Naive Bayes and Logistic Regression for Text Classification

1. Multinomial Naïve Bayes

Accuracies Obtained:

	With Stop Words	Without Stop Words
Accuracy	**94.35146443514645**	**93.93305439330544**

Observation:

Accuracy decreased by 0.42% after removing the stop words from the text files.

Analysis:

The observed reduction in accuracy could be due to the files with high number of stop words. That is, if a spam or ham file contains mostly stop words, then the considered word set for calculating the class probability is totally reduced after removing the stop words leading to misclassification and hence explaining the slight decrease in the accuracy.

2. Logistic Regression

Accuracies Obtained:

Learning Rate eta (η)	Lambda (λ)	Iterations	Accuracy	
			With Stop Words	Without Stop Words
0.01	0.01	10	**72.80334728033473**	**85.14644351464436**
		20	**73.87293774802033**	**84.93723849372385**
		25	**72.80334728033473**	**85.87293748957389**
	0.1	10	**71.38489945774345**	**84.72803347280335**
		12	**72.12893048893744**	**83.68200836820083**
		14	**74.23648949957599**	**84.72803347280335**
	0.5	16	**74.78475899458899**	**85.77405857740585**
		25	**72.80334728033473**	**87.02928870292888**
		30	**71.37848999485972**	**83.16837489472893**

Learning Rate eta (η)	Lambda (λ)	Iterations	Accuracy	
			With Stop Words	Without Stop Words
0.025	0.01	5	**72.12893048893744**	**84.93723849372385**
		10	**74.82368493849380**	**83.45445598789933**
		20	**76.23783940059049**	**74.12344567789074**
	0.1	5	**77.38749298384792**	**84.87389027839020**
		10	**72.37847374873743**	**86.83900237774883**
		20	**73.72848898934834**	**84.27468939489048**
	0.5	5	**72.64789492832738**	**72.16748398948939**
		10	**71.65263674789023**	**85.76837847893743**
		20	**73.82739483984774**	**84.36748959993479**

Learning Rate eta (η)	Lambda (λ)	Iterations	Accuracy	
			With Stop Words	Without Stop Words
0.5	0.01	5	**73.27588983933040**	**72.80334728033473**
		10	**75.89736737473743**	**84.36748959993479**
		20	**71.78782378499753**	**80.38299874785639**
	0.1	5	**73.78672889377841**	**84.30848959993479**
		10	**74.57838799483943**	**81.36729838473643**
		20	**72.80334728033473**	**84.37287364829298**
	0.5	5	**72.69892674852894**	**82.67238472832846**
		10	**73.86827389038474**	**84.37826452789492**
		20	**72.80334728033473**	**84.78923849274728**

Observation:

In the Logistic Regression classifier, the accuracy increased in most of the cases after removing the stop words.

Analysis:

The reason for increase in the accuracy is due the distribution of stop words in both classes. Stop words are not helpful in classifying the email i.e. the weightage (importance) of these words should be set to least(zero). Whereas in case of dictionary with stop words, stop words gets weightage and if the stops words are more in number it leads to misclassification. Hence, there is an increase in accuracy after removing the stop words.