CSCE 5290 - Natural Language Processing

Classification of the content in StackOverflow

Project Proposal

Project Title:

Classification of the content in StackOverflow.

Team:

Venkata Saran Praneeth Koduru(11440082)

Md Farhad Mokter(11336535)

Goals and Objectives:

- **Motivation**

Out of the many finest forums for discussing programming-related errors and issues in the current environment is Stack Overflow. The user must manually explain the related topic when posing a query on this platform in order to get the issue resolved or directed to the appropriate group of individuals who are working in the similar field. For instance, if we are looking for a Stemming query and we say that the category is Natural Language Processing, it would take some time for the right community to react. Instead, if we used tag stemming, the conversion would start sooner or the search for the solution would be addressed quickly.

- **Significance**

Creating a model that automatically tags the queries, for example, automatically assigning a VLSI query a VLSI tag and a Natural Language Processing problem a Natural Language Processing tag, etc., then this would be really helpful. By doing this, the queries would be rapidly answered because the right community would be involved, and solutions would be found more quickly.

- **Objectives**

Our major objective is to create a model that can automatically tag questions submitted to Stack Overflow. We created a multi-categorization system that automatically tags discussion forums. To build and test our classifier, we use a dataset of Stack Overflow queries.

- **Features**

By locating n-grams and calculating the total frequency of words, the most frequent words—such as "and," "or," "between," and so on—that have no bearing on sentence classification will be eliminated. We wish to use Term Frequency and Inverse Document Frequencies to rebalance the count based on how it contributes meaning to or is relevant to the topic we want to classify.

References:

- https://www.researchgate.net/publication/266850100_A_Hybrid_Auto-tagging_System_for_StackOverflow_Forum_Questions
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. CoRR, abs/1301.3781
- https://www.kaggle.com/competitions/multilabel-bird-species-classification-nips2013/data
- https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff
- https://towardsdatascience.com/auto-tagging-stack-overflow-questions-5426af692904

Github

https://github.com/praneethk6795/NLP_Classification_of_the_content_in_StackOverflow