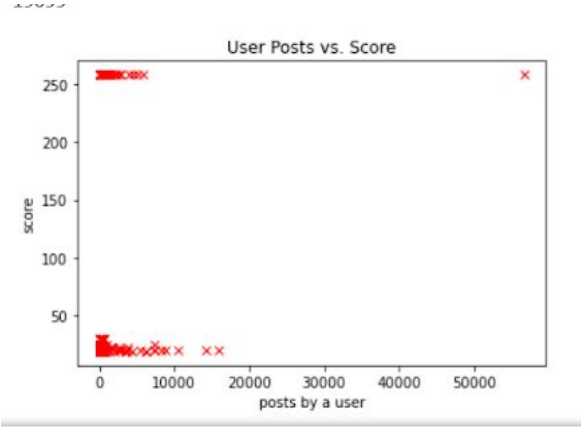
- This graph is of the polarity vs subjectivity of these posts. It makes sense that these posts have positive polarities and intermediate subjectivity as they generally deal with questions a user asks. The red dot is the mean subjectivity, polarity.
  (polarity: x, subjectivity:y)



- This is a graph of the top 10 users along with their ids. Most active user takes up more than 50% of the graph



This is a graph of the scores of the top10 users



- This graph shows posts by top 10 users vs score. It shows how more posts result in more scores.

- From the analysis carried out, we can tell that most use