Problem 1.

Threshold = 0.01: Tree size = 197, Accuracy = 74.97%

Threshold = 0.05: Tree size = 250, Accuracy = 74.43%

Threshold = 1: Tree size = 9780, Accuracy = 68.02%

1. We can observe that the tree size increases as the threshold increases. When the threshold is 1, the tree is the full tree and has the largest tree size. However, as the threshold increases from 0.01 to 0.05, the accuracy is still similar. When the threshold becomes 1, the accuracy drops 6% for overfitting.

2. Both threshold = 0.01 and threshold = 0.05 work very well in this dataset. They all have small tree sizes and good accuracy. The tree stopped at a small threshold and did not cause overfitting. A high threshold might cause the overfitting of the decision tree.

Problem 2.

We keep a spam dictionary and a ham dictionary to get P(w|spam), P(w|ham). We also calculated prior P(Y) by counting the total number of spams and hams.

When classifying a new email, we apply apply Naive bayes with log, namely, $P(Y|x1,x2,..) = P(Y)*\Sigma\log(P(xi|Y))$. We also used Laplace smoothing with smoothing parameter alpha = 1. Finally we get the posterior and keep the P(Y=spam|W), P(Y=ham|W) and keep the bigger one as our predicted result.

For alpha = 0.1, or 1, the accuracy would be 77.8% in my program.

For alpha = 10, the accuracy would be 68.6% in my program.

For alpha = 100, the accuracy would be 42%

Therefore, alpha = 0.1 or 1 works best for my Naive Bayes learner.