CS 780 Assignment 2
Due 11/14 at 11:59pm

Download the Reuters datasets from the website. This is a subset of the Reuters-21578 dataset (the original dataset can be found here: http://www.daviddlewis.com/resources/testcollections/reuters21578/). There are 8 classes, and the dataset is split into the training set and the test set. The dataset is already pre-processed (e.g. it's been stemmed by using the Porter Stemmer; short words and stop words were removed). The training set contains 5485 documents, and the test set contains 2189 documents. Both the training and the test sets are text files containing one *document* per line. Each document begins with its class label, followed by a TAB character and then a sequence of "words" delimited by spaces, and ends with the word "reuter". In this exercise, you will compare two of the classifiers that we discussed in class: Naïve Bayes and k-NN (using k = 9).

1. Read the files and convert the documents into the bag-of-words representation. Construct a doc-by-term matrix using the training set. For simplicity, you can use raw frequency in the matrix.
2. Implement Naïve Bayes and k-NN classifiers. Note: GMU Honor Code will be strictly enforced.
3. Classify the test documents using Naïve Bayes and report the classification accuracy.
4. Now classify the test documents using k-NN and report the classification accuracy.

Class	# correctly classified documents by Naïve Bayes	# correctly classified documents by k-NN
acq		
crude		
earn		
grain		
interest		
money-fx		
ship		
trade		
Total		
Accuracy (% correctly classified documents)		

Extra Credit: Use tf-idf for term frequency. Report the classification accuracy by constructing another table similar to the one shown above.

Turn in: In addition to the table above, you will also need to turn in all your code and classification results (i.e. predicted class labels vs. true class labels) for the first 100 documents.