Task 3B: StackOverflow Analysis

Report

T. H. Arjun

IIIT-Hyderabad

Overview

In Task 3b, I used the python script xml_2_db.py to convert the .xml files to mongodb Database. The mongodb dump is here. For the code the stackoverflow.com folder is required which was given as part of the task. The code parses the files and inserts the data to mongodb in chunks (to avoid large times). StackOverflowAnalysis.ipynb is the Jupyter Notebook containing the analysis code.

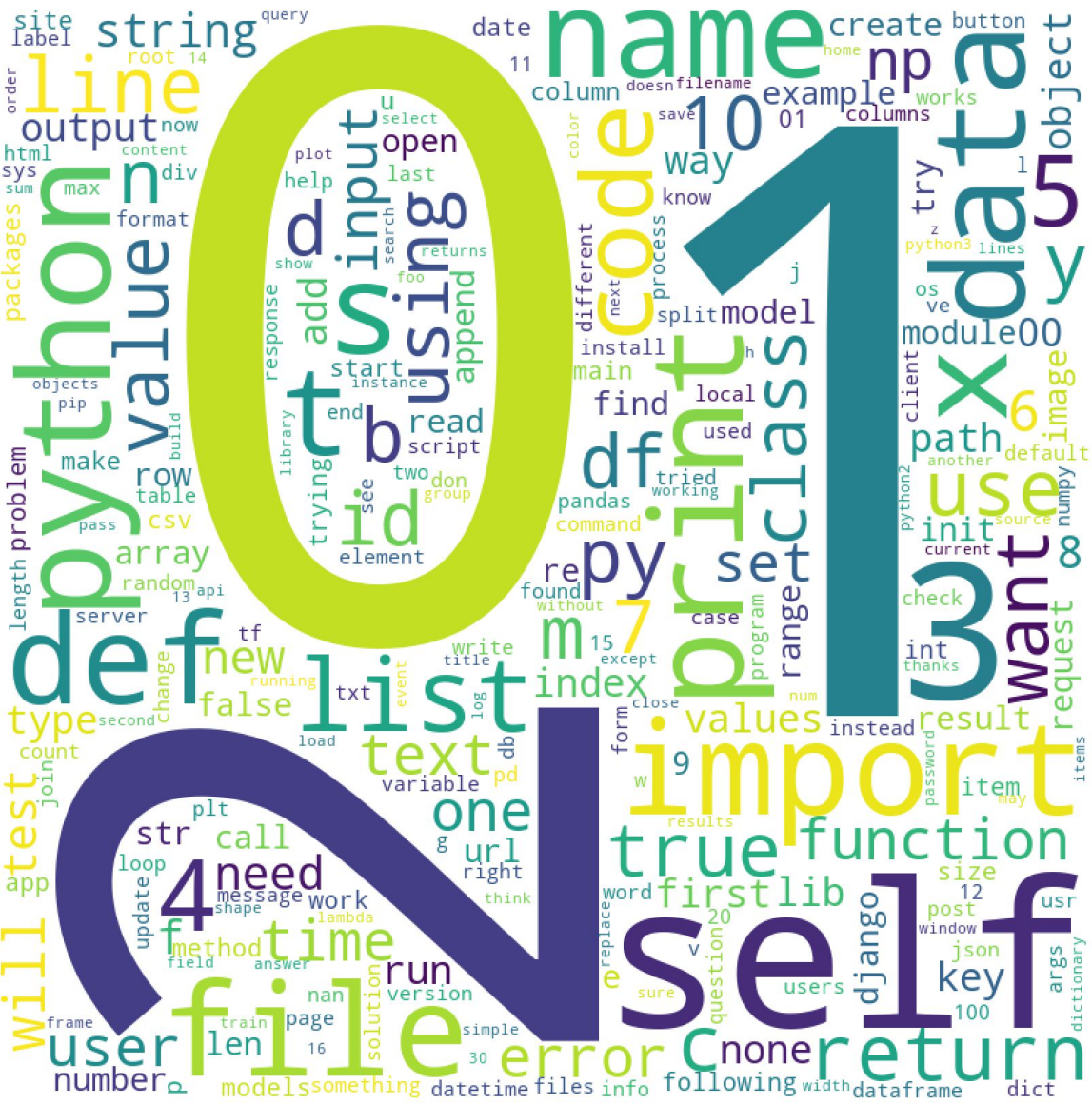
Analysis

I. What criteria was used for subsampling?

