PRML ASSIGNMENT-3

-By Bhumika Gupta
Roll No. EE21S060

Email Classifier:

STEP 1: DATA-SET CREATION:

There are various email datasets that are publicly available. Some of them are huge. Out of these, 4 datasets from Kaggle were selected. Spam and non-spam emails from each were mixed to create a dataset of 8000 mails with 4000 spam mails and 4000 non-spam mails.

Details:

DATASET (.CSV)	**Total Mails**	**Total Ham mails**	**Total Spam mails**	**Selected Ham mails for dataset**	**Selected Spam mails for dataset**
completeSpamAssassin	6046	4150	1896	1000	1000
lingSpam	2605	2172	433	1000	433
enronSpamSubset	10000	5000	5000	1000	1567
spamham	5728	4360	1368	1000	1000
total				4000	4000

The datasets mentioned above were available in .csv format but had different number of columns and headers. The extra columns except the ones with content and label were dropped from the datasets and the headers were changed to common headers named ‘Body’ and ‘Label’. These individual datasets were then appended and the dataset named ‘dataset.csv’ was created.
It contains 8000 rows and two columns named ‘Body’ and ‘Label’.

‘Body’: contains the raw (unprocessed) email content
‘Label’: contains the true label of the email. 1→ Spam, 0→ Ham

Equal number of ham and spam mails were selected to avoid the model to bias prediction towards one type of classification which is a very common problem in imbalanced dataset classification. Another remedy is to use stratified test-train-split and ensure high recall for all categories of classification.

The ‘dataset.csv’ looks like:

	Body	Label
0	Date: Wed, 21 Aug 2002 10:54:46 -05...	0
1	Martin A posted:\nTassos Papadopoulos, the Gre...	0
2	Man Threatens Explosion In Moscow Thursday Aug...	0
3	Klez: The Virus That Won't Die\n \nAlready the...	0
4	> in adding cream to spaghetti carbonara, whi...	0

The dataset was first shuffled and then was sent for preprocessing.

After shuffling:

	Body	Label
976	\nOn Mon, 09 Sep 2002 12:05:55 PDT,\n\tRick Ba...	0
7677	Subject: entrust your visual identity to us t...	1
7118	Subject: notification from sky bank # 6521 - 3...	1
367	empty	0
2245	Subject: discourse conference final call\n \n ...	0

The dataset.csv is also submitted along with the code.

STEP 2: FEATURE EXTRACTION:

Libraries used:

- Regex for matching links/punctuations/mailids etc.
- Contractions for making short forms to full forms
- Unidecode for identifying special characters
- Nltk for stemming, tokenisation and stop words removal
- Spacy for lemmatisation
- WordCloud for visualisation of words in corpus
- Sklearn TFIDF for vectorization of text tokens

1. Lowercasing
   [for equivalence of “Super”,”SUPER”,”SupeR”,”super” etc.]

2. Contractions fix
   [eg. “wasn’t” to “was not”, “don’t” to “do not” etc.]

3. Removing numbers
   [“12” , “1304000” etc replaced by white spaces]

4. Replacing urls by ‘links’
   [“https:wwwmvdjrgn/fvjrnf/go” replaced by “links”]

5. Removing MailIds
   [“abscd@xyz.com” …etc]

6. Replacing Currency Signs by ‘dollars’
   [“$” to “dollars”]

7. Removing Punctuations
   [.(_$:;”. Etc.]

8. Replacing Accented characters with ASCII
   [“ã”, “ï”,” ÿ” to “a”, “i” , “y”...]

9. Removing multiple occurrences of a character in a string
   [“awwwwww” , “xxxxxxxggggg” to “aww”,”xxgg”]

10. Remove common words like ‘subject and enron’ since these occur frequently in our dataset
    [“subject” , “enron” to “ “]

11. Multiple white spaces and ‘\n\t’ removal
    [“ “ ,”\n” , “\t” to “ “]

12. Lemmatisation and Tokenisation
    [“bats” to “bat”]

13. Stemming
    [“programmer” “programs” “programming” to “program”]

14. Removing Stopwords
    [“a” , “the” , “and” etc]

Before preprocessing the email content looks like this.

