1. Attempt No 1 Basic model
   1. Model Details
      1. Using one directory as test and 9 directories as training, word dictionary contains word and its occurrence is. All these words are selected without choosing the most occurring. The reason for no choosing most frequent words is that spam mails are 5,6 time less than non spam. So when I chose most frequent, representation of spam files become even more less. I select folder 8 for testing, rest of 9 are for training. Non alpha and single word characters are removed
   2. Error analysis
   3. Improvement suggestion

It takes a lot of time in finding the features for the whole corpus

* 1. Precision Recall F1 score
     1. Precision: 97%
     2. Recall: 95%
     3. F1-score: 96%
     4. Accuracy :98%

1. Attempt No 2 Stop words removed
   1. Model Details
      1. Using one directory as test and 9 directories as training, word dictionary contains word and its occurrence is. I remove stop words. I select folder 8 for testing, rest of 9 are for training. Stop words are removed.
   2. Error analysis

When I examined mis predicted files, I found they are non-English emails.

Here is bit of lines of those emails

professur fuer allgemeine sprachwissenschaft an der universitaet heidelberg : die ausschreibung ist erschienen in " ausschreibungsdienst des deutschen hochschulverbandes "

* 1. Improvement suggestion

We need to remove non English emails.

* 1. Precision Recall F1 score
     1. Precision: 100%
     2. Recall: 98%
     3. F1-score: 99%
     4. Accuracy :100%

1. Attempt No 3 Stemming
   1. Model Details
      1. Using one directory as test and 9 directories as training. Stemming is done, but stop words are not removed.
   2. Error analysis

We predicted 5 out of 289 emails as wrong. 2 non spam email as spam and 3 spam emails as non-spam. Two emails are related to students, one email is spam but present with name of non spam. (so it is marked as correct, but its label is wrong). One email is non English

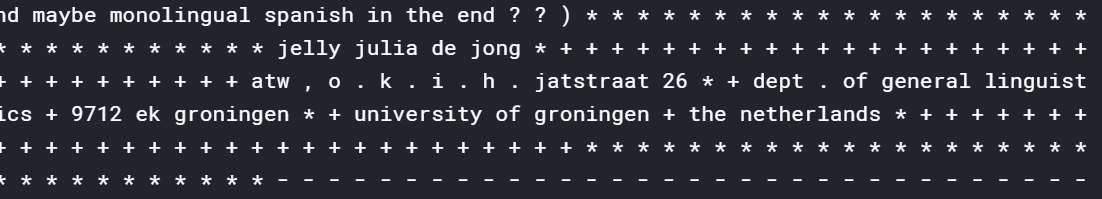
* 1. Improvement suggestion

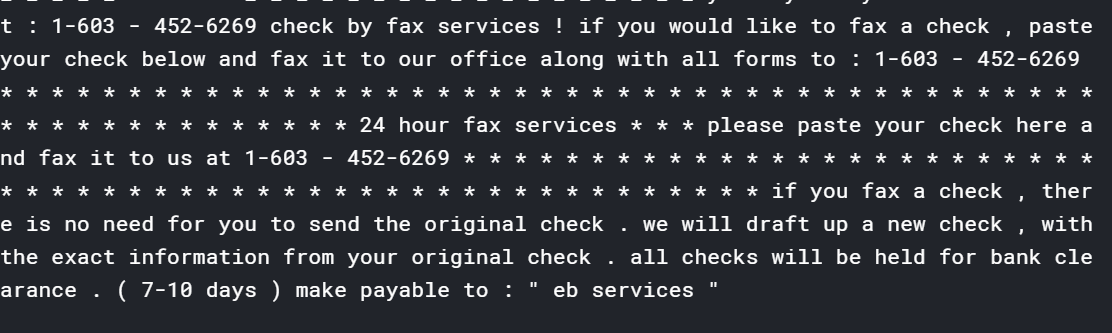
It takes a lot of time in finding the features for the whole corpus

* 1. Precision Recall F1 score
     1. Precision: 97%
     2. Recall: 95%
     3. F1-score: 96%
     4. Accuracy :98%

1. Attempt No 4
   1. Model Details
      1. Using one directory as test and 9 directories as training. Stemming is applied and stop words are removed
   2. Error analysis

Classifier predicted 7 out of 289 emails as wrong. There is a pattern in these emails, most of these emails contains \* sign and dashes –





* 1. Improvement suggestion

It takes a lot of time in finding the features for the whole corpus

* 1. Precision Recall F1 score
     1. Precision: 97%
     2. Recall: 95%
     3. F1-score: 96%
     4. Accuracy 96%

10 Fold Cross validation

Here I made each portion as validation one by one and run the training on others.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Precision |  | 97 | 100 | 99 | 97 | 97 | 99 | 97 | 98 |
| Recall |  | 93 | 99 | 100 | 97 | 96 | 98 | 95 | 98 |
| F1 Score |  | 95 | 99 | 99 | 97 | 98 | 98 | 96 | 98 |
| Accuracy |  | 97 | 99 | 99 | 98 | 98 | 98 | 98 | 98 |