1 Data Set

From given training spam and ham data sets, we find, there are 10,450 words or tokens. Here we split the whole data set using white space. Among them 6,289 tokens from spam and 5,903 tokens from ham. Thus, both spam and ham data sets have 1,742 common words.

2 Naive Bayes (NB) Algorithm

2.1 With stop words

Table 1 presents the corresponding accuracy statistics for NB classifier with stop words. We find, out of 130 spam 125 and out of 348 ham 328 are detected by NB. Overall accuracy is about 94.76 detecting 5 and 20 spam and ham incorrectly respectively. For all tables in this report, total number of spams and hams in test data set are given in paranthesis in *Classification* column.

Classification	**Correct Detection**	**Incorrect Detection**	**Accuracy**
Spam (130)	125	5	0.9615
Ham (348)	328	20	0.9425
Total (478)	**453**	**25**	**0.9476**

Table 1: Accuracy using Naive Bayes classifier using all words.

2.2 Without stop words

As like table 1, table 2 shows the performace of NB without stop words, collected from `http://www.ranks.nl/stopwords`. We find there is a small drop of accuracy.

Classification	**Correct Detection**	**Incorrect Detection**	**Accuracy**
Spam (130)	122	8	0.9384
Ham (348)	329	19	0.9454
Total (478)	**451**	**27**	**0.9435**

Table 2: Accuracy using Naive Bayes classifier without stop words.

The reduction of accuracy is not significant. The accuracy falls can be concidental regularities that means it may happen by chance. Another reason can be: frequency of some stop words is high and their reduction reduce overall information.

3 Logistic Regression (LR) for classification

In this homework, we use LR with L2 regularization. LR has more parameters that can be optimized than NB. And running time of LR is $\mathcal{O}(ndi)$ where n is number of examples, d is number attributes or vocabulary or token, and i is number of iterations. On the other hand, for NB running time is $\mathcal{O}(nd)$ which as least as small when iteration is 1 for LR. To understand impact of various parameters, we set up several experiments. Our focused for this project prameters are:

1. regularization parameter, λ
2. number of iteration
3. learning rate, η
4. weight initialization value

Findings from are experiments are presented in following subsections.

3.1 Impact of regularization parameter (λ)

To understand the impact of regularization parameter or strength of penalty term, we keep iteration value, learning rate, and initial values of weights in fixed at 100, 0.01, and 0 respectively. And experiments are run from 0 to 20 for λ values.

Tabel 3 and 4 shows accuracy for various λ with and without stop words.

From table 3 we can see, small value for λ provides better results. As the value of this parameter is increased, we see a decrease in accuracy. Here need to mention, total accuracy does not give the big picture of the performance of the classifier. For example, for $\lambda = 20$ LR can only detect the ham and over all accuracy is 72.80% which is not a good hypothesis. For this, we include numeric values for correct and incorrect detection as well as corresponding accuracy.

When we eliminate stop words, we see almost same trend (see table 4), decreasing accuracy. However, For some cases, removing stop words increase accuracy for both spam and ham.

3.2 Impact of iteration

To understand the impact of iteration, we start iteration from 10 to 500 by increasing 10. We fix other parameters at $\lambda = 2$, $\eta = 0.01$, and initial weight value = 0. For iteration number (as convergence is very time consuming to get), as we increase it, the accuracy increment is not monotonic (see Table 5 and Table 6). For example, we receive the best result for LR with iteration 60 and 450 which is 93.51%.

3.3 Impact of learning rate (η)

To understand the influence of learning rate, η, we start our experiments from 0 to 0.1 by increasing 0.01. We set other parameters as: iteration number = 280 (as it receives good result when we use stop words), $\lambda = 2$ (as we receive best result for this), and initial weight value = 0.

The findings of the impact of learning rate are given in table 7 and table 8. When we set $\eta = 0$, we receive the hypothesis can detect only spam. And when hypothesis is very high, the generated hypothesis can detect only ham. Still we are not getting a linear relation for increasing learning rate. Like for $\eta = 0.01$, we get best result (93.30%). And for $\eta = 0.09$, we get second best result (93.09%) for using stop words. When we remove stop words, we comparatively receive bad result in compare with that we receive with stop words.

3.4 Impact of weight initialization

Tabel 9 and table 10 show the impact of initialization of weights with various values. Other parameters like number of iteration, learning rate, and regularization factor are set to 280, 0.01, and 2 respectively. For initial weight value = 0, we receive best result (93.30%) when we use stop words and with initial weight value = -3, we receive best result (92.46%) when we eliminate stop words.

Overall, NB's performance is better than LR. But performance defference is not too high like best accuracy from NB is 94.76% and best accuracy from LR is 93.51%. Though the different is not high, NB is performancing better with small training size than its counterpart LR. This helps us to understand, we need less amount samples for examples for NB than LR for non-asymptotic analysis.

