Stack Overflow Question Tag Prediction

Project Members
Mudra Sahu 193050007
Muttineni Navya 1930500017
Ankita Singh 19305R002
Tejal Topre 19319R002

June 11, 2021

Indian Institute Of Technology Bombay
Powai, Mumbai, India

Contents

1 Objective

The project is aimed at automatically assigning relevant tags to the questions posted on stack overflow . This will help stack overflow in notifying the appropriate users who can answer the given question. The problem says that we will be provided with a bunch of questions. A question in Stack Overflow contains three segments Title, Description and Tags. By using the text in the title and description we should suggest the tags related to the subject of the question automatically. These tags are extremely important for the proper working of Stack Overflow.

2 Related Literature

2.1 Stack overflow

Stack Overflow is a question and answer site for professional and enthusiast programmers.It is the largest, most trusted online community for developers to learn, share their programming knowledge, and build their careers.It features questions and answers on a wide range of topics in computer programming. The website serves as a platform for users to ask and answer questions, and, through membership and active participation, to vote questions and answers up or down and edit questions and answers.

2.2 Multi-label Classification

It is a variant of classification problem in machine learning where multiple labels may be assigned to each instance.It is a generalization of multiclass classification, which is the single-label problem of categorizing instances into precisely one of more than two classes; in the multi-label problem there is no constraint on how many of the classes the instance can be assigned to.

2.3 Multiclass Classification

Multiclass or multinomial classifcation is the problem of classifying instances into one of three or more classes. The existing multi-class classification techniques can be categorized into (i) Transformation to binary (ii) Extension from binary and (iii) Hierarchical classification. Binary Classification problems are classified as one vs Rest and One vs One. One vs Rest strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negative.

2.4 Support Vector Machine Classifier

It is a discriminative classifier which uses separating hyperplane.Given labeled training data,the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

2.5 Logistic Regression Classifier

Logistic Regression is a ‘Statistical Learning’ technique categorized in ’Supervised’ Machine Learning Methods dedicated to classification tasks. In a classification problem, output variable, y, can take only discrete values for given set of input features,X. But Logistic Regression is a Regression Model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”.

2.6 TFIDF

TFIDF stands for Term frequency-inverse document frequency. TFIDF are word frequency scores that try to highlight words that are more frequent in a document.The higher the TFIDF score, the rarer the term is. Tf-idf term is composed of two terms, Normalised Term Frequency (tf) and Inverse Document Frequency (idf). Term frequency is the frequency of term t in document d and Inverse Document Frequency - idf is computed by taking log of number of documents containing the term t divided by document frequency of a term t. Finally,TFIDF is computed by taking product of term frequency and inverse document frequency.

2.7 Performance Metrics

Various performance metrics can be used to evaluate performance of ML algorithms, classification as well as regression algorithms.For multiclass classification problems,we can use performance metrics like Precision, Recall, F1 Score, Logloss,etc. Here for binary multiclass classification,we have used Micro Averaged F1 score, Macro Averaged F1 Score and Hamming loss.

1. **Macro Averaged F1 score :** Macro Averaged F1 score is simple average of each F1 score for each label ‘k’.It doesn’t take frequency of label or tag into consideration.

2. **Micro Averaged F1 Score :** Micro Averaged F1 score is weighted F1.To get Micro F1 Average for a given set of tags, we compute precision and recall for each tag.It is not preferred when some tag occurs lot of times and some occurs few times.

3. **Hamming Loss :** Hamming loss is the fraction of labels that are incorrectly predicted. It is fraction of wrong labels to the total number of labels.

3 Description Of Approaches Tried

- One vs rest classifier with Stochastic Gradient Descent with bag of words features and logistic loss
- one vs rest classifier with Stochastic Gradient Descent with term frequency features and logistic loss
- one vs rest classifier with Stochastic Gradient Descent with bag of words features and hinge loss (SVM)
- one vs rest classifier with Stochastic Gradient Descent with term frequency features and hinge loss (SVM)
- one vs rest classifier with Gradient Descent with with term frequency features and logistic loss
- one vs rest classifier with Gradient Descent with bag of words features and logistic loss

