Project Title: Tag Recommendations on StackOverflow

Team members:

Abhijith Shreesh (1213204276), Channabasava Gola (1213222619), Jagdeesh Basavaraju (1213004713), Namrata Nayak (1212408767), Abishek Ravichandran (1213090253), Sahan Vishwas (1213094049), Siddhant Sudan (1212890391), Shoor Veer Singh (1211334810).

Motivation:

We always find ourselves on sites such as StackOverflow and FreeCode for solutions, when we stumble upon a technical hiccup. The tags we associate with questions are generally the driving factor in getting responses. To aid this process of associating the right tags with questions on StackOverflow, we want to come up with a tag recommender system.

Project Description:

For tag recommendations on StackOverflow data, we implement an ensemble of the following techniques:

- Predicting a ranked list of tags for the content using Multinomial Naive Bayes classifier.
- Generating a ranked list of tags in the order of similarity to the content in consideration.
- Recommending tags based on the affinity of tags to terms in the content.

Possible milestones and corresponding roles:

- Going through papers/literature survey: 1st February to 23rd February.
- Collection of data: 23rd February to 3rd March.
- Model and feature selection: 4th March to 25th March.
- Finalizing data model and storage technologies: 25th March to 5th April: Continuous evaluation and reiteration of models involving everyone on the team.
- Scope updation checkpoint (5th April). Decide on additional tasks which can be part of the project scope.
- Code Implementation: 6th April to 22nd April
  - Data pre-processing of raw data: Channabasava, Abhijith.
  - Feature selection: Abhishek, Shoor.
  - Model implementation: Jagdeesh, Namrata.
  - Model Tuning: Sahan, Siddhant.
- Bug fixing, module tuning, testing, analysis and improvements if required: 22nd April to 28th April: Everyone to be involved.

Assumptions:

- We will be using Naive Bayes model in the process and we assume that tags that we consider are independent of one another.
- The training and testing set is different and the two are strongly correlated.
- The order in which the data is fed to create the model does not matter. All data is treated in the same manner.

Expected Results:

- Based on the ensemble technique mentioned, untagged content will be assigned suitable tags.
- Compare and contrast the results and accuracy of different machine learning algorithms.

Paper List:

– Tag Recommendation in Software Information Sites. Xin Xia*‡, David Lo†, Xinyu Wang*, and Bo Zhou*§ (http://ieeexplore.ieee.org.ezproxy1.lib.asu.edu/stamp/stamp.jsp?tp=&arnumber=6624040&tag=1)