-----Mail Filter----

Name: Aditya Borde

Net Id: asb140930

Naïve Bayes:

Accuracy before Removing Stop Words:

1) Accuracy on Spam Mails : 98.46%
2) Accuracy on Ham Mails : 94.83%
3) Overall Accuracy on Test Set : 95.82%

Accuracy after Removing Stop Words:

1) Accuracy on Spam Mails : 98.46%
2) Accuracy on Ham Mails : 93.96%
3) Overall Accuracy on Test Set : 95.18%

Conclusions:
Observed slight decrease in overall Accuracy in Naïve Bayes.

Reason: Naïve Bayes is using the principle that all features are conditionally independent of each other. Here, each word in the mail is acting as a feature for Spam/Ham detection. Also, Bag of words suggests that position of the word in doesn't matter. So, all words are treated as equally important.
After removing the stop words, the information on the basis of which Naïve Bayes is calculating the accuracy is decreased (As stop words are not considered while every calculation. Hence information gain reduced.)Even if they are common and generally language required common words, we can see a slight decrease in the accuracy due to their occurrence is making slight impact on calculation as every word has equal priority.

Logistic Regression:

Accuracy - Lambda and Iterations:

For η = 0.0008

	No of iterations	Value of λ	Accuracy After Removing Stop Words (%)			Accuracy Before Removing Stop Words (%)		
			SPAM	HAM	Overall	SPAM	HAM	Overall
1	100	0.0003	80.77%	85.63%	84.31%	73.08%	83.33%	80.54%
2	100	0.001	75.38%	89.36%	85.56%	77.69%	84.19%	82.42%
3	200	0.0003	72.58%	90.34%	85.14%	74.33%	87.67%	90.43%
4	200	0.001	78.43%	89.08%	87.68%	78.67%	85.24%	84.64%
5	300	0.0003	76.92%	93.39%	89.91%	77.64%	92.81%	90.56%
6	300	0.001	83.84%	89.08%	87.65%	75.38%	89.65%	85.77%
7	400	0.0003	82.51%	93.39%	90.16%	79.23%	86.20%	84.30%
8	400	0.001	78.92%	91.09%	89.23%	76.15%	84.54%	85.38%
9	500	0.0003	79.12%	89.16%	89.13%	78.15%	87.97%	89.64%
10	1000	0.0003	83.32%	92.81%	91.68%	82.46%	93.97%	92.74%

Accuracy - Lambda and Iterations:

For η = 0.001

	No of iterations	Value of λ	Accuracy After Removing Stop Words (%)			Accuracy Before Removing Stop Words (%)		
			SPAM	HAM	Overall	SPAM	HAM	Overall
1	100	0.0006	69.84%	93.10%	85.14%	65.69%	89.08%	80.54%
2	100	0.002	68.46%	90.80%	84.72%	69.23	91.04%	80.23%
3	200	0.0006	76.15%	89.94%	86.19%	74.76%	92.52%	85.89%
4	200	0.002	73.07%	90.51%	85.77%	82.30%	89.08%	87.23%
5	300	0.0006	73.07%	93.96%	88.28%	70.15%	92.81%	85.56%
6	300	0.002	71.03%	90.79%	84.36%	7.00%	93.67%	86.40%
7	400	0.0006	74.84%	92.52%	88.44%	78.56%	89.35%	85.39%
8	400	0.002	72.00%	93.39%	87.03%	70.00%	93.67%	86.40%
9	500	0.002	77.69%	91.66%	87.86%	83.07%	87.19%	86.79%
10	1000	0.002	84.34%	94.17%	92.32%	87.42%	92.14%	90.89%

Conclusions:

Observed accuracy may change every time depending upon the weights that has been assigned by the program randomly and number of iterations. So, no exact pattern is observed but, sometimes, overall accuracy will increase after removing the stop words in Logistic Regression. From the above observations, accuracy over Spam folder is always lesser than its accuracy over Ham. Detection of Spam is always have lesser accuracy.

Reason: Every iteration, Logistic Regression is updating the weight vector. So, removal of stop words may increase the accuracy due to the fact that they are having no particular meaning and have been assigned some weight while choosing the spam/ ham class. It can't be said surely that stop words removal will always increase the accuracy. For lager iterations accuracy value for is increased after removing stop words due to the convergence. Main thing is Ham accuracy always increases over the number of iterations.