What Defines a Good Stack Overflow Answer Post
An Automated Post Rater Based on NLP

Project Category
Natural Language

Group Members:
Yanpei Tian; yanpeit@stanford.edu 06079698
Yanhao Jiang; jiangyh@stanford.edu 006340964
Chunyue Wei; weicy18@stanford.edu 06288308

Motivation
Major question and answer sites for professional and enthusiast programmers, like Stack Overflow, have gained phenomenal popularity during recent years. Programmers are relying on these websites to seek inspiration, find solutions, and help others. However, as all of the common open online communities, the level of professionality, correctness, and sophistication varies among different posts and users. In this project, we want to identify the good posts, particularly the answer posts, that received most upvotes. A complication of related features could also potentially serve as a guide to people about how to write better (more helpful and interesting) answer posts on such Q&A sites, like Stack Overflow.

Method
We plan to apply supervised learning to the dataset of Stack Overflow. We will define features that we think may be helpful to identify better answer posts and extract them from the dataset of Stack Overflow. These extracted features (e.g. length, technical keywords frequency, references, post time, and user profile, etc.) will be the input of our training model. The label of our dataset will be measured as a function of a post's upvotes and downvotes. The training set will only include posts before a certain date and we plan to use later posts to test out implementation. Through the training, we will explore the relationship between the features and the label, identify the features that lead to a better post.

Intended experiments
We plan to first start with the regression algorithms to identify the features that are more likely to contribute to our model. After we get a reasonable understanding of the problem, we would move forward to consider more sophisticated models, such as SVM and neural networks, to obtain a better solution to our problem.

Evaluation
After we get a satisfactory feature extractor and a trained model, we will use the model to predict later posts' ratings (measured by upvotes and downvotes, etc.).

Source of Data
https://data.stackexchange.com/stackoverflow/query/new

On this website, we can use SQL to easily get the data we want from Stack Overflow.