4 Extra Credit

We check with two approaches for improving the accuracy.

4.1 Higher value for smoothing

In this approach, we tried to improve accuracy by increasing prior frequency for each word. This approach only affect the result of NB. Instead of Laplace Smoothing, we increase prior or hallucinated value of zero frequency words. However, this approach does not increase the accuracy. For example, if we use 2 instead of 1. We receive 93.09% accuracy which is still below than 94.76%. And 126 spam out of 130 and 319 ham out of 348 are correctly detected. The possible reason can be we are overestimating our confidence about the occurance of the word.

4.2 Eliminating less frequent words

In this approach we elminated the word, if frequency of words is less than 5. The effect of this approach is on both Naive Bayes and Logistic Regression. However, still we are not getting better result. Possible reason can be we are losing important information from eliminating them. The accuracy for NB is 93.51% while detecting 126 spam out of 130 and 321 ham out of 348. And LR receives accuracy with 77.19% while detecting 28 spam out of 130 and 341 ham out of 348. For LR performace degrade much specially for spam. And the parameters value for this results is: $\lambda = 2$, number of iteration $= 100$, $\eta = 0.01$, and initial weight value $= 0$.

λ	**Classification**	**Correct Detection**	**Incorrect Detection**	**Accuracy**
0	Spam (130)	110	20	0.8461
	Ham (348)	326	22	0.9367
	Total (478)	**436**	**42**	**0.9121**
1	Spam (130)	112	18	0.8615
	Ham (348)	328	20	0.9425
	Total (478)	**440**	**38**	**0.9205**
2	Spam (130)	113	17	0.9000
	Ham (348)	329	17	0.9454
	Total (478)	**442**	**34**	**0.9246**
3	Spam (130)	100	30	0.7692
	Ham (348)	330	18	0.9482
	Total (478)	**430**	**48**	**0.8995**
4	Spam (130)	36	94	0.2769
	Ham (348)	347	1	0.9971
	Total (478)	**383**	**95**	**0.8012**
5	Spam (130)	72	58	0.5538
	Ham (348)	326	22	0.9367
	Total (478)	**398**	**80**	**0.8326**
7	Spam (130)	8	112	0.0615
	Ham (348)	348	0	1.000
	Total (478)	**356**	**122**	**0.7447**
8	Spam (130)	116	14	0.8923
	Ham (348)	172	176	0.4942
	Total (478)	**288**	**190**	**0.6025**
12	Spam (130)	1	129	0.0076
	Ham (348)	348	0	1.0000
	Total (478)	**349**	**129**	**0.7301**
13	Spam (130)	9	121	0.0692
	Ham (348)	347	1	0.9971
	Total (478)	**356**	**122**	**0.7447**
20	Spam (130)	0	130	0.0
	Ham (348)	348	0	1.000
	Total (478)	**348**	**130**	**0.7280**

Table 3: Accuracy using LR classifier with stop words for various values of λ.

λ	Classification	Correct Detection	Incorrect Detection	Accuracy
0	Spam (130)	111	19	0.8538
	Ham (348)	333	15	0.9568
	Total (478)	**444**	**34**	**0.9288**
1	Spam (130)	112	18	0.8615
	Ham (348)	334	14	0.9597
	Total (478)	**446**	**32**	**0.9330**
2	Spam (130)	112	18	0.8615
	Ham (348)	334	14	0.9597
	Total (478)	**446**	**32**	**0.9330**
3	Spam (130)	92	38	0.7076
	Ham (348)	330	18	0.9482
	Total (478)	**422**	**56**	**0.8828**
4	Spam (130)	116	14	0.8923
	Ham (348)	144	204	0.4137
	Total (478)	**260**	**218**	**0.5439**
5	Spam (130)	16	114	0.1230
	Ham (348)	347	1	0.9971
	Total (478)	**363**	**115**	**0.7594**
7	Spam (130)	3	127	0.0230
	Ham (348)	348	0	1.0000
	Total (478)	**351**	**127**	**0.7343**
8	Spam (130)	0	130	0.0000
	Ham (348)	348	0	1.0000
	Total (478)	**348**	**130**	**0.7280**
11	Spam (130)	47	83	0.3615
	Ham (348)	282	66	0.8103
	Total (478)	**329**	**149**	**0.6882**
12	Spam (130)	8	112	0.0615
	Ham (348)	346	2	0.9942
	Total (478)	**354**	**124**	**0.7405**
20	Spam (130)	112	18	0.8615
	Ham (348)	7	341	0.0201
	Total (478)	**119**	**359**	**0.2489**

