1. Introduction

The project aims for the experimental evaluation of the linear SVM's and nearest neighbor algorithm data set having many irrelevant features. The task is to learn which Reuters articles are on corporate acquisitions. To cater this requirement each features represents the frequencies of the words. The training and validation data set has 300 samples each having 20,000 features that include only 9947 actual features and rest of all are noise.

2*. Filter Methods*

The filter methods that can be directly used on the dataset is

a) Pearson Correlation Coefficient
b) Signal to Noise Ratio
c) T Test

Mutual Information and Chi-Square methods can be neglected because most of the features are 0 and some others are having a considerable values which could make the methods based towards the high parity features and it well known that these methods are good only at nominal features which is not the case here. Whereas, the above stated three methods are well suited because the data set is sparse and they try to measure how clustered the data points are so that we could rank and neglect the noise.

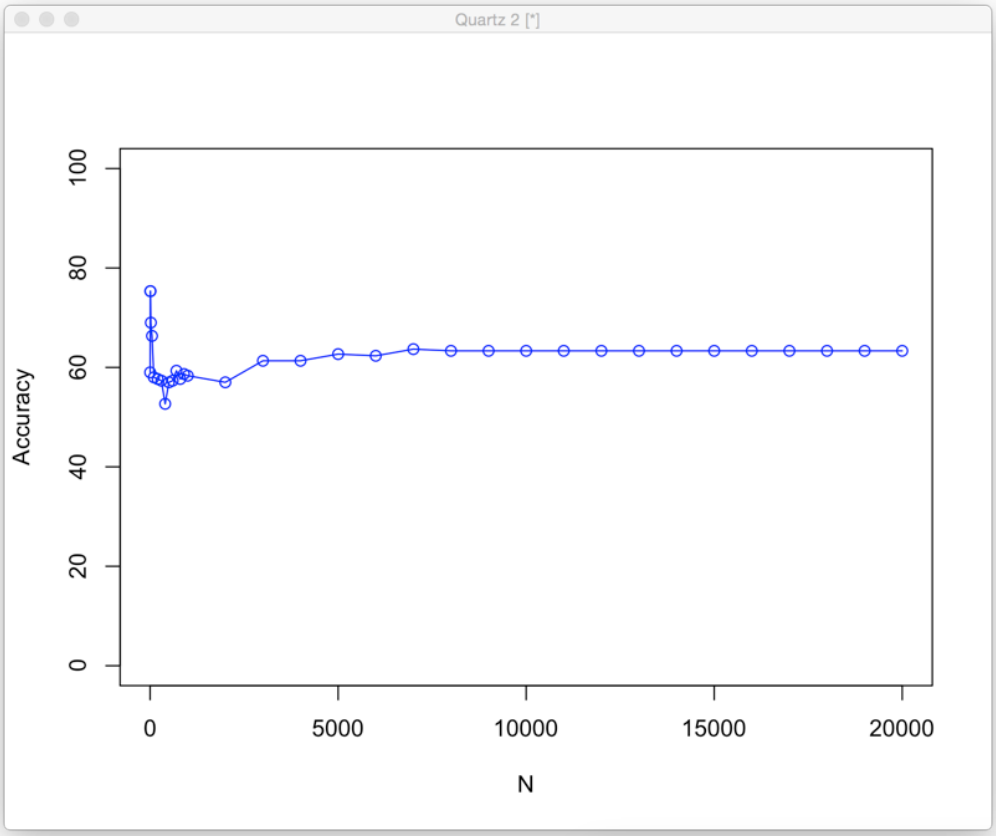
3. Linear SVM Implementation and Discussion

The three filter methods are implemented in java to compare and analyze which one works better on this data set. The algorithm's performance has been tested by taking multiple values for N starting from 1 till 20,000 on sparse intervals on samples of size 300 each for train and validation data. The features are ranked with Linear Kernel SVM having c=1 using SVM Light package.

All the results that are captured here are after normalization of the features.

