**EE-628**

**Final Project Proposal – Rohan Chandra**

**Autonomous Tagging of StackOverflow Questions**

**Problem Statement**

* Given a dataset with the text of 10% of questions and answers from the Stack Overflow programming Q&A website, the goal is to implement an autonomous tagging system that will come up with appropriate tags for each new question
* Tasks:
  + Clean the data
  + Figure which model to use and appropriately preprocess the data for the model
  + Adjust the parameters, complexity and hyper-parameters of the model to produce the most accurate model
  + The results of this accurate model would be tag suggestions for each question it encounters

**Description of Data Set**

The dataset consists of 3 CSV files with the following formats:

1. Questions:
   1. Contains 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.
   2. Has the format: Id, OwnerUserId, CreationDate, ClosedDate, Score, Title, Body
2. Answers:
   1. Contains the body, creation date, score, and owner ID for each of the answers to these questions. The ParentId column links back to the Questions table.
   2. Has the format: Id, OwnerUserId, CreationDate, ParentId, Score, Body
3. Tags
   1. Contains the tags on each of these questions
   2. Has the format: Id,Tag

**Implementation Plan**

1. Data Preprocessing and Cleaning, with customization for each model being used.
2. Model Training,Testing and Optimization(for each type of Model separately).
3. Evaluating results.
4. Choosing the best Model/Models.
5. Consolidating the Results.