Table 4: Accuracy using LR classifier with stop words for various values of λ.

iteration Number	Classification	Correct Detection	Incorrect Detection	Accuracy
60	Spam (130)	104	26	0.8000
	Ham (348)	330	18	0.9482
	Total (478)	**434**	**44**	**0.9079**
100	Spam (130)	113	17	0.8692
	Ham (348)	329	19	0.9454
	Total (478)	**442**	**36**	**0.9246**
120	Spam (130)	114	16	0.8769
	Ham (348)	331	17	0.9511
	Total (478)	**445**	**33**	**0.9309**
140	Spam (130)	112	18	0.8615
	Ham (348)	332	16	0.9540
	Total (478)	**442**	**36**	**0.9246**
190	Spam (130)	97	33	0.7461
	Ham (348)	333	15	0.9568
	Total (478)	**430**	**48**	**0.8995**
250	Spam (130)	113	17	0.8692
	Ham (348)	329	19	0.9454
	Total (478)	**442**	**36**	**0.9246**
270	Spam (130)	126	4	0.9692
	Ham (348)	319	29	0.9166
	Total (478)	**445**	**33**	**0.9309**
310	Spam (130)	64	66	0.4923
	Ham (348)	288	60	0.8275
	Total (478)	**352**	**126**	**0.7364**
420	Spam (130)	115	15	0.8846
	Ham (348)	329	19	0.9166
	Total (478)	**444**	**34**	**0.9288**
450	Spam (130)	112	18	0.8615
	Ham (348)	332	16	0.9540
	Total (478)	**444**	**34**	**0.9288**
500	Spam (130)	60	70	0.4615
	Ham (348)	332	16	0.9540
	Total (478)	**392**	**86**	**0.8200**

Table 5: Accuracy using LR classifier with stop words for various iteration numbers.

iteration Number	Classification	Correct Detection	Incorrect Detection	Accuracy
60	Spam (130)	118	12	0.9076
	Ham (348)	329	19	0.9166
	Total (478)	**447**	**31**	**0.9351**
100	Spam (130)	112	18	0.8615
	Ham (348)	333	15	0.9568
	Total (478)	**445**	**33**	**0.9309**
120	Spam (130)	112	18	0.8615
	Ham (348)	334	14	0.9597
	Total (478)	**446**	**32**	**0.93305**
140	Spam (130)	36	94	0.2769
	Ham (348)	310	38	0.8908
	Total (478)	**346**	**132**	**0.7238**
190	Spam (130)	100	30	0.7692
	Ham (348)	333	15	0.9568
	Total (478)	**433**	**45**	**0.9058**
250	Spam (130)	2	128	0.0153
	Ham (348)	348	0	1.000
	Total (478)	**350**	**128**	**0.7322**
270	Spam (130)	21	109	0.1615
	Ham (348)	347	1	0.9971
	Total (478)	**368**	**110**	**0.7698**
310	Spam (130)	108	22	0.8307
	Ham (348)	333	15	0.9568
	Total (478)	**441**	**37**	**0.9225**
420	Spam (130)	113	18	0.8692
	Ham (348)	332	16	0.9540
	Total (478)	**445**	**33**	**0.9058**
450	Spam (130)	114	16	0.8769
	Ham (348)	333	15	0.9568
	Total (478)	**447**	**31**	**0.9351**
500	Spam (130)	120	10	0.9230
	Ham (348)	233	115	0.66954
	Total (478)	**353**	**125**	**0.7384**

Table 6: Accuracy using LR classifier without stop words for various iteration numbers.

η	**Classification**	**Correct Detection**	**Incorrect Detection**	**Accuracy**
0	Spam (130)	130	0	1.0000
	Ham (348)	0	348	0.0
	Total (478)	**130**	**348**	**0.2719**
0.01	Spam (130)	115	15	0.8846
	Ham (348)	331	17	0.9511
	Total (478)	**446**	**32**	**0.9330**
0.02	Spam (130)	65	65	0.5
	Ham (348)	303	45	0.0.8706
	Total (478)	**368**	**110**	**0.7698**
0.03	Spam (130)	117	13	0.9
	Ham (348)	199	149	0.5718
	Total (478)	**316**	**162**	**0.6610**
0.04	Spam (130)	113	17	0.8692
	Ham (348)	123	225	0.3534
	Total (478)	**236**	**242**	**0.4937**
0.05	Spam (130)	5	125	0.0384
	Ham (348)	348	0	1.0000
	Total (478)	**353**	**125**	**0.7384**
0.06	Spam (130)	47	83	0.3615
	Ham (348)	305	43	0.8764
	Total (478)	**352**	**126**	**0.7364**
0.07	Spam (130)	1	129	0.0076
	Ham (348)	348	0	1.0000
	Total (478)	**349**	**129**	**0.7301**
0.08	Spam (130)	97	33	0.7461
	Ham (348)	225	123	0.6465
	Total (478)	**322**	**156**	**0.6736**
0.09	Spam (130)	114	16	0.8769
	Ham (348)	58	290	0.1666
	Total (478)	**172**	**306**	**0.9309**
0.1	Spam (130)	2	128	0.0153
	Ham (348)	348	0	1.0000
	Total (478)	**350**	**128**	**0.7322**

Table 7: Accuracy using LR classifier with stop words for various values of η.

