Project Proposal

Xinyue Zhang(xz2139), Hanlu Zhang (hz1625), Yiyan Chen (yc2462), Ben Zhang (bz957)

We will use the yelp dataset. The business problem we are trying to solve here is to find out which yelp review contains useful information and recommend it to users. This is important because the order of reviews that displayed on the first page of a business will certainly affect the most. When someone writes a new review, the website will need to know whether it contains useful information and determine where to put it. And if the website can successfully recommend the reviews that contain most useful information, it will increase the reliability of the website and attract more users. There are a lot of factors that may affect the usefulness of a certain review. In our dataset, we have a mix of information on the specific review, the business and the user. The content of the review(ie. text & pics), the user who wrote it, the time it came out, the type of the business etc. are all potential factors that may influence whether people think it is useful or not.

The supervised data mining problem here is to build a predictive model which gives the best accuracy of the prediction on whether a newly written yelp review will be useful or not. In the dataset, we know how many times a review was chosen as useful from actual users, hence we will convert that to binary using a threshold and use that as our target variable. After cleaning up the dataset, we will try different kinds algorithms, including logistic regression, decision tree classifier and SVM to find the "best" model from the training set and test its accuracy. Each review in our dataset will be our data instance. We want to limit our instance to just the review on restaurants in the US for minimizing the number of outliers. Overall, the prediction model should give out a better understanding of what makes a review useful and when a new review came out, whether we can recommend it. This will not only benefit the user, but also improve the performance of the website.