International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Question Tagging System

Harsh Parikh[1] , Parth Patel[2] , Vatsal Sanghrajka[3] , Chintan Savla[4] , Manya Gidwani[5]

[1] Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India
[5]Assistant Professor Department of Information Technology, Shah and Anchor Kutchhi Engineering College, Mumbai, India.

Abstract: *With the advent of modern technology electronic goods are getting cheaper and better. Mobile phones and computers have become basic necessities. As internet services are becoming faster and cheaper, almost everyone is connected to the web. It's a human tendency to take opinions from different people before taking a decision. Earlier a person could ask their family, friends or colleagues if they had doubts about something or wanted to acquire information about a specific topic. This human tendency of consulting people before taking decisions and easily accessible web has led to the generation of Q&A forums. People can post their Question here and other experienced users can provide answers to them. Over a period of time, it may happen that same question may be asked again and again which has already been answered earlier. To save the trouble of answering the same question again and again it is advisable to tag the questions. So if another user searches the web, he/she can find the same question and the related answer to that. In this paper a prototype of a system aimed to assist the users while tagging their questions is proposed.*

Key words: Word2Vec, Skip-gram, F1 Score, Tokenization, Tagging.

1. INTRODUCTION

To tag the question we need to understand the question first, question are in the form of text. So we need to perform Text Mining. Text Mining is a subdomain of data mining. Data mining is the analysis step of the knowledge discovery in databases process, or KDD. The overall goal of data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Text Mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics. Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analysed with traditional data mining techniques.

Here in this research we apply text mining to tag the question posted by users on Q&A forums to make it easier for users to read and save the efforts of answering the same question again and again.

We are provided with a data from different domains wiz. Cooking, DIY, Travel and Biology which consist of title, content and tags for each and every question. We are asked to predict the tags for the Physics domain. In the Physics dataset we have the task to predict the tags from the title and content already present in the data.

id	title	content	tags
1	What is the criticality of the rib	translation, how critical	ribosome binding
2	How is RNAse contamination in	<p>Does anyone have any	rna biochemistry
3	Are lymphocyte sizes clustered	<p>Tortora writes in	immunology cell-
4	How long does antibiotic-dose(	<p>Various people in our	cell-culture

Fig. -1: Data Overview

A major assumption in many machine learning and data mining algorithms is that the training and the future data must be in the same feature space and have the same distribution. However, in many real life applications this may not hold true. For example, we sometimes have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter data may be in a different space or follow a different distribution. In such cases knowledge must be transferred and if done successfully this may improve the performance of learning.

2. RELATED WORK

With the advent of technologies electronic goods are getting cheaper and better. They have become a common necessities. Online Q&A forums have also gained a lot of interest from the user. Recent study on Question tagging and related work involves use of algorithms like Multi-, SVM Label Classification, etc.

Francisco Charte, Antonio J. Rivera, Maria J. del Jesus and Franciso Herrera[1] have the used their own software known as QUINTA i.e. QUestIon Tagging Assistant. Here the software QUINTA uses the algorithm of K-NN(K-Nearest Neighbor). Using this software they have also tried to predict the tags of various question as asked by their user. To accomplish this task, firstly the text of each post is processed to produce a multi label dataset then a lazy nearest neighbour multi label classification algorithmis used to predict the tags on new posts.

Avigit K. Saha, Ripon K. Saha and Kevin A. Schneider[2] they have also predicted tags for the Stack OverFlow Questions and have also tagged this particular question to their respective domain using the algorithm of SVM (Support

Vector Mechanism). Here they have used a discriminative model approach and they have got an accuracy of around 68.47%. They have suggested that predicting the tags for a particular question dynamically can be a very difficult task and have suggested that not all the question may have the same way of predicting the tags as compared to some. So to provide a solution to the problem they have suggested a discriminative model approach with various kinds of solution of various type of domain oriented questions.

Sinno Jialin Pan and Qiang Yang[3] give a basic survey about transfer learning. This paper helps us in understanding the various types of transfer learning. It also helps us to understand that there might be difference in the test dataset and the training dataset and there might be instances that both of them will be the same. It helps in choosing the transfer learning technique by presenting a comparative analysis of different transfer learning techniques.

Jyoti S. Deshmukh and Amiya Kumar Tripathy[4] have used transductive transfer learning to transfer the knowledge of one domain to another. They have used two different approaches wiz. SSKMDA1 and their own proposed method to produce a proper example of opinion mining on Amazon Dataset.

