CSCI 720 Big Data Analytics

Assignment 5

In this assignment, you will use a Naïve Bayes classifier on the 20 newsgroup data set.

Data

Download the dataset in 20news-bydate.tar.gz from http://qwone.com/~jason/20Newsgroups/. If you are using scikit-learn, the dataset is already available to you.

```
#read data
newsgroups_train = fetch_20newsgroups(subset='train', categories=categories,
                                      shuffle=True, random_state=42,
                                      remove=remove)
newsgroups_test = fetch_20newsgroups(subset='test', categories=categories,
                                     shuffle=True, random_state=42,
                                     remove=remove)
```

The data set consists of about 1000 articles in each of the 20 newsgroups split roughly into 60% training and 40% test set.

For starters consider only the following 4 categories:

```
['alt.atheism', 'talk.religion.misc',
 'comp.graphics', 'sci.space']
```

Once you are satisfied with your implementation and results, you can consider the entire dataset.

Preprocessing

You need convert the data into a bag of words representation. You need to remove stop words and convert the corpus into a document matrix (sparse representation to save on memory). Consider using the `sklearn.feature_extraction.text.HashingVectorizer`. You may wish to play with the size of the vocabulary. The performance of your classifier will depend quite a bit on your preprocessing.

Analysis

Implement a Naïve Bayes classifier like we discussed in the class or use Multinomial Naïve Bayes implementation from `sklearn.naive_bayes.MultinomialNB`. Train your classifier on the training data and evaluate performance on test dataset. First test your implementation on the small 4 category data before trying it on the entire data.

Results

Answer the following for the small (4 category) dataset as well as the full dataset

1. What is size of your vocabulary? What was the average sparsity of your training data (average number of unique words in a document as a fraction of size of the vocabulary).
2. Play with the smoothing parameter (alpha). How does it change the accuracy of your classification?

3. Plot the confusion matrix (either a plot or a table). Which group achieved the best accuracy? Which was the worst? Why? Discuss your results

Include your results (answers to questions above) as well any code you developed with your submission.