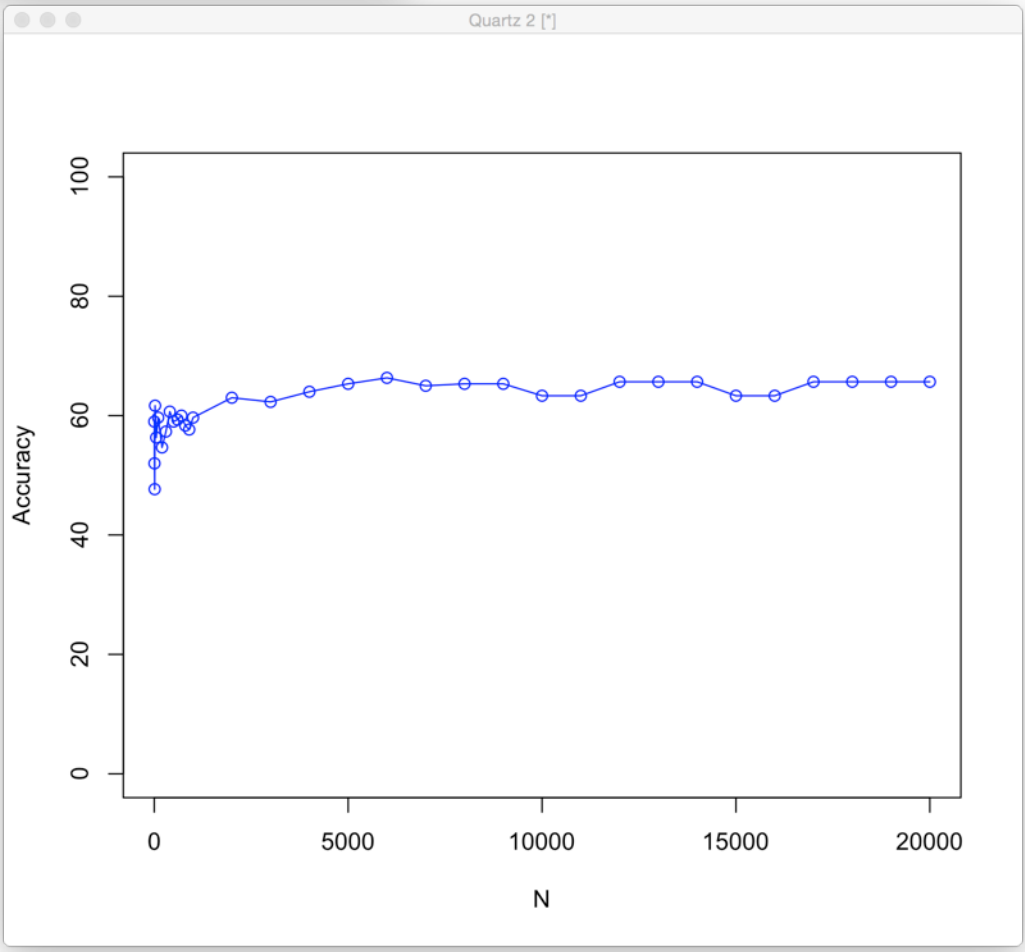
Using Pearson Correlation Coefficient, The best accuracy of 75.33% is obtained when top ranked 10 features are chosen on all of the samples. It is also observed that accuracies are decreased as the number of the features are increased because there are more irrelevant features.

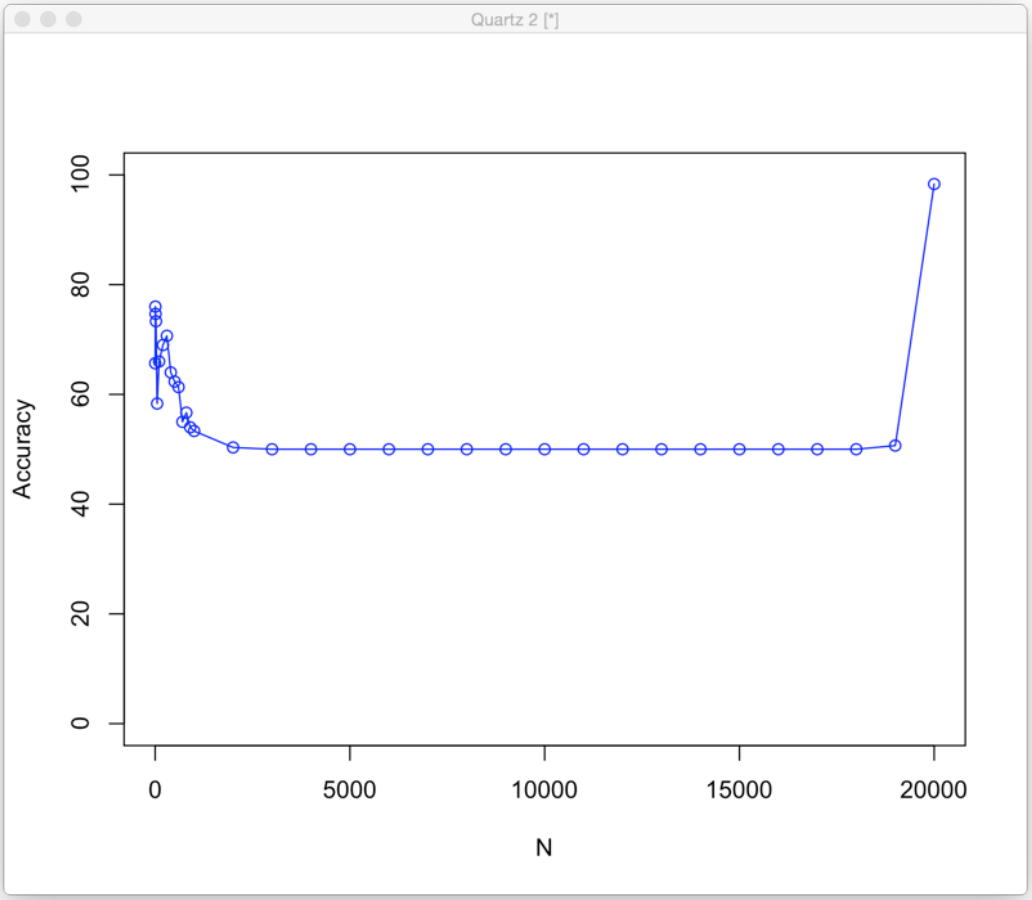
Using Signal to Noise Ratio, The best accuracy of 66.33%is obtained top 2000 features are chosen from all of the samples. For this method its observed that as the no of features are increased the accuracies doesn't change much.



Plot of N Values vs Accuracy for Pearson Correlation Coefficient Filter method.

Plot of N Values vs Accuracy for Signal to Noise Ratio Filter method.





The Plot of N Vs Accuracy of T Test marks a highest accuracy of 98.33% for 20,000 features. This accuracy clearly shows that that T Test might not be a good performer for feature selection on this data set using Linear SVM

I would like to say that best SVM-filter-N for this data set could be Pearson Correlation Coefficient with 20 features/sample as it has got the best reasonable accuracy among all.