Final Project

IST 664 – Natural Language Processing

Christina Ling

2019

# Introduction

For the final project of this course, the detection of SPAM in email was selected. The dataset includes several emails from the Enron public email corpus. The emails were labeled as spam for spam emails or ham for real emails. The dataset includes 3,672 regular emails and 1,500 spam emails – for a total of 5,172 emails.

This project uses Naïve Baye’s machine learning technique to classify emails. The contents of the emails were examined to determine the classification of the email. A base classification was conducted followed by some experiments. The experiment used additional preprocessing to increase the accuracy of the algorithm.

# Base Pre-Processing

After import all 5,172 emails, some pre-processing was conducted before applying the Naïve Bayes tool. First the numbers were removed from the text, so the program only analyzes the words. Next a lemmatizer was applied. By applying a lemmatizer the program is able to compare similar root words. For example, normally a computer treats the words “going” and “go” as two different entities. By lemmatizing the text, the word “go” and “going” are transform to their base form and allow the computer to treat them as the same entity.

# Base Algorithm

The initial machine learning algorithm uses the top 500 most frequent words to determine whether the email is spam or ham. While the shuffling function may vary the results each time the program is used. For this report, the following is the result of the base algorithm

Accuracy = 0.9301366331528744

|  |  |
| --- | --- |
| Cross Validation | |
| 1 | 0.9197292069632496 |
| 2 | 0.9313346228239845 |
| 3 | 0.9187620889748549 |
| 4 | 0.9235976789168279 |
| 5 | 0.9274661508704062 |
| Mean Accuracy = 0.9241779497098644 | |

|  |  |  |
| --- | --- | --- |
|  | Ham | Spam |
| Ham | 64.3% | 6.7% |
| Spam | 0.3% | 28.7% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Ham | 0.905 | 0.996 | 0.949 |
| Spam | 0.991 | 0.810 | 0.891 |

# Experiments

Three experiments were conducted. The first experiments with the number of most frequency words. The second examine the affect of stop words. The third experiment looks at the affect of names.

For the first experiment many different numbers of most frequent words. The numbers were commented out after experimenting with the code. The first numbers used was 1000 of the most common words. This yielded better results than the base 500. Since the increase of number of words used increased the accuracy, the next number was higher than 1000 – the next number used was 1500. This yield worst results. The following experiments choose number between 1000 and 1500 to optimize the accuracy. Finally, 1400 words were chosen after yielding the most accurate results. The results is as follow:

Accuracy = 0.9479247228667183

|  |  |
| --- | --- |
| Cross Validation | |
| 1 | 0.9555125725338491 |
| 2 | 0.9535783365570599 |
| 3 | 0.9371373307543520 |
| 4 | 0.9497098646034816 |
| 5 | 0.9410058027079303 |
| Mean Accuracy = 0.9473887814313345 | |

|  |  |  |
| --- | --- | --- |
|  | Ham | Spam |
| Ham | 66.5% | 4.35% |
| Spam | 0.7% | 28.3% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Ham | 0.976 | 0.863 | 0.916 |
| Spam | 0.936 | 0.990 | 0.962 |

The second experiment removed stop words from the pre-processing and built on the first experiment. So, the second experiment also uses 1400 of the most common words. The results yield nearly identical results. The following is the results of this experiment

Accuracy = 0.9494715132766177

|  |  |
| --- | --- |
| Cross Validation | |
| 1 | 0.9516441005802708 |
| 2 | 0.9458413926499033 |
| 3 | 0.9526112185686654 |
| 4 | 0.9468085106382979 |
| 5 | 0.9400386847195358 |
| Mean Accuracy = 0.9473887814313346 | |

|  |  |  |
| --- | --- | --- |
|  | Ham | Spam |
| Ham | 65.9% | 4.9% |
| Spam | 0.2% | 29.1% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Ham | 0.931 | 0.997 | 0.963 |
| Spam | 0.994 | 0.856 | 0.920 |

The last experiment conducted removed names in the preprocessing. This experiment also built on the last one. So, this also removed the stop words and used 1400 of the most frequent words. This experiment yielded slightly better results. The results are as following.

Accuracy = 0.9409641660221707

|  |  |
| --- | --- |
| Cross Validation | |
| 1 | 0.9545454545454546 |
| 2 | 0.9419729206963250 |
| 3 | 0.9506769825918762 |
| 4 | 0.9497098646034816 |
| 5 | 0.9458413926499033 |
| Mean Accuracy = 0.948549323017408 | |

|  |  |  |
| --- | --- | --- |
|  | Ham | Spam |
| Ham | 67.0% | 4.3% |
| Spam | 1.6% | 27.1% |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| Ham | 0.940 | 0.976 | 0.958 |
| Spam | 0.943 | 0.864 | 0.902 |