Capstone Project proposal

Stack Overflow – Tag Prediction

Student Name: Aashik Sujaudeen Sikkandher Babu

Course Name: Machine Learning Track

Cohort: March 2, 2020

Mentor: Ashwin Kumar Kannan

Problem Statement:

Stack Overflow is a largest, most trusted online community for program developers to learn, share their programming knowledge, and build their careers. It is something which every programmer uses one way, or another. Primarily it contains three segments namely Title, Description and Tags.

It is very important and critical for Stack overflow to predict the accurate tags for the question posted based on the title and description of the question posted as the posted question will be routed to appropriate experts or set of people based on the tags to get the appropriate answer with quick turnaround time. Any wrong tagging would drastically reduce the time taken to get the answer to the questions. The wrongly tagged question can also be left unanswered at times. It can also suggest other questions like the posted questions as helpful Q8A threads to bring good customer experience which is again based on the predicted tags. Also, many times, users add wrong tags to their posted question which also need to be taken care when predicting the accurate tag. Any inaccurate tagging would cause a mess and bring down the user traffic to the website. Therefore, tag prediction accuracy rate is business critical for Stack Overflow website.

Data:

Stack Overflow Q & A dataset is planned to download from Kaggle.

Dataset: Sample Stack Overflow Q & A [Size 3 GB & Format: csv]

Dataset with the text of 10% of questions and answers from the Stack Overflow programming Q&A website.

This is organized as three tables:

- Questions contain the title, body, creation date, closed date (if applicable), score, and owner ID for all non-deleted Stack Overflow questions whose Id is a multiple of 10.

- Answers contain the body, creation date, score, and owner ID for each of the answers to these questions. The Parent_Id column links back to the Questions table.

- Tags contain the tags on each of these questions

Project Approach:

- The tag prediction problem that we are planning to solve is a supervised problem as our approach will use ‘Title’ and ‘Question’ as inputs which means labelled input and predict the ‘Tag’ as output by training the model with 80% data in the above dataset and testing the model with 20% data of dataset.

- As the approach will go through question and title classification, it is expected to use supervised classifiers to solve the problem.

- We may use regression algorithms in case the project is extended further on other user cases related to Stack Overflow like predicting whether the question will be upvoted or not, predicting how much time a question can take to get answered, etc.

Project Deliverable:

‘Stack Overflow - Tag Prediction’ application that will be deployed as a web service API. This section will be updated further upon progressing further in the ML track course.

Computation Resource Requirement:

The computation resource required for this capstone project will be updated on reaching computation resource calculation chapters and after discussing with the mentor.