Gerel Tumenbayar and Hung Yu Kao[5] helps us in predicting the domain of the particular question posted. However this paper can be applied only when we have the tags for each and every question. Here they used Kruskal's algorithm to predict the domain of the respective question from the tags already present in the question. Here they have the tags already present in the test as well as the training dataset. With the help of Kruskal's algorithm they have created a Bayesian network from the training dataset and use it to predict the domain of the test dataset.

Fanming Dong, Shifan Mao and Weiqiang Zhu[6] have predicted the tags of a test datasets by training a model created using the training dataset that has got various question from different domains all together. Here tags were only provided in the training dataset and they have predicted the tags for the test dataset. After all the initial pre-processing they have first tokenized the dataset. After the tokenization they have converted each and every word in to vector values using Glove. After converting the words to vector they have formed their model using these words. After the formation of a model they have performed RNN(Recurrent Neural Network) on it. Using this they have predicted the tags of the respective questions.

3. QUESTION TAGGING USING WORD EMBEDDINGS

3.1 Pre-Processing

The dataset available to us in a raw form, i.e. the title, content and the tags in the training dataset is unstructured data. We then process this unstructured data.

i. Cleaning HTML Tags, Punctuations and link replacement: The HTML Tags, Punctuations and the links like 'www', 'http', 'https', etc. are removed.
ii. Stop Words: Stop Words are a certain English words that are not important for the natural language processing.
iii. Tokenization: It is a process of splitting input string into words or tokens.
iv. Word Embedding: For implementing word embeddings, we have used Word2Vec neural network model available in genism toolkit [7].

3.2 Word Embedding

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean [8] proposed CBOW (Continuous Bag of words) and Skip Gram models for generating word vectors.

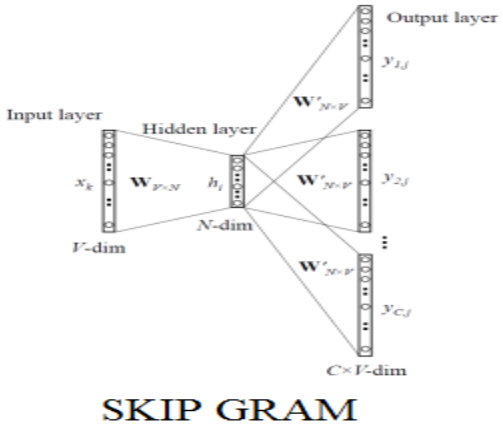

Fig. -2: Skip-Gram

Skip Gram will basically take each and every word in the large corpus and will also take the input of other words one-by-one. These words will surround the corpus within a defined window. These inputs will be then be given to the neural network. After processing it will then predict the probability for each word to actually appear in the window around the focus word [9].

APPROACH 1

We have implemented Skip Gram model and have kept the window size of 5. The training data for this model comprised of all the dataset of ours. We have trained the model with the Cooking, Biology, DIY, Travel and Physics dataset. After the training dataset is loaded the vector space models for each and every word is generated separately.

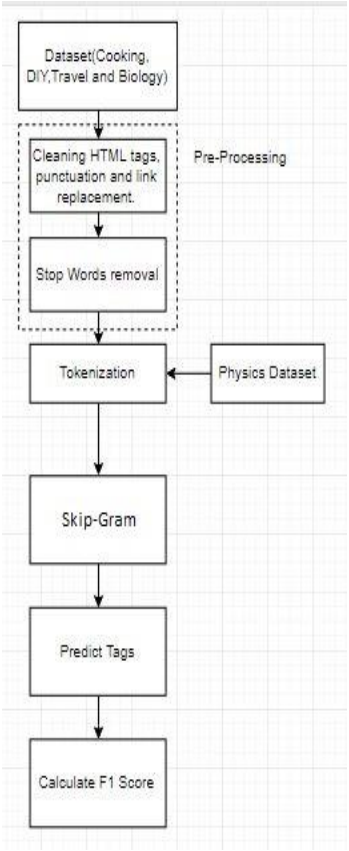

Fig. -3: Approach 1

After training our model we pass the context words as an input to our trained model to get center words which would be our tags. Now we calculate the F1 Score as a measure to test our model[10].

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Fig. 4: F1 Score

APPROACH 2

In this approach we have used the model from the above approach. After predicting the tags from the above models in order to achieve a better F1 score, we have performed lemmatization on the tags predicted.