4 Experiment

4.1 Preprocessing

4.1.1 Data overview

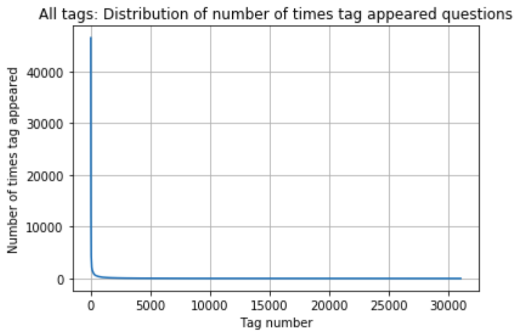
Train.csv contains 4 columns

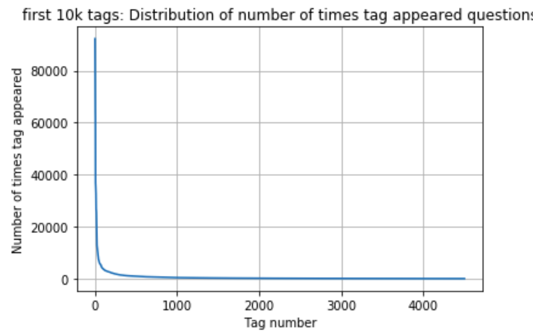
1. Id : Uniquely identifies a question
2. Title : Title of the question
3. Body : Body of the question
4. Tags : Tags associated with the question

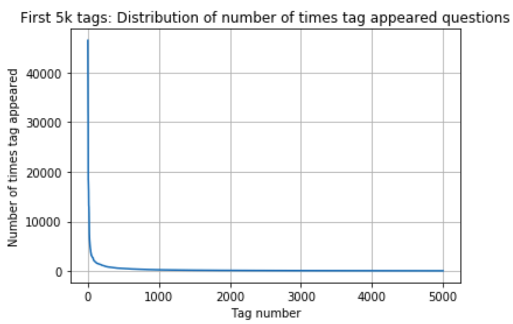
Number of rows in Train.csv = 603419

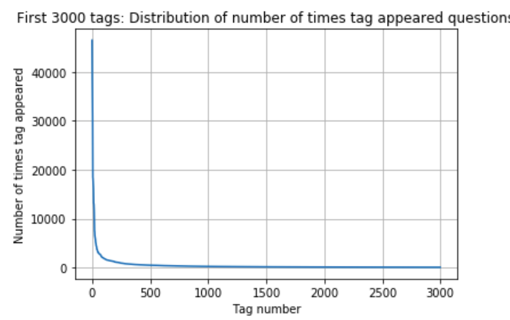
4.1.2 Analysis of Tags

It is important to understand how many times a tag has appeared, as it will help in curtailing the non significant tags. There were around 31k tags in total. It can be observed from graphs that only a small fraction of them has appeared frequently. Below we have shown plotted distribution of number of times tags appeared in question in descending order :









4.1.3 Remove stop words

The word cloud before and after pre-processing have been shown in figures 1 and 2

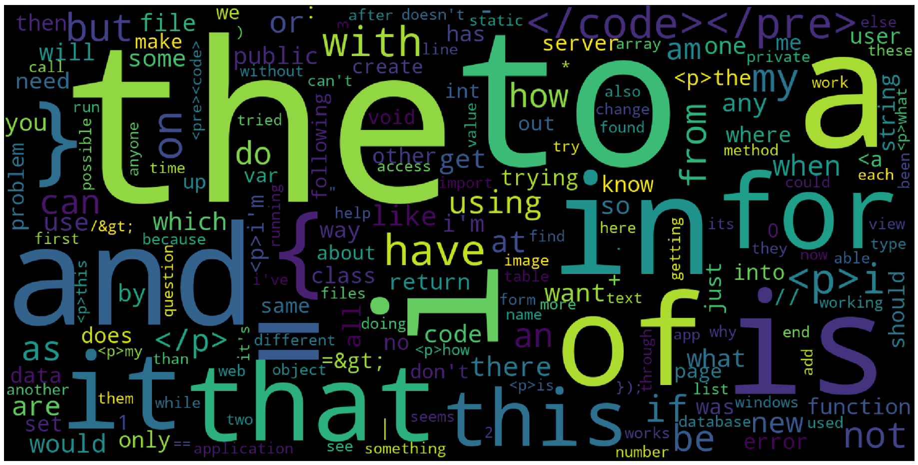


Figure 1: Word cloud on (2 x title + body of question) before removing stop words

