A. Attempt No 1 Basic model
    a. Model Details
        i. 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
    b. Error analysis
    c. Improvement suggestion
        It takes a lot of time in finding the features for the whole corpus
    d. Precision Recall F1 score
        i. Precision: 97%
        ii. Recall: 95%
        iii. F1-score: 96%
        iv. Accuracy :98%
B. Attempt No 2 Stop words removed
    a. Model Details
        i. 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.
    b. 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 unive
        rsitaet heidelberg : die ausschreibung ist erschienen in
        " ausschreibungsdienst des deutschen hochschulverbandes "
        ```

    c. Improvement suggestion
        We need to remove non English emails.
    d. Precision Recall F1 score
        i. Precision: 100%
        ii. Recall: 98%
        iii. F1-score: 99%
        iv. Accuracy :100%
C. Attempt No 3 Stemming
    a. Model Details
        i. Using one directory as test and 9 directories as training. Stemming is done, but stop words are not removed.
    b. 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
    c. Improvement suggestion
        It takes a lot of time in finding the features for the whole corpus
    d. Precision Recall F1 score

i. Precision: 97%
ii. Recall: 95%
iii. F1-score: 96%
iv. Accuracy :98%

D. Attempt No 4
   a. Model Details
      i. Using one directory as test and 9 directories as training. Stemming is applied and stop words are removed
   b. 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 –

```
nd maybe monolingual spanish in the end ? ? ) * * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * jelly julia de jong * + + + + + + + + + + + + + + + + + + + + +
+ + + + + + + + + + atw , o . k . i . h . jatstraat 26 * + dept . of general linguist
ics + 9712 ek groningen * + university of groningen + the netherlands * + + + + + + +
+ + + + + + + + + + + + + + + + + + + + + + + + + * * * * * * * * * * * * * * * * * *
* * * * * * * * * * - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```
t : 1-603 - 452-6269 check by fax services ! if you would like to fax a check , paste
your check below and fax it to our office along with all forms to : 1-603 - 452-6269
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * * * * 24 hour fax services * * * please paste your check here a
nd fax it to us at 1-603 - 452-6269 * * * * * * * * * * * * * * * * * * * * * * * * *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * if you fax a check , ther
e is no need for you to send the original check . we will draft up a new check , with
the exact information from your original check . all checks will be held for bank cle
arance . ( 7-10 days ) make payable to : " eb services "
```

   c. Improvement suggestion
      It takes a lot of time in finding the features for the whole corpus
   d. Precision Recall F1 score
      i. Precision: 97%
      ii. Recall: 95%
      iii. F1-score: 96%
      iv. 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