In this problem, we will use the naive Bayes algorithm to learn a model for classifying an article into a given set of newsgroups. The data and its description is available through the UCI data repository.

We have processed the data further to remove punctuation symbols, stopwords etc. The processed dataset contains the subset of the articles in the newsgroups rec.* and talk.*. This corresponds to a total of 7230 articles in 8 different newsgroups. The processed data is made available to you in 5 equal sized splits of size 1446 each with each row representing one article. Each row contains the information about the class of an article followed by the list of words appearing in the article.

A) Implement the Naive Bayes algorithm to classify each of the articles into one of the newsgroup categories. Perform 5-fold cross validation (train on each possible combination of 4 splits and the test on the remaining one). Report average test set accuracies.

• Make sure to use the Laplace smoothing for Naive Bayes (as discussed in class) to avoid any zero probabilities.

• You should implement your algorithm using logarithms to avoid underflow issues.

• You should implement naive Bayes from the first principles.

B) What is the accuracy that you would obtain by randomly guessing one of the newsgroups as the target class for each of the articles. How much improvement does your algorithm give over a random prediction?

Note:

1. You can refer to: http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

2. See attachment gda.pdf for Laplace Smoothing.

3. This is a single-person assignment.

4. Your submission should include (i) Your code, (ii) A doc containing answers to questions (A) and (B). Please upload your zip file as <entry_number>.zip