Ruchin Patel
Email: ruchinpa@usc.edu

Project proposal towards EE 503 (MW 8:00 - 9:50)

OVERVIEW

Today large amount of information is available online. The information being reviews for books, tweets on twitter, news on news website, blogs etc. All of these information have one thing in common and that is they are written texts. These written texts convey the emotions or sentiments of the writer who wrote them . For instance:

1.) The reviews of books give us some information about how well the characters are written, if the storyline is novel in nature or whether it all comes together coherently.
2.) The tweets on twitter regarding a certain topic like the United states presidential election gives us information how was a candidate's speech received, whether a recent revelation about the candidate affected the view of the masses of that candidate.
3.) Several news article on a particular subject from different news website would let us know about the gravity of the situation of that particular subject.

All these various degrees of information obtained from the online sources are in one way or the other used to classify the concerned subject matter in a some single concrete domain which can be anything like, positive or negative approval ratings, a good or a bad book, a rainy or a sunny day etc.

GOAL

All this information needs to be first digested by the person reading the information before the human mind can perceive how to classify that information. However, there are times when the reader wouldn't necessarily need all the details in the text, but would rather just need actionable information. Sentiment Classification is the process of determining the emotional tone behind a series of words; gaining a very succinct understanding of the opinions expressed within a vast pool of information.

In this project, we will attempt to apply Sentiment Classification to reviews of various books. This will be accomplished by using Naive Baye's Machine Learning algorithm. We will be using the 'Amazon Books Review Data Set' from the UCI Machine Learning Repository. There are 213335

reviews in this dataset for different books. We will be training our model on the subset of these reviews and then finally we will be testing our model on the remaining subset.

Our Naive Bayes algorithm would classify a new document into a particular class by estimating the product of probability of each word occurring in the document given a particular class(likelihood), which is nothing but the conditional probability of a word in a document given a particular class. This conditional probability is then multiplied by the particular class prior. After calculating the above for all classes we would would select the class with the highest probability.Furthermore we will classify each particular review in the test set stating whether that particular review is good or bad. We will then compare this result with the actual value in the test set to determine the accuracy of our algorithm on review-per-review basis. Also, we will judge from the test set whether each individual book is readable or not. We will then compare this information with the net result computed from the predicted results of our algorithm to find its accuracy on a book-per-book basis as well.