Matching Learning Homework 2 Report

To build vocabulary used in the algorithms, I chose to split contents of the files with space.

Part 1 Multinomial Naive Bayes Algorithm

The accuracy on test set is 0.9477.

Part 2 MCAP Logistic Regression Algorithm

For different value of λ, the accuracies on test set are as following. To balance the running time and the accuracy, choose the number of iteration as 30.

Table 1 the accuracies on test set using MCAP Logistic Regression with different λ
(initial weight = 0, η = 0.01, hard limit on the number of iteration = 30)

Value of λ	0.001	1	5
Accuracy	0.9100	0.8682	0.8138

Part 3 Throwing away Stop Words with The Two Algorithms

In this part I use the Long Stop Word list, about 660 words, from the given website.

Before running on the test set, I supposed that the accuracy would increase. Because these words in the list are which people used most in daily life and have less information than other words. Thus they can be treat as noises for the content and when we remove the noise from the data, we should get a high accuracy.

However, the practical result is that the accuracy does not improve. And Table 2 shows the results on the test data by throwing away stop words. I think the reasons are maybe as following.

However, the practical result is that the accuracy does not improve. And Table 2 shows the results on the test data by throwing away stop words. I think the reasons are maybe as following.

First assumption is that the goal of this experiment is to classify emails between ham and spam. And the words in the list are the key attributes to do this classification. In other word, these words have high weight to decide an email belonging to which class. And the function of λ is regularization, which means smoothing the weights of the attributes.Thus after we removing these words from vocabulary, the accuracy would decrease. To verify this assumption, we can do the training and test with other kind of files in which the stop word do not have hight weights to classify. And the expected result is the accuracy do not change

Second assumption is that the degree of convergent throwing away stop words or not is not the same. Since to balance the running time and the accuracy, choose the number of iteration as 30. And compare the result of Logistic Regression and Naive Bayes, we can get that the convergent of Logistic Regression is not so good with 30 times. If use a greater number of iteration as 50 (λ =1, η = 0.01), the accuracy can be better as 0.912. And to solve this problem, we can use a hight iteration time.

Table 2 the accuracy by throwing away stop words
(initial weight = 0, η = 0.01, hard limit on the number of iteration = 30)

	NaiveBayse	**Logistic Regression**		
λ	——	0.001	1	5
Accuracy	0.9247	0.8766	0.8870	0.7748