Assignment 3
Sentiment Analysis

Your goal for this homework is to perform **Sentiment Analysis**: classifying movie reviews as positive or negative. Recall from lecture that sentiment analysis can be used to extract people's opinions about all sorts of and at many levels of granularity (the sentence, the paragraph, the entire document). Our goal in this task is to look at an entire movie review and classify it as positive or negative.

You will be using **Naïve Bayes** with Laplace smoothing. Your classifier will use words as features, add the logprob scores for each token, and make a binary decision between positive and negative. You will also explore the effects of stop-word filtering. This means removing common words like "the", "a" and "it" from your train and test sets. A stop list with the starter code is provided in the file:

`data/english.stop`

Your will have to train a **Naïve Bayes** classifier on the **imdb1** data set provided with the starter code. The starter code comes already set up for 10-fold cross-validation training and testing on this data. Recall that cross-validation involves dividing the data into several sections (10 in this case), then training and testing the classifier repeatedly, with a different section as the held-out test set each time. Your final accuracy is the average of the 10 runs. When using a movie review for training, you use the fact that it is positive or negative (the hand-labeled "true class") to help compute the correct statistics. But when the same review is used for testing, you only use this label to compute your accuracy. The data comes with the cross-validation sections; they are defined in the file:

`data/poldata.README.2.0`

Your first task is to implement the classifier training and testing code and evaluate them using the cross-validation mechanism. The **expected accuracy** is 0.8165 (testing with all words).

Next, evaluate your model again with the stop words removed (set nb.FILTER_STOP_WORDS = True). Does this approach affect average accuracy (for the current given data set)?

To ensure that your code works properly, you should limit your changes to addExample() and classify(). You're free to add other elements further invoked from addExample() or classify(), but you will be evaluated only on these two methods, so you cannot rely on anything added elsewhere.