COMP 6721 Project 2 Report

Zhongxu Huang[1] and Yixuan Li[2]

[1] 40052560 realdonald9@gmail.com
[2] 40079830 liyixuan4030614@gmail.com

1 Experiment 1 (½ page)

1.1 Baseline experiment

In experiment #1 (baseline experiment), we compute the prior for each class and conditional probability for each word in the vocabulary, and build the model based on the emails in the training set. While computing the conditional probability, the smoothing factor $\delta = 0.5$. In the next step, we use the prior and the likelihood to compute *overrall_ham* (score(ham)), and *overrall_spam* (score(spam)) on all test set files and label the emails.

1.2 Results and analysis

By analyzing the result, we evaluate the baseline experiment with the test set by these values: Accuracy, Precision, Recall, F_1-measure.

Table 1. Analysis of baseline experiment

Evaluation \ Class	Ham	Spam
Accuracy		
Precision		
Recall		
F_1-measure		

(explain the Table 1)

Table 2. Confusion matrix of baseline experiment

Correct class (that should have been assigned)	Classes assigned by our classifier		
	Ham	Spam	Total
Ham			
Spam			

(explain the Table 2)

2 Experiment 2 (½ page)

2.1 Stop-word filtering

In the experiment of stop-word filtering, we remove the given stop words from the vocabulary and ignore the stop words in emails. Then we re-compute the conditional probabilities for each word, and use the same smoothing factor ($\delta = 0.5$) and prior is same as in experiment #1. Next, we use the updated likelihood to classify emails on the test set.

2.2 Results and analysis

After removing the stop words in vocabulary and experimenting with the classifier, we show the results in the same way as in experiment 1.

Table 3. Analysis of stop-word filtering experiment

Class / Evaluation	Ham	Spam
Accuracy		
Precision		
Recall		
F_1-measure		

(explain the Table 3)

Table 4. Confusion matrix of stop-word filtering experiment

Correct class (that should have been assigned)	Classes assigned by our classifier		
	Ham	Spam	Total
Ham			
Spam			

(explain the Table 4)

(if the accuracy goes up, describe the possible reasons, if not, explain why removing stop words doesn't help?)

3 Experiment 3 (½ page)

3.1 Word-length filtering

In the word-length filtering, instead of removing the given stop words, we will remove all words which length is equal or shorter than 2 and words which length is equal or longer than 9 from the vocabulary. And still, we use the same smoothing factor (δ = 0.5), and prior is same as in the previous experiments.

3.2 Results and analysis

After we re-classify again on the emails with test set, here are the same 2 tables which show the results and performance changes:

Table 5. Analysis of word-length filtering experiment

Class / Evaluation	Ham	Spam
Accuracy		
Precision		
Recall		
F_1-measure		

(explain the Table 5)

Table 6. Confusion matrix of word-length filtering experiment

Correct class (that should have been assigned)	Classes assigned by our classifier		
	Ham	Spam	Total
Ham			
Spam			

(explain the Table 6)

(if the accuracy goes up, describe the possible reasons, if not, explain why removing short-length and long-length words doesn't help?)

4 Experiment #1, #2, #3 comparison (1 page)

In this section, we will analyze the result (Accuracy, Precision, Recall, F_1-measure) given by baseline model, stop-word filtering and word-length filtering. We show the confusion matrix of 3 experiments in a same table and we will discuss the performance on the three classifiers.

Table 7. Analysis of experiment #1, #2, and #3

Experiment / Evaluation	#1 Baseline	#2 Stop-word	#3 Word-length
Accuracy			
Precision			
Recall			
F_1-measure			

Table 8. Confusion matrix of the first three experiments

Correct class (that should have been assigned)	Classes assigned by our classifier						
	Ham			Spam			Total
Experiment No.	#1	#2	#3	#1	#2	#3	
Ham							
Spam							

Compare the 3 experiments follows this steps:
1. compare baseline and stop-word, 在section 2.2应该有所提及
2. compare baseline and word-length, 在section 3.2应该有所提及
3. compare stop-word and word-length (which method helps more? And why?)
4. 需不需要分析 remove stop-word and word-length 组合的实验？
5. 其他分析

5 Encountered difficulties and interest on experiment #1, #2, #3 (½ page for difficulties + ½ page for interests)

5.1 Difficulties

Difficulty 1.
(Describe)

Solution:
(Describe)

Difficulty 2.
(Describe)

Solution:
(Describe)

5.2 Interests and inspirations

(I think one choice is we could talk about how to choose a better hyper-parameter, such as smoothing factor δ and β in F_1-measure etc., or it there a good way to use AI/ML to detect better hyper-parameters)

(我在这个网站https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/ 查naïve bayes classifier时，文章在最后有介绍advanced techniques，我觉得也可以写一写) [3]

6 Experiment 4 (½ page)

6.1 Infrequent word filtering

In experiment 4, firstly, we will compute frequency for all words in emails, and then we gradually remove the infrequent words and a certain proportion of the top most frequent words from the vocabulary. Next, we re-compute the conditional probability of each word in vocabulary using the same smoothing factor ($\delta = 0.5$), and use the updated probability to calculate the scores and classify the emails in the test set.