η	**Classification**	**Correct Detection**	**Incorrect Detection**	**Accuracy**
0	Spam (130)	130	0	1.0000
	Ham (348)	0	348	0.0
	Total (478)	**130**	**348**	**0.2719**
0.01	Spam (130)	94	36	0.7230
	Ham (348)	336	12	0.9655
	Total (478)	**430**	**48**	**0.8995**
0.02	Spam (130)	110	20	0.8461
	Ham (348)	215	133	0.6178
	Total (478)	**325**	**153**	**0.6799**
0.03	Spam (130)	97	33	0.7461
	Ham (348)	251	97	0.7212
	Total (478)	**348**	**130**	**0.7280**
0.04	Spam (130)	88	42	0.6769
	Ham (348)	269	79	0.7729
	Total (478)	**357**	**121**	**0.7468**
0.05	Spam (130)	2	118	0.0153
	Ham (348)	348	0	1.0000
	Total (478)	**350**	**128**	**0.7322**
0.06	Spam (130)	84	46	0.6461
	Ham (348)	281	67	0.8074
	Total (478)	**365**	**113**	**0.7635**
0.07	Spam (130)	111	19	0.8538
	Ham (348)	14	334	0.0402
	Total (478)	**225**	**253**	**0.2615**
0.08	Spam (130)	2	118	0.0153
	Ham (348)	348	0	1.0
	Total (478)	**350**	**128**	**0.7322**
0.09	Spam (130)	45	85	0.3461
	Ham (348)	286	62	0.8218
	Total (478)	**331**	**147**	**0.6924**
0.1	Spam (130)	113	17	0.8692
	Ham (348)	5	343	0.0143
	Total (478)	**118**	**160**	**0.2468**

Table 8: Accuracy using LR classifier without stop words for various values of η.

w_i	**Classification**	**Correct Detection**	**Incorrect Detection**	**Accuracy**
-4	Spam (130)	111	19	0.8538
	Ham (348)	330	18	0.9482
	Total (478)	**441**	**37**	**0.9225**
-3	Spam (130)	112	18	0.8615
	Ham (348)	329	19	0.9454
	Total (478)	**441**	**37**	**0.9225**
-2	Spam (130)	126	4	0.9692
	Ham (348)	319	29	0.0.9166
	Total (478)	**445**	**33**	**0.9309**
-1	Spam (130)	0	130	0.0
	Ham (348)	348	0	1.0000
	Total (478)	**348**	**130**	**0.7280**
0	Spam (130)	115	15	0.8846
	Ham (348)	331	17	0.9511
	Total (478)	**446**	**32**	**0.9330**
1	Spam (130)	115	15	0.8846
	Ham (348)	330	18	0.9482
	Total (478)	**445**	**33**	**0.9309**
2	Spam (130)	13	117	0.1
	Ham (348)	348	0	1.0000
	Total (478)	**361**	**117**	**0.7552**

Table 9: Accuracy using LR classifier with stop words for various initial values of w_i.

w_i	Classification	Correct Detection	Incorrect Detection	Accuracy
-4	Spam (130)	63	67	0.4846
	Ham (348)	302	46	0.8678
	Total (478)	**365**	**113**	**0.7635**
-3	Spam (130)	110	20	0.8461
	Ham (348)	332	16	0.0.9540
	Total (478)	**442**	**36**	**0.9246**
-2	Spam (130)	109	21	0.8384
	Ham (348)	332	16	0.9540
	Total (478)	**441**	**37**	**0.9225**
-1	Spam (130)	52	78	0.4
	Ham (348)	309	39	0.8879
	Total (478)	**361**	**117**	**0.7552**
0	Spam (130)	94	36	0.7230
	Ham (348)	336	12	0.9655
	Total (478)	**430**	**48**	**0.8995**
1	Spam (130)	36	94	0.2769
	Ham (348)	345	3	0.9913
	Total (478)	**381**	**97**	**0.7970**
2	Spam (130)	92	38	0.7076
	Ham (348)	321	27	0.9224
	Total (478)	**413**	**65**	**0.8640**

Table 10: Accuracy using LR classifier without stop words for various initial values of w_i.