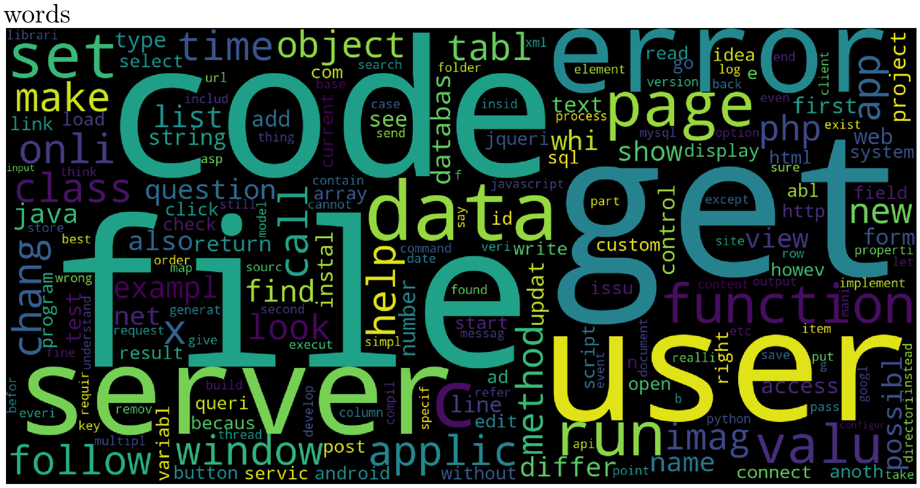


Figure 2: Word cloud after removing stop words

4.1.4 Remove HTML and code using Regular Expressions

The training data was collected by web scraping stack overflow so the data contains HTML Tags which needs to be removed. Also a lot of user post code snippets in the question body. The content contained within these tags(<code> </code>) is generally specific to their application and does not contribute much in training the model. It would rather confuse the model so the code tags needed to be removed.

4.1.5 Stemming of words

The project uses Snowball Stemmer to reduce words to their root form.

4.2 Details about code structure

- The code can be found in the links below:
  (i)Code for Modelling (ii)Code for Preprocessing
- The language used is python - approximately 200 lines
- Environment - Jupyter Notebook in Google collab
- Memory allocated via GPU

4.3 Experimental Platform

Our experimental platform was google collab . We tested our approach via various graph plots and initially training 10k rows of data considering 100 tags . This was done on various classifiers .Pre-processing techniques were modified based on word cloud results. The code run took approx 6 hours to complete (all models).

4.4 Modelling

To get the individual word features for the Tags we have considered two of the most popularly used featurizations :

1. Bag of words
2. Term Frequency Inverse Document Frequency (tfidf)

For each of these features we used one vs rest framework of sklearn module , this strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes. In our case each class denotes a tag and every question is classified against all the tags .

Now for both the featurization techniques mentioned above, we applied stochastic gradient descent on log loss as well as hinge loss to get the optimum values. We also tried the gradient descent for the same. The results obtained in terms of F1 scores and hamming loss are as follows :

Classifier	Macro f1 score	Micro f1 score	Hamming loss
OVR with SGD, log loss, using Bag of Words	0.25682086818252037	0.33851336665942183	0.006052199850857569
OVR with SGD, log loss, using TfIDF	0.3479273223107815	0.47727973364572407	0.002809843400447427
OVR with Logistic Regression, using Bag of Words	0.3754717190494653	0.47662404928713575	0.0031815394813157678
OVR with Logistic Regression using TfIDF	0.34285440513546256	0.4650769808486669	0.0028327119065374097
OVR with SGD Classifier --> Linear SVM using Bag of Words	0.2529228710277107	0.3350882848035529	0.006152953848703289
OVR with SGD Classifier --> Linear SVM using TfIDF	0.31649590575276954	0.4891131847901987	0.0026673295219156517

Also we used sklearn package GridSearchCV to tune the hyperparameter values.The F1 score we obtained against each classifier metric are as follows:

Classifier	**Featurization**	**Micro f1 score**	**Hyperparameter alpha/C value**
OVR with SGD, log loss,	Bag of words	0.33851336665942183	0.00001
OVR with SGD, log loss,	TfIDF	0.47727973364572407	0.000001
OVR with Logistic Regression	Bag of words	0.47662404928713575	1
OVR with Logistic Regression	TfIDF	0.4650769808486669	1
OVR with SGD Classifier --> Linear SVM	Bag of Words	0.3350882848035529	0.001
OVR with SGD Classifier --> Linear SVM	TfIdf	0.4891131847901987	0.001

Thus we saw that logistic regression (both via stochastic gradient decent / gradient decent) were giving the best results among other classifications. Hence we chose the same to run on the final test data .

It was initially considered that LSTM might be used for making tag predictions. LSTM adds several more weights to each of the preceding words considered in the title+body part which is not needed for our modelling, since our predictions depend mostly on individual words rather than word dependency. It also takes a lot of time to train and the model we considered provides an improved time threshold over LSTM.

4.5 Effort

The time we spent on different parts of the project can be summarised as follows :

1. Understanding and visualising the problem statement :20%
2. Pre-processing of data : 35%
3. Modelling the data via various approaches and hyper-parameter tuning:40
4. Discussing approaches ,format and problem solving : 5%

The most challenging part was to train the data in the given time constraints and determine which model was best suited for our data.

Division of work:

1. Mudra : Data analysis and pre-processing
2. Ankita: Pre-processing with different approaches ,hyper parameter tuning
3. Navya : Modelling the data via various approaches and hyper-parameter tuning
4. Tejal : Pre-processing and data recording