From the word cloud generated I think the criteria used for Sub Sampling was to collect Posts related to Data Science. As the words that appear are Data Science Related Technologies and Terms.

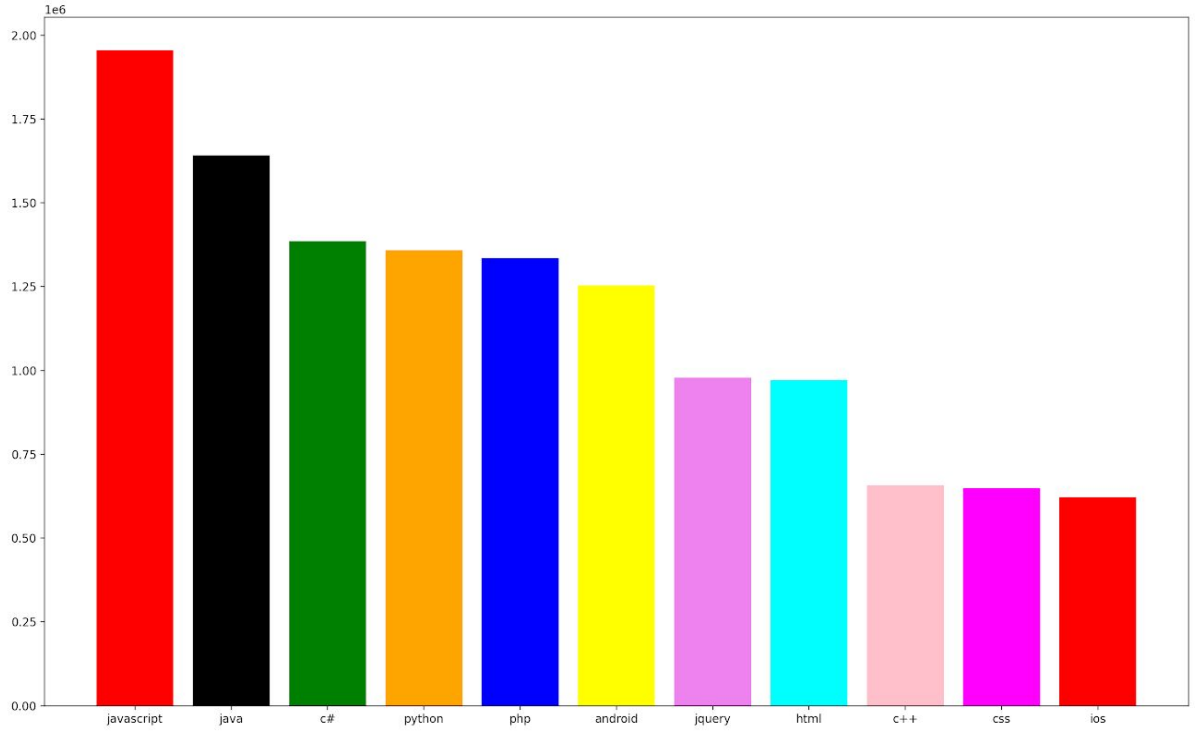
II. Generating a WordCloud

The word cloud is generated in the Jupyter Notebook. The word cloud shows words from python syntax and other Data Science related topics such as numpy, xml, files, re, pandas dataframe etc. We also see our favourite numbers 0,1 . This also leads me to believe that the criteria used for the dump was Data Science related Technologies.



