CS361 Assignment5

In this assignment, we need to improve the prediction accuracy of the Naive Bayes algorithm, and we know the formula of the algorithm is P(Y|X) = P(X|Y)*P(Y)/P(X), these three probabilities in algorithm are likelihoods, prior probability and posterior probability separately. Concerning about the likelihood(P(X|Y)=P(word|A)/P(word|B)/P(word|E)/ P(word|V)), we need to calculate the sum of each high frequency of occurrence word's probability(p(word1|A)*p(word2|A)*p(word3|A)*p(word4|A)…=sumP(word|A)) for different 4 classes with A, B, E and V in training datasets. In relation to the prior probability, we can directly calculate the probability by taking the number of corresponding classes dividing the whole number of all classes still in training datasets. And then, by using P(class|data) = P(data|class)*P(class) we can get four posterior probabilities which related to four classes, they are P(A|word), P(B|word), P(E|word) and P(V|word), the maximum P value will determine the class of the abstract.
Since the Naive Bayes algorithm doesn't require as much training data to predict and can handle both continuous and discrete data, we can obtain better predictive accuracy by preprocessing the given training and testing data. In the training data, we can reduce the amount of datasets or omitting the redundant data such as deleting the data is not string or exceeding 9 letters or less than 4 letters, then acquire a list with 1000 the most common words and utilize in calculating probabilities, this can not only save run-time but also lessen the possibilities of overfitting. As for the testing datasets, we can make a classification for each abstract's word and find each word occurrence numbers with reference to the 1000 words list to predict class accurately.