Danhua Yang
SID: 010769681

Rank: 11; F-1 score: 0.7429

Approach:

The “train.dat” file is read line by line using “.readlines()” function. The first element of each line is saved separately as into the “res” list; the rest element of each line are saved in the “doc” list.

For the “test.dat” file, the entire file is read line by line using “.readlines()” function. And because it does not contain the class result, it only needs to split each line and save into “test” list.

After having the “doc” list and “test” list ready, they both are converted into two separated sparse matrix.

In my best result, the “ExtraTreeClassifier” tree-based feature selection function from skLearn library has been used to reduce the feature number to about 600 from 100000. Then the “RandomForestClassifier” function from skLearn library has been applied to predict the result for each test data. The best F-1 score I got is 0.7429.

Methodology:

Before come to the final best result, I tried Chi2, Mutual info classifier, Extra tree classifier and TruncatedSVD packages from skLearn to perform the feature reduction.

For the classifiers, I’ve implemented K neighbors classifier, Decision tree classifier, Random forest classifier, Adaboost classifier, Naive bayes classifier, Neural network classifier, and Support vector machine. All the functions and packages used for these classifier are imported from sklearn.

To determine the best combination of feature reduction method and classifier, I fixed the feature reduction method then run all the classifiers through the cross validation to check which classifier has the best F-1 score. As the result, the Random forest gave out the best F-1 score among all other classifiers.

After choosing the Random forest as the classifier, I tried select best K features feature reduction function with chi2 and mutual info classifier with K from 5000 reduced

to 200 features. And I also tried with tree-based feature selection which the algorithm reduced the number of features to around 600.

After determined the combination of tree-based feature reduction functions and Random forest classifier. I tried different maximum tree height from 2 to 12 with increment of 1. And the maximum tree height of 10 gave out the best result.

With fixed maximum tree height of 10, I then tried to change the number of n_estimators which is the tree numbers in the forest. And with 200 as the n_estimator gave out the best result.

After tried all the combinations through cross validation. The tree-based feature selection function followed by random forest classifier gave out the best F-1 score result which is 0.7429.