REPORT | TASK 3B

Analysing StackOverflow Data Dump

The Task

The task is to analyse an abridged version of the StackOverflow Data Dump. The dump consists of 5 files, namely `Posts.xml`, `Votes.xml`, `Tags.xml`, `Badges.xml` and `Users.xml`. The files have to be first converted into a Mongo Database and then analysed to find patterns and insights.

One of the major challenges is the size of the data. Totalling to 7.7 GBs, with individual files as large as 5.2 GBs, this dataset requires that all code that handles it be written keeping memory consumption in consideration.

Parsing

Usual XML parsers construct the entire Element Tree at once, essentially loading the entire file in memory. Because this is not possible with the given dataset, we resort to iterative parsing. The elements are parsed one-by-one, and each row is read, converted to JSON and saved to MongoDB. Before moving onto the next row, we delete this row, as well as the path from the parent to this row from memory to keep the memory consumption within limits. This allows the parsing time to scale roughly linearly with file size, and memory consumption to remain constant.

References: *High-performance XML parsing in Python with lxml* by Liza Daly [IBM Developer Resources]

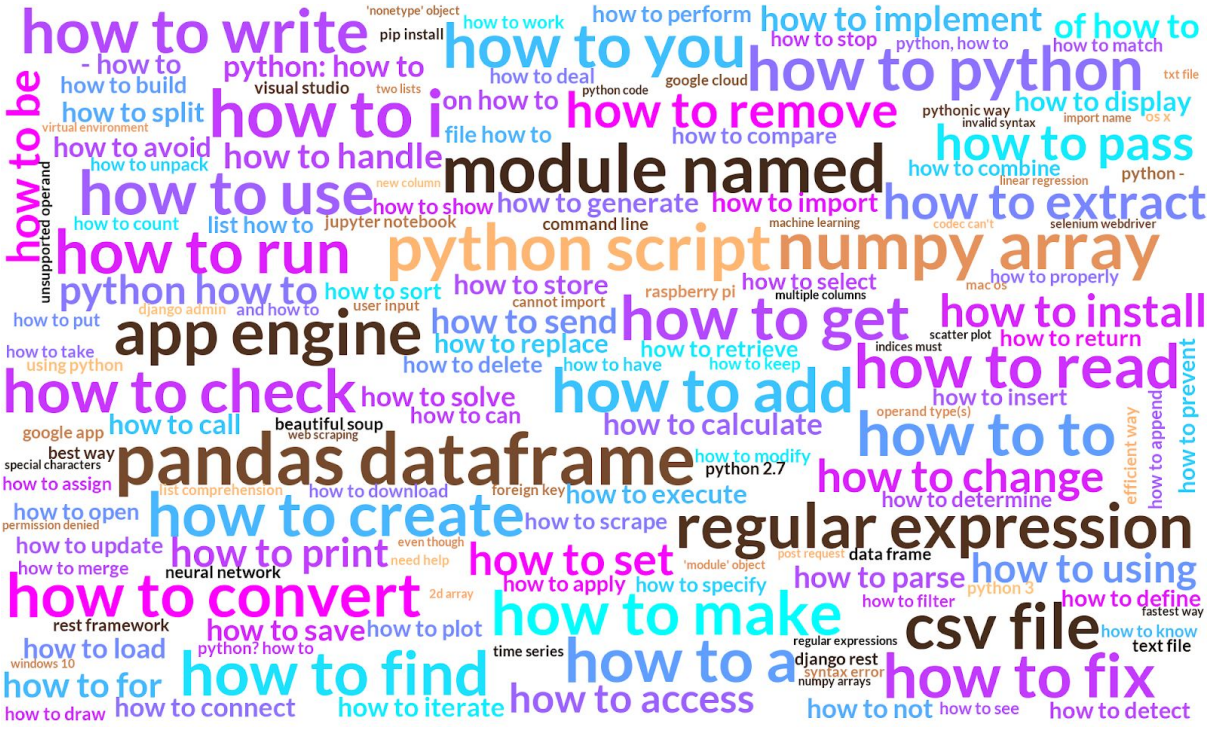
Analysis

Subsampling Condition

After plotting the top 10 occurring tags as well as the word cloud as detailed below, we notice that almost all terms correspond to the Python language. While Python is popular, other languages aren't so far behind. Using the StackExchange Data Explorer, we find that there are roughly 17k questions with the tag "Python" and 40k questions with the tag "Java". We'd expect the subsampled dataset to maintain a similar ratio (in spite of the dataset being only till March 2020) unless this was the subsampling criteria.