III. Top 10 Occuring Tags

Here we find out the Top 10 Occurring Tags on Stack Overflow. We loop through tag.xml to do this.



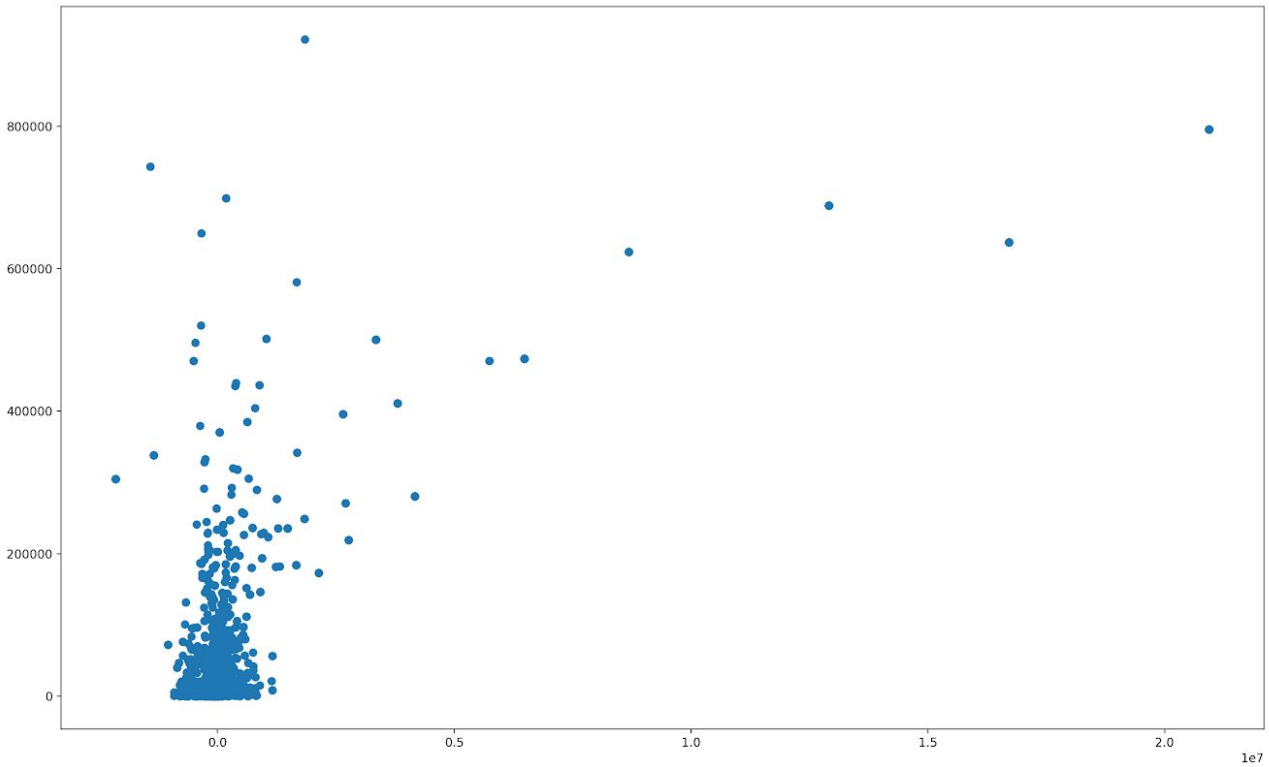
IV. Analysis of Reputation vs Skill

StackOverflow uses a metric called Reputation for each user. It is a number specifying the useful contributions of a user on StackOverflow. This metric results in the gamification of the site. We can analyse the relation between Reputation and Skill level of a user using Elo_rating_system . Each question is seen as a "game", the responders were seen as "players" and the outcome of each game was determined by the votes. The player's skill was then estimated using the Elo rating system. Initially everyone was given a skill of 1500 and the game was started. When a user answers a question, his/her skill will get a positive update if the answer gets more upvotes than the other answers on the same question, and a negative update if it gets less upvotes. The magnitude of the skill update is determined by the reputation of the other users and whether or not the outcome was expected. I used the absolute difference in reputation as the metric for seeing if it is expected or not and updated accordingly. Because the process takes a lot of computing power I did the study for a random set of 5000 posts from the database

