**Review Rater**

Joseph Finnegan, Joe Martinez

**Overview**

The idea of this project is to build a data mining program that will determine whether or not a particular yelp review has the attribute of being useful. This could be useful in through the fact that many of the reviews contained within the corpus provided by yelp provide few details as to why the review was given a particular star rating. It’s also a safe assumption that the people would much rather see a review on a yelp page that is more useful and relevant than one that simple has a star rating without any particularly relevant information in the text of the review.

**Tasks**

The data mining task performed in this project will be taking the information given in a yelp review, and predicting/determining whether or not that particular review should have the attribute of being “useful” using the given information.

**Deliverable**

At the end of the semester, we plan to deliver a program that can interpret the corpus of yelp reviews and predict with a percentage of accuracy whether or not that review is useful, or has been deemed useful by a number of users.

**Challenges**

The primary challenge of this project is that finding a way to interpret the natural language used within each review could be extremely tedious through the simple fact that it IS natural language. Right now we are thinking that there should be key words that we can pick out in order to determine whether a review is useful, or perhaps words that indicate that a review is not useful. There is also the length of the review to be considered, and likely a good number of other factors that we have not thought of yet. We will address these problems through simple experimentation and logical deduction of how to go about interpreting each review.

**Evaluation**

The efficacy of the solution could be evaluated by the accuracy of the results given by our application as compared to the data given on whether or not a review has been rated as useful. The primary problem with this is that many reviews may be useful, but have not had the chance to be seen by yelp users. Without being seen by a sufficient number of people, a comparison of our results against the number of positive AND negative results of whether a review is useful would be unfair, so it would be best to do an accuracy comparison between our results and ONLY the positive review results from the corpus. (If a yelp review has been deemed useful, then our review should also be deemed useful.)

**Partitioning**

The project will be partitioned through the fact that each of us will explore separate avenues of data analysis. We will likely write the majority of the code together, but we will each come up with our own tweaks that will result in a difference of accuracy in prediction. We will pick out and combine the methods that have the most success in order to make a more functional program for mining this information. Also, documentation and optimization of the code base will be assigned out to either of us.