DEEP
LEARNING ACADEMY

Today we will implement a spam filter using a Multinomial Naive Bayes classifier on a bag of words model.
This exercise will loosely follow the exercise from Andrew Ng's online Machine Learning course:
http://openclassroom.stanford.edu/MainFolder/DocumentPage.php?course=MachineLearning&doc=exercises/ex6/ex6.html
However, we will not use the feature data and first use different probability formulas.

It also worth taking a look at the wiki page:
https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering

1. Download the prepared pre-processed emails from (do not download the features data from the exercise page):
http://openclassroom.stanford.edu/MainFolder/courses/MachineLearning/exercises/ex6materials/ex6DataEmails.zip
2. To create our classifier, it is recommended to create the following helper functions:
    a. Read all words from a file into a list
    b. Count number of instances of words in a list

    Optional : Try using Python's defaultdict class which provides a convenient dictionary with default value, thus simplifying the counting instances code.
3. Compute the probability of word appearing in a spam or a ham (not a spam) message by counting the number of times it appears in all appropriate messages divided by the number of words in all those messages.
4. To combine the probability for all words in a message we inspect, we need to multiply them. Since we are multiplying many small numbers we are likely to run into floating pointing underflow. Hence we will calculate the log probabilities instead, and add them up!
5. Now build the classifier which will simply compare the probability of message being ham or spam, and check the accuracy of prediction for the test data.
6. Alternatively to the method in 3 you can calculate the probabilities of each word being part of a ham or spam message, using other methods such as the one suggested in the exercise link, or the wiki page. Try to experiment with a few methods, as well as dealing with issues such as words which didn't appear in train data, or rare words.

Good Luck!