Project Report – part2

Member: Mingxin Lu Diwen Hu

1. the accuracy of your classifiers, and description of the method you used to measure this;

In this project, we use two classifiers in this lab, which are Naïve Bayes and K-nearest neighbors. We construct a list of TOPIC words and deleted the articles which don't have any topics. Finally, we use 11367 articles as our data set. We use TF/IDF as our important metric. We choose frequency words as our feature vectors.

For K-nearest neightbors, since K-nearest neighbor method doesn’t need train data set. We default choose 5 nearest neighbors. We use cosine algorithm to calculate the distance between two vectors.

For Naïve Bayes method, we firstly remove the most and least frequency words in the articles since they are unnecessary vectors. Based on our metric (value of TF/IDF), we choose the attributes that has highest value of TF/IDF. Then we choose 100 attributes to train the data set.

We have total 11367 articles, we use 8557 articles as training datasets and use 2810 articles as test articles for both classifiers.

The accuracy of K-nearest neighbors (with 100 attributes) is 0.639857651246

The accuracy of K-nearest neighbors (with 500 attributes) is 0.662277580071

The accuracy of Naïve Bayes (with 100 attributes) is 0.65978476868

The accuracy of Naïve Bayes (with 500 attributes) is 0.691814946619

1. the offline efficiency cost (time to build model) for each classifier using each feature vector;

Using KNN method:

The time to build model is 28.95 seconds

Using Naïve Bayes method:

The time to build model is 29.38 seconds

1. the online efficiency cost (time to classify) for each classifier using each feature vector;

Using KNN method:

The time to classify is 2 mins 31 seconds.

Using Naïve Bayes method:

The time to classify is almost 1 second.