6.2 Results and analysis

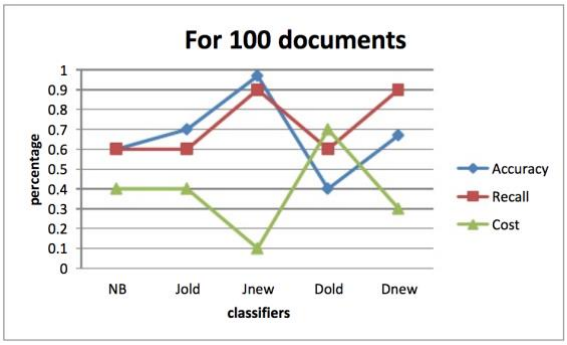
Stop words are words that show up a lot in documents, such as prepositions, pronouns, etc.[4]. At first, as analyzing the previous experiments, we will show the same 2 tables that represent the performance of the classifier as before. Next, we will compare the different removal strategies and discuss by removing how much proportion based on the word frequency, the performance achieves the best. In the end, we will plot the performance of the classifier against the number of words left in the vocabulary.

Table 9. Analysis of infrequent word filtering experiment

Evaluation \ Class	Ham	Spam
	Removing words with frequency = 1	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 5	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 10	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 15	
Accuracy		
Precision		
Recall		
F_1-measure		

	Removing words with frequency ≤ 20	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 20 and the top 5%	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 20 and the top 10%	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 20 and the top 15%	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 20 and the top 20%	
Accuracy		
Precision		
Recall		
F_1-measure		
	Removing words with frequency ≤ 20 and the top 25%	
Accuracy		
Precision		
Recall		
F_1-measure		

(explain Table 9)
(这样的表格是不是好的方式？换成线性图或柱状图会不会更好？)
比如：



图片来自：
https://www.academia.edu/9040601/NAIVE_BAYES_CLASSIFIER_WITH_MODIFIED_SMOOTHING_TECHNIQUES_FOR_BETTER_SPAM_CLASSIFICATION

Table 10. Confusion matrix of infrequent word filtering experiment

Removing words with frequency	Correct class (that should have been assigned)	Classes assigned by our classifier		
		Ham	Spam	Total
= 1	Ham			
	Spam			
≤ 5	Ham			
	Spam			
≤ 10	Ham			
	Spam			
≤ 15	Ham			
	Spam			
≤ 20	Ham			
	Spam			
≤ 20 and the top 5%	Ham			
	Spam			
≤ 20 and the top 10%	Ham			
	Spam			
	Ham			

≤ 20 and the top 15%	Spam			
≤ 20 and the top 20%	Ham			
	Spam			
≤ 20 and the top 25%	Ham			
	Spam			

(explain Table 10, compare performance of all removal strategies)

(Then give the output plot of the program and describe)

7 Experiment 5 (½ page)

7.1 Experiment 5: change smoothing

In this experiment, we will do different smoothing factor trials from no smoothing to $\delta = 1$ by increasing 0.1 at each step. For each step, we re-calculate the conditional probability for each word in the vocabulary and use it to compute the scores.

7.2 Results and analysis

As we known, for the reason that a word, especially for person names, proper noun etc., which does not appear in any training text but is in a test set leads to score of 0 for all classes. To balance this, additive smoothing is commonly a component of Naïve Bayes classifiers. Most cases, we use add-one smoothing, but in practice, a smaller value is preferred[1]. Smoothing methods plays important roles in Naïve Bayes classifiers, it not only improves the accuracy of the language model, but also accommodates the generation of new words and non-informative words[2].

In this section, we will analyze the results and evaluate which smoothing factor suits the best for the model.

Table 11. Analysis of smoothing changing experiment

Class / Evaluation	Ham	Spam	Smoothing factor
Accuracy			0
Precision			
Recall			
F_1-measure			
Accuracy			0.1
Precision			
Recall			
F_1-measure			
Accuracy			0.2
Precision			
Recall			
F_1-measure			
Accuracy			0.3
Precision			
Recall			
F_1-measure			
Accuracy			0.4
Precision			

Recall			
F_1-measure			
Accuracy			0.5
Precision			
Recall			
F_1-measure			
Accuracy			0.6
Precision			
Recall			
F_1-measure			
Accuracy			0.7
Precision			
Recall			
F_1-measure			
Accuracy			0.8
Precision			
Recall			
F_1-measure			
Accuracy			0.9
Precision			
Recall			
F_1-measure			
Accuracy			1.0
Precision			
Recall			
F_1-measure			

(explain Table 11)
(Again,
表格在这里是不是比较好的表现方式？换成线性图或柱状图会不会更好？)

Table 12. Confusion matrix of smoothing changing experiment

Smoothing factor	Correct class (that should have been assigned)	Classes assigned by our classifier		
		Ham	Spam	Total
0	Ham			
	Spam			
0.1	Ham			
	Spam			
0.2	Ham			
	Spam			
0.3	Ham			
	Spam			
0.4	Ham			
	Spam			
0.5	Ham			
	Spam			
0.6	Ham			
	Spam			
0.7	Ham			
	Spam			
0.8	Ham			
	Spam			
0.9	Ham			
	Spam			
1.0	Ham			
	Spam			

(explain Table 12)

8 Comparisons on all experiments (1 page)

(这个需要从哪几个什么方面来比较？什么表现方式比较好？)

9 References

1. Wikipedia, Additive smoothing, https://en.wikipedia.org/wiki/Additive_smoothing
2. IJCSMC Jounrnal, Vol. 3. Issue. 10, October 2014, page 869-878, https://www.academia.edu/9040601/NAIVE_BAYES_CLASSIFIER_WITH_MODIFIED_SMOOTHING_TECHNIQUES_FOR_BETTER_SPAM_CLASSIFICATION
3. A practical explanation of a Naïve Bayes classifier, https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/
4. Multinomial Naïve Bayes Classifier for Text Analysis, https://towardsdatascience.com/multinomial-naive-bayes-classifier-for-text-analysis-python-8dd6825ece67
5.

最后根据正文调整一下 reference 的顺序，先引用的在前，后引用的在后