“‘

```
   Date:        Wed, 21 Aug 2002 10:54:46 -0500\n    From:        Chris
Garrigues \n    Message-ID:  <1029945287.4797.TMDA@deepeddy.vircio.com>\n
| I can\'t reproduce this error.For me it is very repeatable... (like
every time, without fail).This is the debug log of the pick happening
...18:19:03 Pick_It {exec pick +inbox -list -lbrace -lbrace -subject ftp
-rbrace -rbrace} {4852-4852 -sequence mercury}\n18:19:03 exec pick +inbox
-list -lbrace -lbrace -subject ftp -rbrace -rbrace 4852-4852 -sequence
mercury\n18:19:04 Ftoc_PickMsgs {{1 hit}}\n18:19:04 Marking 1
hits\n18:19:04 tkerror: syntax error in expression "int ...Note, if I run
the pick command by hand ...delta$ pick +inbox -list -lbrace -lbrace
-subject ftp -rbrace -rbrace  4852-4852 -sequence mercury\n1 hitThat\'s
where the "1 hit" comes from (obviously).  The version of nmh I\'m\nusing
is ...delta$ pick -version\npick -- nmh-1.0.4 [compiled on
fuchsia.cs.mu.OZ.AU at Sun Mar 17 14:55:56 ICT 2002]And the relevant part
of my .mh_profile ...delta$ mhparam pick\n-seq sel -list\nSince the pick
command works, the sequence (actually, both of them, the\none that\'s
explicit on the command line, from the search popup, and the\none that
comes from .mh_profile) do get created.kreps: this is still using the
version of the code form a day ago, I haven\'t\nbeen able to reach the cvs
repository today (local routing issue I
think)._______________________________________________\nExmh-workers
mailing
list\nExmh-workers@redhat.com\nhttps://listman.redhat.com/mailman/listinfo
/exmh-workers
```

“‘

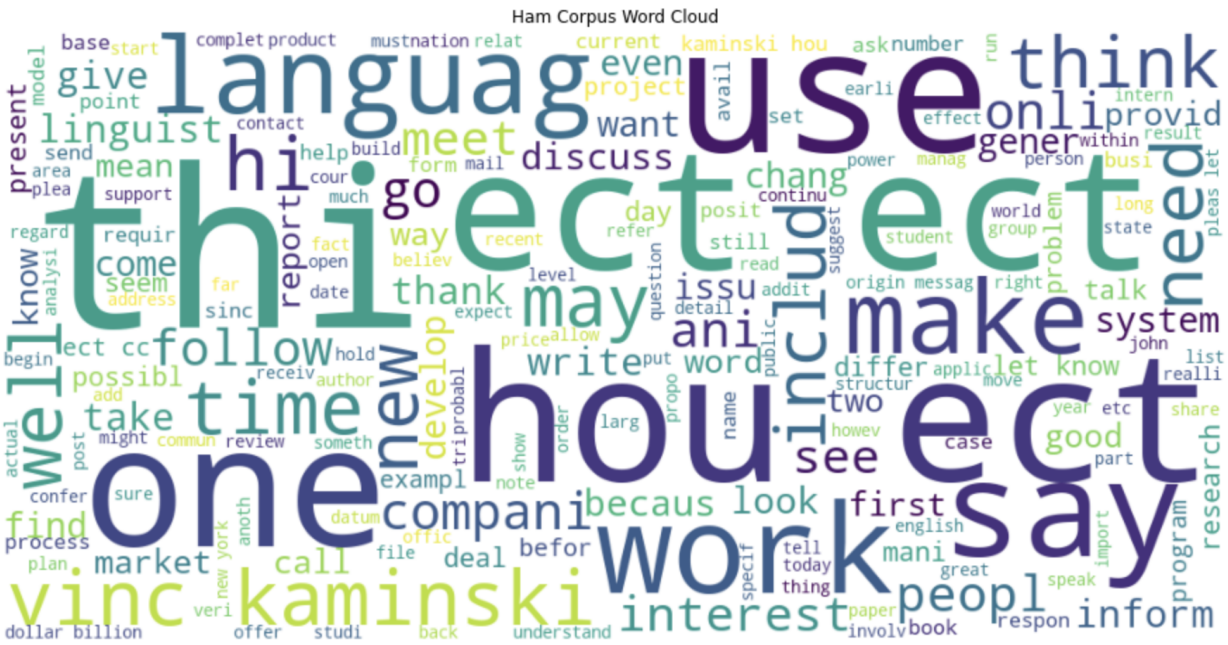
After preprocessing the email content looks like this.

“

```
date aug chri garrigu messag tmda deepeddi vircio com reproduc thi error
veri repeat like everi time without fail thi debug log pick happen pick
exec pick inbox list lbrace lbrace ftp rbrace rbrace sequenc mercuri exec
pick inbox list lbrace lbrace ftp rbrace rbrace sequenc mercuri ftoc
pickmsg hit mark hit tkerror syntax error express int note run pick
command hand delta dollar pick inbox list lbrace lbrace ftp rbrace rbrace
sequenc mercuri hitthat hit come obvious version nmh use delta dollar pick
version pick nmh compil fuchsia cs mu oz au sun mar ict relev part mh
profil delta dollar mhparam pick seq sel list sinc pick command work
sequenc actual one explicit command line search popup one come mh profil
get creat krep thi still use version code form day ago abl reach cv
repositori today local rout issu think exmh worker mail list exmh worker
redhat com link
```

“

After preprocessing steps, wordcloud library was used to visualize the words and their frequency in spam corpus and Ham corpus



The above wordclouds show the frequency of words in respective ham and spam corpus.

