**User Query Identification Stack Overflow**

* **Uniqueness/Novelty of project:-**

In our approach, we will be using tags information for predicting similar information and also will be showing that tags, title tokens, and context of the post plays a significant role in ranking the relevant posts. We will be using OneVsRest Classifier and LinearSVC for predicting the topic of the posts and then assigning higher scores to similar topic scores.

* **Impact on business:-**

A lot of content is present in form of stack overflow questions and answers, various studies point that developers face problems while development life cycles and they ask questions on stack overflow which gets answered by fellow developers across the globe. In order for a new developer to understand a concept or solve an issue, it is very difficult to identify the problems. It involves domain experts in form of experienced software developers. The information present is overwhelming and at times can be too much to handle for a budding developer.

Hence developing a project like this can help:-

1. To identify most relevant questions to a query [text similarity]

2. Identify the matching tags and pick top relevant questions from stack overflow.

3. To identify top k solutions of the problem.

* **The architectural flow of the proposed solution:-**

We used the stack overflow dump data and preprocessed the data to be used as the dataset for training. Train.csv was processed into various text files such as cleaned.txt, tags.txt, clf.txt and multibin.txt . After training the model successfully we used OneVsRest classifier along with TfidfTransformer and countVectorizer to predict tags and recommend questions and top answers based on user query. The predicted results were displayed through a user interface where the query was given by the user.

* **Scope of work: -**

In our project, we extracted topics and combined it with LinearSVC model for recommending similar posts. We used OneVsRest classifier but just considered features like tag length, body length, view count, etc. But it can further be extended by using textual features of the post and combining it with feature vector from TF-IDF and LDA models. It would be interesting to observe the results of a similar post using that model. These models can also be extended to recommend the most relevant questions to a user according to his technical interests. This helps in finding relevant questions for experts that they are more likely to answer.