X-Axis is Skill and Y-Axis is Reputation

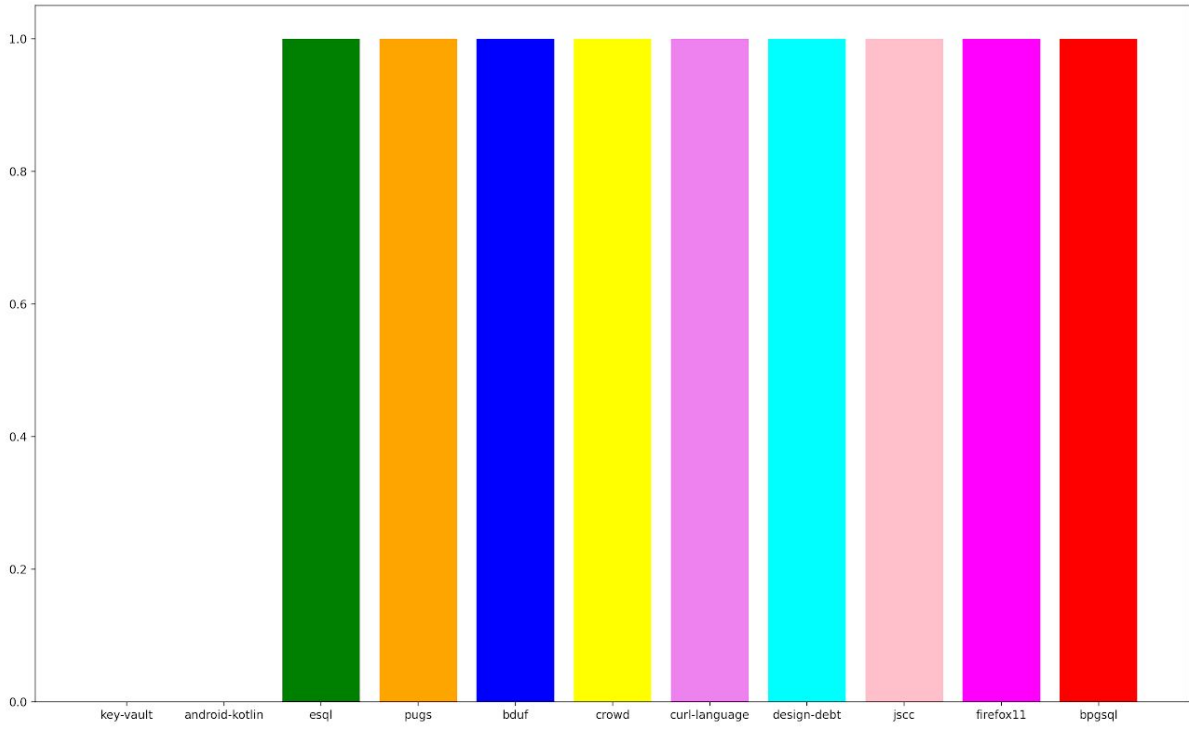
Observations Made

1. Reputation and Skill shows a Relation.
2. Users with high reputation have high skill levels as expected.
3. Most of the Stack Overflow Users have similar skill level and they have similar reputations (Cluster in plot)
4. A low reputed user can be skilled or unskilled.



V. Least Occurring Tags

By looking at the top 10 least occuring tags we can see a relation. The community doesn't like vague topics and tags. They like tags that are common and they can answer fast and easily.



VI. Most Occuring Badges