It takes around 6 minutes for the whole dataset to get preprocessed.

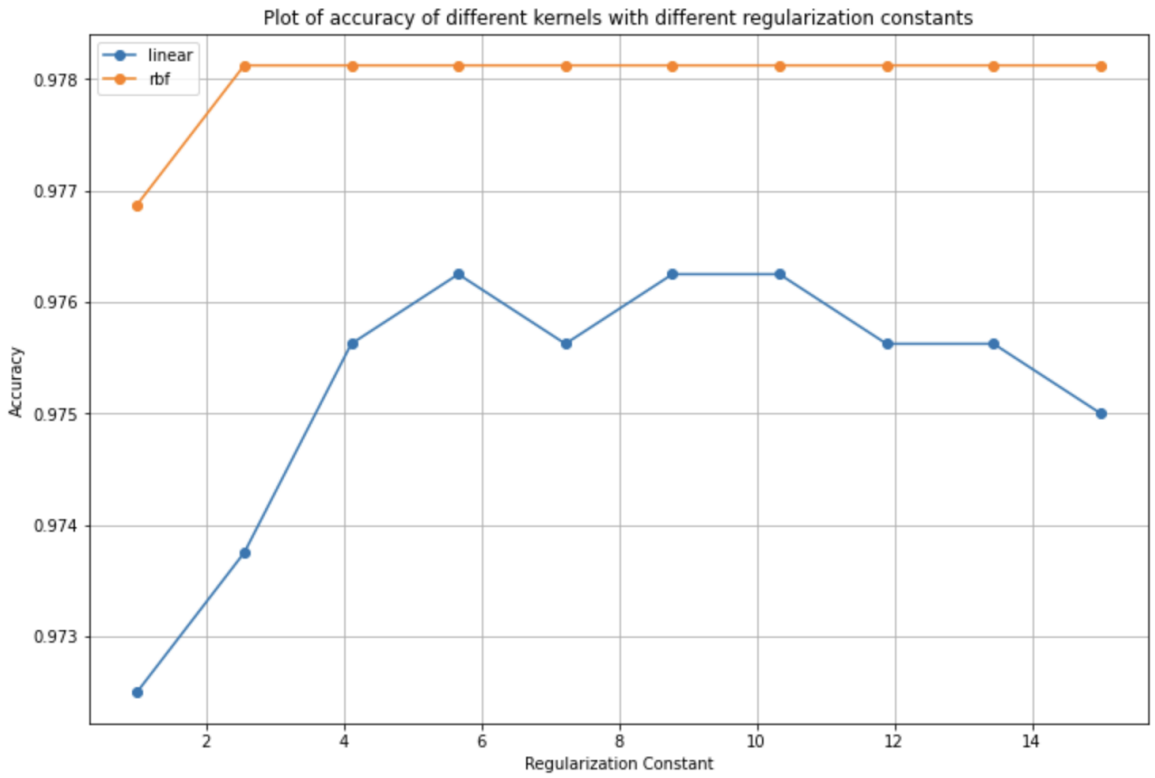
STEP 3: MODEL TRAINING:

SVM- Support Vector Machines was chosen for this task. As per the question, SVM need not be coded by scratch therefore, SVM from sklearn was used for the training.

Libraries used:

- Sklearn for calculating prediction performance, test-train split, and for SVM model.

Performance of SVM with two kernels ‘linear’ and ‘rbf’ with different regularization constants were compared.



From the comparison Plot obtained for the two kernels, it is observed that there is no such significant difference between them.

```
Model: SVM
        Regularization Constant:  2.5555555555555554
        Kernel:  rbf
Confusion Matrix:
 [[778  13]
 [ 22 787]]
Accuracy        :       0.978125
Precision       :       0.9781882812500001
Recall          :       0.978125
```

SVM Classifier with Kernel = ‘rbf’ with regularization constant = 2.55
Was found to have achieved better performance.

‘Sigmoid’ and ‘poly’ kernels were not performing up to the mark that is why they were not considered for comparison .

STEP 4: MODEL TESTING:

For the purpose of testing the model, folder named test can be present in the current directory containing two test emails named ‘email#.txt’.

To get the model prediction: we need to call the function `model_testing()` which takes no arguments. It will print the filename and its prediction as 0 (ham) or 1 (spam).

It takes 15-20 minutes for the code to run completely.