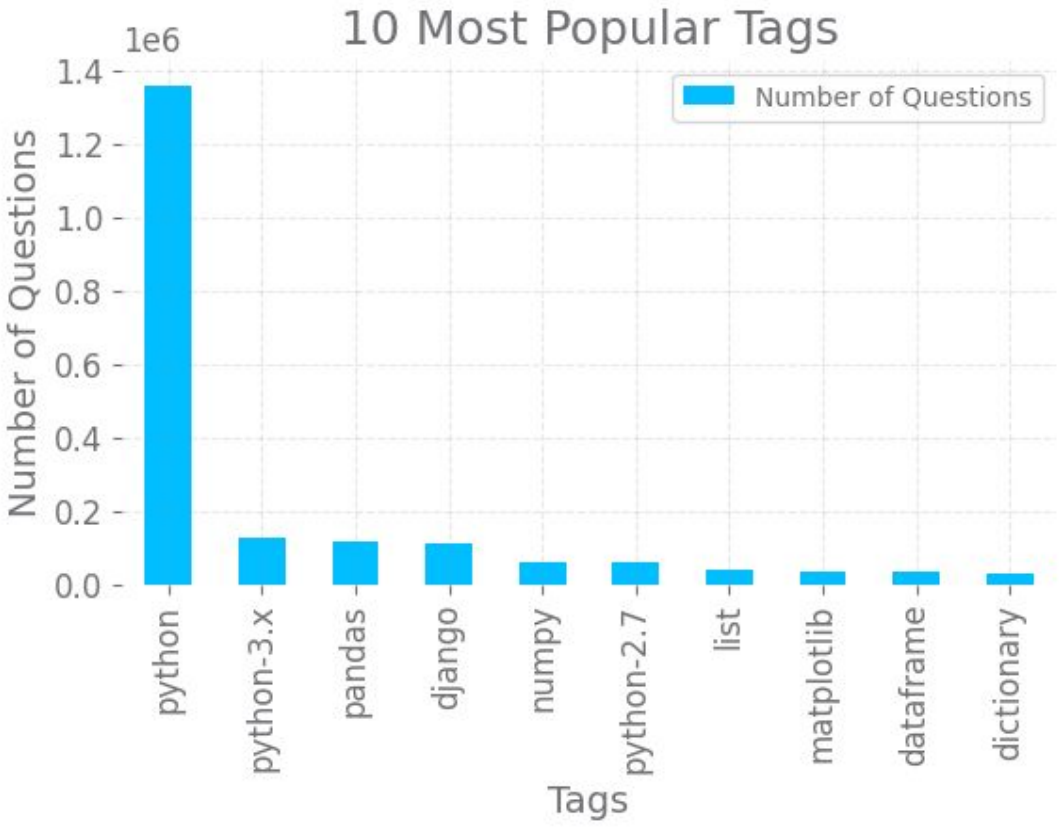
On closer analysis, we realise that all the questions in Posts.xml are tagged "Python", and the answers correspond to only these questions (other post types have not been included). Further, the users in Users.xml are only those who have contributed to at least one post from Posts.xml, as revealed by a join between the User's Id and Post's OwnerUserId. Other files correspond to these users and posts. The dataset only contains data up till March 2020.

Word Cloud



This word cloud consists of popular collocations present within the titles of the posts. The collocations chosen are of two types: Bigrams, after removal of stopwords (displayed in shades of brown), which reflect common terms and technologies being discussed in the posts (such as "pandas dataframe" or "import error") and Trigrams, without removing stopwords (displayed in shades of blue and magenta), which reflect popular questions being asked (such as "how to fix", "how to iterate" or "how to display"). The Bigram "how to" doesn't show up in brown because stopwords were removed. The collocations are ranked using their likelihood ratio.

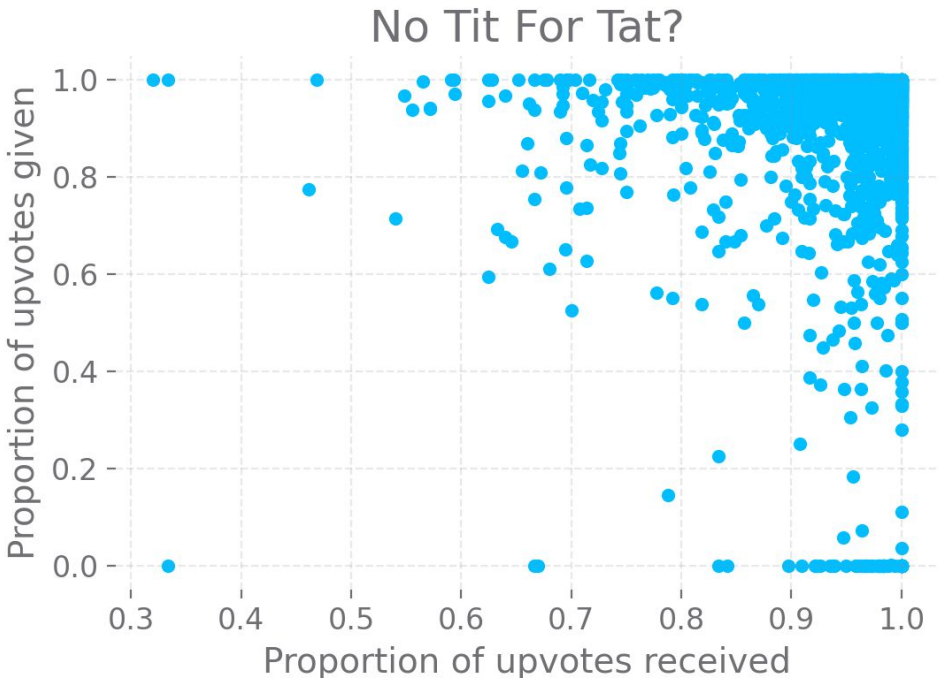
Popular Tags



Other Visualisations

Taking reference from the paper discussed in Task-1, we plot the ratio of positive votes given versus received for each user, hoping for a tit-for-tat effect (linear correlation). However, we find none. While this may well be due to deficiencies in the data collection and cleaning process, it could also reflect that

StackOverflow users take upvoting and downvoting much more sincerely, as it plays a crucial role in the StackOverflow moderation process.



We also plot the score (difference of up- and down-votes) densities of the accepted and other answers and notice that accepted answers have a slightly higher score on average.