Program Descriptions

a. Computing Features
   - The spam classifier first creates a lexicon of words by including every word in every text file within a specified parent folder (all the emails).
   - If a value for **k** is specified, words seen less than k times are dropped from the lexicon.

b. Training
   - The logarithms of both priors and likelihoods are calculated and stored.
   - Priors are calculated from the proportions of spam and ham in the training sets.
   - Likelihoods are calculated for both spam and ham for every word in the lexicon.
     - Each time a word is encountered in a training email, the running tally for that word in that class of email is incremented.
     - The likelihoods are calculated as the tally for that word in that class over the total tallies for that class.
     - If a value for **m** is specified, tallies begin at **m**.

c. Testing
   - Classes are assigned to each email in the test sets using a MAP decision.
     - $\text{Class} = \text{argmax}_c \text{lgP}(c) \sum \text{lgP}(w_i|c)$
   - Because logs are used, the class with the least negative sum is assigned.

d. Measuring Performance
   - A confusion matrix is made of all the assignments in the test set.
   - Overall accuracy is calculated as the ratio of correct assignments over both classes.
   - Spam accuracy is the ratio of spam correctly assigned to spam.
   - Ham accuracy is the ratio of ham correctly assigned to ham.

Results

Tuning on the training set (see extra credit for description) resulted in default values of **m = 0.5** and **k = 5**. With these parameters:

- Overall Accuracy: 92.5%
- Spam Accuracy: 91%
- Ham Accuracy : 94%

Note that **ML classification and MAP classification yield the same results for this set**. The given training set includes equal ratios of spam and ham, so **the priors for spam and ham are equivalent**. Therefore ignoring the priors will not change the assignment.

Examples

- Spam Identified Well: 0195.2004-01-12.GP.spam.txt
  - This email is an advertisement for penis enhancement. The content and words in this email are likely never going to be found in ham.
- Ham Identified Well: 2966.2000-11-29.farmer.ham.txt

  - This is a personal email between coworkers. The use of “I,” “me,” and “you” was probably a major contributing factor to the ham identification.
- Spam Misidentified: 3110.2004-12-08.GP.spam.txt
  - This email is a nonsensical series of “two-dollar” words. These words are unlikely to be found in ham but have almost zero chance of being found in spam.
- Ham Misidentified: 4731.2001-07-09.farmer.ham.txt
  - This email is an order confirmation from Amazon.com. The many “$” and “amazon” probably contributed most to the spam misidentification.

Extra Credit

- Tuning
  - The parameters **k** and **m** can be tuned on the training set by including the flag “-t” on the command line. This will tune the values of **k** and **m** around the initial values given (or defaults).
  - The tuning function evaluates the accuracy of different value pairs for **k** and **m**.
    - The training set is randomly divided into a **sub-training** set and a **hold-out** set.
    - The **likelihoods** are calculated from the sub-training set and using **m**.
    - The accuracy of the model using **k** and **m** is **evaluated on the hold-out** set.
  - The average accuracy over 4 splits is used to compare parameter value pairs.