For this homework you will implement a naive Bayes classifier, and compare its performance on classifying spam and non-spam with that of the decision trees you used in Homework 2.

1. Using the same training set of e-mail messages that you used for Homework 2, choose a set of features that you think will be suited for a naive Bayes classifier. Write a program to extract these features, and create a new training set using these features. (We will discuss in class what kinds of features are well-suited.) Use this training set to implement a naive Bayes classifier that classifies any new example. You may use any high-level language to write this code. You may re-use some or all of your features from Homeworks 1 and 2 if you feel they will be appropriate. Explain what your features are, and why you think they will be appropriate for a naive Bayes classifier.

2. Use this same training set to construct a decision tree, using ID3.

3. Create a new test set with your chosen features, extracted from the test set of e-mail messages given for Homework 2. Run your naive Bayes classifier on all the examples in this new test set, and run your decision tree on the same test set. Compare the classification accuracy of the two algorithms on the test set. Does the naive Bayes classifier perform better, worse, or the same as the decision tree? Explain briefly (1-2 paragraphs) why you think you got these results.

4. Now, introduce noise into your training set by changing the class of a randomly chosen 5\% of the training examples. Construct a new naive Bayes classifier using this noisy training data. Also construct a decision tree with ID on this same noisy training data. Run the resulting classifiers on the test data. How does the accuracy on the test data compare between these two classifiers and with that of your original results (from part 3)? Explain, in one or two paragraphs, why you think you got these results.

Here is what you need to turn in: **Hard copy**: Your answers to the questions above. **Electronic**: Your (well-commented) code for implementing the naive Bayes classifier, instructions for running it, and the data sets and test set you used in the experiments above (in the form of feature values and classifications, not e-mail messages). In short, everything that is needed to run your code and replicate your results.

Your homework will be graded based on (1) your answers to the questions given above; and (2) the correctness and clarity of your naive Bayes classifier code.