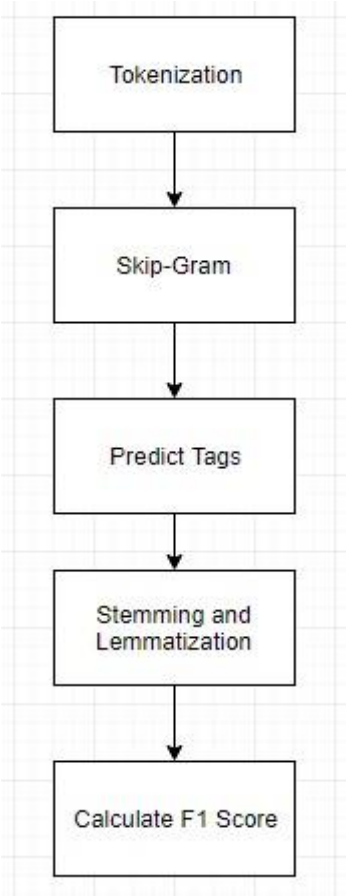

Fig. -5: Approach 2

APPROACH 3

Here we just trained our model using physics dataset and an external source of physics tags [11]. At first we tokenize the physics dataset. After training our model we pass the context words as an input to our trained model to get center words which would be our tags.

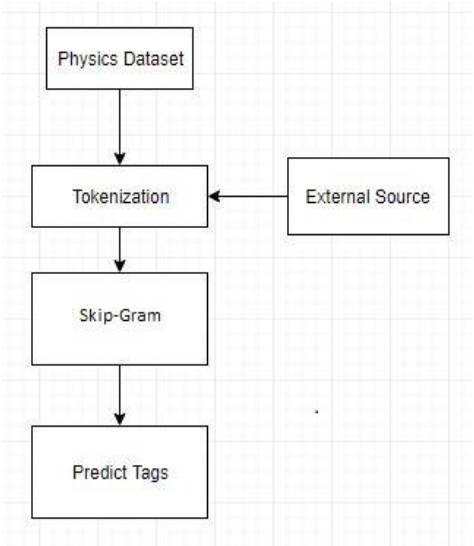

Fig. -6: Approach 3

4. RESULT AND ANALYSIS

F1 score calculated in the approach 1 is given in the following table:

Table. -1: F1 Score of approach 1

Dataset	For First 100 Rows	For First 1000 Rows
Biology	10.9834	131.54
Cooking	35.834	360.13
DIY	40.78	381.96
Travel	11.633	158.883

Here is the graph depicting the above values:

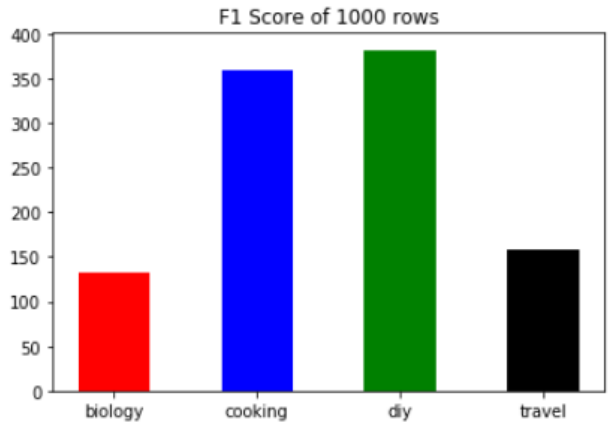

Fig. -7: F1 score of approach 1

F1 score calculated in the approach 2 is given in the following table:

Table. -2: F1 Score of approach 2

Dataset	For the first 100 rows	For the first 1000 rows
Biology	10.9834	131.78
Cooking	35.834	361.38
DIY	41.1167	391.19
Travel	12.71	164.95

Here is the graph depicting the above values:

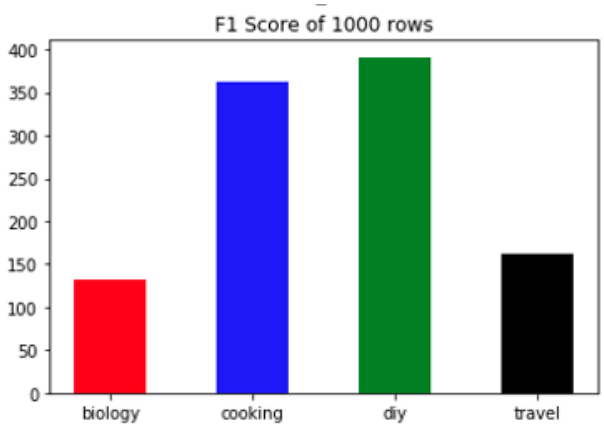

Fig. -8: F1 Score of approach 2

From the above two tables we can state that we get a better F1 Score by using the 2nd approach. Hence after looking at the result and after comparing them we have used the 2nd approach to predict the tags for the physics dataset. After predicting the tags using the 2nd approach we have the following tags:

```
['spin', 'particle', 'electron', 'elementary']
['theory', 'string', 'witten', 'superstring']
[]
[]
['principle', 'uncertainty']
['sound', 'noise', 'wave', 'loud', 'hearing']
['string', 'theory', 'superstring']
['color', 'red', 'blue', 'colour', 'sky', 'dark', 'bright', 'white', 'green']
['energy', 'kinetic', 'conservation', 'mass', 'particle', 'collision']
['physic', 'introductory', 'mathematics', 'mathematical']
```

Fig. -9: Physics tags using approach 2

And now using the approach 3 we have predicted the tags for the physics tags that are:

```
['spin', 'particle', 'charged', 'massive', 'electron']
['theory', 'qft', 'study']
['qft']
['weak-interaction', 'determinism', 'beyond-the-standard-model', 'proton', 'elementary-partic
les', 'stability', 'magnetic-moment', 'fluid-statics', 'ising-model', 'distance']
['classical', 'principle']
['wave', 'speed', 'sound', 'effect', 'source', 'plasma']
['theory', 'string', 'qft', 'brane', 'topological', 'bosonic']
['hole', 'black', 'horizon', 'star', 'red', 'observer', 'white', 'night', 'event']
['kinetic', 'level', 'conservation', 'high', 'internal', 'total', 'free']
['physic', 'math', 'book', 'knowledge']
```

Fig. -10: Physics tags using approach 3

4. CONCLUSION

After implementing the process we can conclude that we can predict the tags using Word2Vec Model. It can help in predicting the tags using the Word2Vec model but the biggest drawback to using this model is that it will not provide a better accuracy as compared to other methods to

do so. One of the biggest reasons behind this is that this model uses a training and test dataset. Sometimes it is possible that the training and test dataset might be of the same domain. And sometimes this may be the total opposite i.e. the training dataset might be a non-technical domain and the test domain might be of the technical domain. But for the current situation we have also used some separate external tags that helps us to determine tags for test domain with a better accuracy.

REFERENCES:

[1] Charte, Francisco, et al. "QUINTA: A Question Tagging Assistant to Improve the Answering Ratio in Electronic Forums." IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), 2015, doi:10.1109/eurocon.2015.7313677.

[2] Saha, Avigit K., et al. "A Discriminative Model Approach for Suggesting Tags Automatically for Stack Overflow Questions." 2013 10th Working Conference on Mining Software Repositories (MSR), 2013, doi:10.1109/msr.2013.6624009.

[3] Pan, Sinno Jialin, and Qiang Yang. "A Survey on Transfer Learning." IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, 2010, pp. 1345–1359., doi:10.1109/tkde.2009.191.

[4] Deshmukh, Jyoti S, and Amiya Kumar Tripathy. "Text Classification Using Semi-Supervised Approach for Multi Domain." 2017 International Conference on Nascent Technologies in Engineering (ICNTE), 2017, doi:10.1109/icnte.2017.7947982.

[5] Tumenbayar, Gerel, and Hung Yu Kao. "Topic Suggestion by Bayesian Network Enhanced Tag Inference in Community Question Answering." 2016 Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2016, doi:10.1109/taai.2016.7880110.

[6] Dong, Mao and Zhu. "Transfer Learning on Stack Exchange Tags." Stanford University.

[7] "Word Embedding." Wikipedia, Wikimedia Foundation, 12 Apr. 2018, en.wikipedia.org/wiki/Word_embedding

[8] arXiv: 1301.3781 [cs. CL]

[9] "Word2Vec Tutorial - The Skip-Gram Model." Word2Vec Tutorial - The Skip-Gram Model · Chris McCormick, mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/.

[10] "F1 Score." Wikipedia, Wikimedia Foundation, 5 Apr. 2018, en.wikipedia.org/wiki/F1_score.

[11]Frequent Words Model | Kaggle, www.kaggle.com/ymcdull/frequent-words-model.

[12] Transfer Learning on Stack Exchange Tags | Kaggle, kaggle.com/c/transfer-learning-on-stack-exchange-tags.