Spam Filter Report

Taehoon Kwon

1. Any decisions specific to spam filter :
   1. Symbols that are attached to the word are ignored.

(except $,%,& - could have possible meaning)

* 1. Some prepositions are excluded. I used the reference below.

(reference : https://www.talkenglish.com/vocabulary/top-50-prepositions.aspx)

* 1. Unnecessary words are excluded (e.g. subject, re, the, a, an, and, or, etc.)
  2. All words are treated as lowercase (Case-insensitive)
  3. Threshold for the probability of spam : 0.5
  4. 50 most spammiest and 50 most hammiest words are used for spam filter

1. The 5 most “spammiest” and the 5 most “hammiest” words from the training stage

|  |  |  |  |
| --- | --- | --- | --- |
| “Spammiest” words | | “Hammiest” words | |
| word | **probability** | **word** | **probability** |
| *adv* | 0.9471510925049372 | *spambayes* | 0.9912658141071086 |
| *money* | 0.9330621453600013 | *zzzzteana* | 0.9905171386516114 |
| *enekio* | 0.9282823584695314 | *razorusers* | 0.9865443555902276 |
| *systemworks* | 0.9087307506036789 | *ouch* | 0.9828321758644969 |
| *mortgage* | 0.8996078029452471 | *sadev* | 0.9825309630291224 |

1. The accuracy, precision and recall rates from the testing stage

|  |  |  |
| --- | --- | --- |
|  | spam | ham |
| predict spam | 29 | 2 |
| predict ham | 96 | 685 |

1. Conclusions on the performance of spam filter and possible improvement steps

The accuracy from the experiment are quite close to the 1, which means most of spams are predicted as spam and most of hams are predicted as ham. And the precision also has high probability that the predicted spam is actual spam. However, the recall value is low.

One possible reason for this problem is the proportion of given data sets are not half and half. The number of given data sets are divided into 2,750 hams and 500 spams, which are also split into 75% of training data and 25% of testing data, resulting in 2061 ham messages and 375 spam messages of training data. There are almost five times more ham messages than spam messages. Therefore, this program performs nice for predicting the ham as ham but not for the spam. Additionally, ham messages accounts for high percentage of calculating the accuracy.

To improve this program, much more samples of spam messages are needed, which makes a set of appropriate words () can be selected for spam filter.