CS534 Project Proposal
Yaonan Zhong
May 16, 2014

Title: Predicting Star Rating base on Yelp User Review Text

Yelp ratings brings us a new way to choose a business as a customer and run a business as an owner in our daily life. We prefer restaurants or hotels with higher ratings which determine our choice most time. However, we know that not all user ratings are objective all the time. It is possible that a very positive review may come with a five-star rating while another similar one just has a three-star rating. And sometimes we may not have enough time to read all the reviews before making a decision. So can we learn a model to predict the rating for a review text? The underlying goal here is to attenuate the effect of subjective reviews by learning rating from a large number of examples. Note that we can consider subjective ratings as noises in our learning model.

The Yelp dataset we will work on has information on businesses, reviews, users and check-ins. We will focus on all of the restaurant reviews. In learning model selection we plan to use Navie Bayes model, decision tree, and support vector machine as our choice, and we can compare the performances between different models. Since we are doing document classification, one of the important thing is feature selection. How do we construct our feature so as to obtain best performance? That would also be a part of our research. We plan to use k-fold cross validation in our training. We will also estimate the sufficient number of examples for PAC learnability. If the time is allowed, we will try to apply Latent Dirichlet Allocation in topic discovering.