Assignment 3: Spam or Ham Classification (using SVM)

The training data was obtained from UCI repository. https://archive.ics.uci.edu/ml/machine-learning-databases/00228/

The algorithm used in the code is as follows

1. Read the training data (messages) in .txt format and store it in the form of table with labels (ham and spam) clearly mentioned.
2. Segregate the messages and the labels.
3. Read the test data one at a time from 'test' folder in .txt format and add it to the last row of the training data table and call it total data. (This is done to make sure that the word vector is the same for both training data and test data).
4. Tokenize each message into set of individual words.
5. Pre-process the total data to remove stop-words, replace '/' with ' ', erase punctuation and normalize the words.
6. Create bag of words which provides information about the word count in each document against the word vector (features).
7. Create the tfidf (time frequency-inverse document frequency) from the bag of words considering all the features (words in word vector). It is possible to reduce the number of features (words) to obtain tfidf using topkwords command.
8. Once the tfidf matrix is obtained, it is necessary to segregate the test data from the training data before performing svm. tfidf of the test data will be the last row of tfidf matrix of total data.
9. Once segregated, use the tfidf of training data and the corresponding labels (ham or spam) to obtain the weights of the features. In the code, fitcsvm is used which works well for binary classification with moderate number of features.
10. Perform cross validation if required. kfoldLoss gives the cross validation error.
11. Predict whether the test message is ham or spam using the predict command which takes in tfidf of the test data and weights calculated for each feature.
12. The code outputs whether the test message was classified as 'ham' or 'spam'.

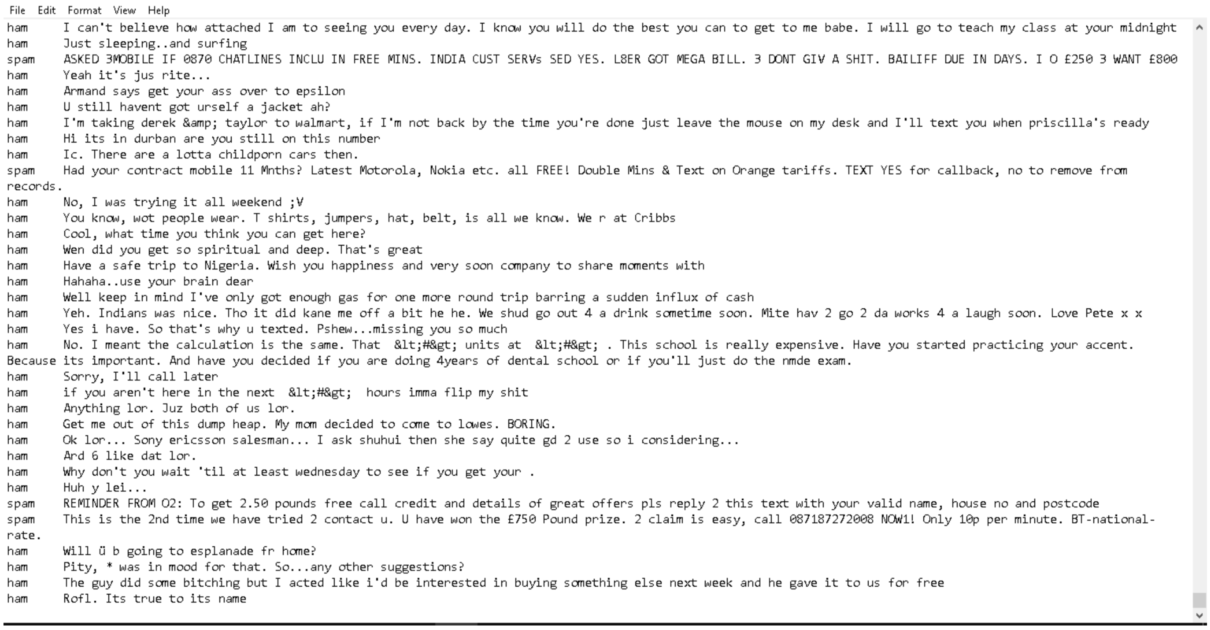
The model loss obtained for two iterations were 1.76% and 1.81%.

Fig(1) shows the text file of the training data.

Fig(2) shows the test file of the testing data email1.

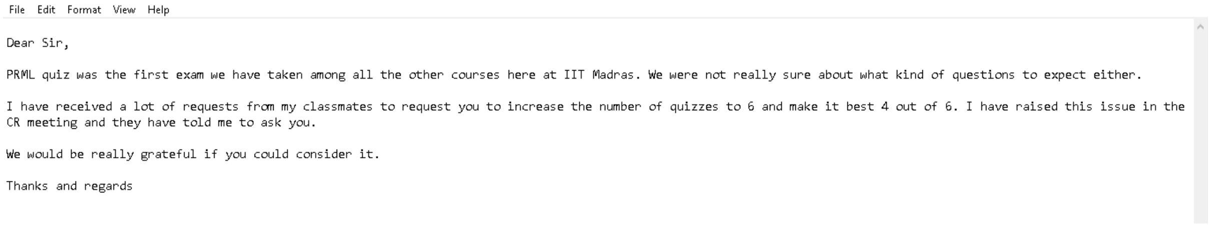
Fig(3) shows the test file of the testing data email2.

Fig(4) shows the result output for the test message email1 and email2.



```
ham     I can't believe how attached I am to seeing you every day. I know you will do the best you can to get to me babe. I will go to teach my class at your midnight
ham     Just sleeping..and surfing
spam    ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE MINS. INDIA CUST SERVs SED YES. L8ER GOT MEGA BILL. 3 DONT GIV A SHIT. BAILIFF DUE IN DAYS. I O £250 3 WANT £800
ham     Yeah it's jus rite...
ham     Armand says get your ass over to epsilon
ham     U still havent got urself a jacket ah?
ham     I'm taking derek &amp; taylor to walmart, if I'm not back by the time you're done just leave the mouse on my desk and I'll text you when priscilla's ready
ham     Hi its in durban are you still on this number
ham     Ic. There are a lotta childporn cars then.
spam    Had your contract mobile 11 Mnths? Latest Motorola, Nokia etc. all FREE! Double Mins & Text on Orange tariffs. TEXT YES for callback, no to remove from
records.
ham     No, I was trying it all weekend ;V
ham     You know, wot people wear. T shirts, jumpers, hat, belt, is all we know. We r at Cribbs
ham     Cool, what time you think you can get here?
ham     Wen did you get so spiritual and deep. That's great
ham     Have a safe trip to Nigeria. Wish you happiness and very soon company to share moments with
ham     Hahaha..use your brain dear
ham     Well keep in mind I've only got enough gas for one more round trip barring a sudden influx of cash
ham     Yeh. Indians was nice. Tho it did kane me off a bit he he. We shud go out 4 a drink sometime soon. Mite hav 2 go 2 da works 4 a laugh soon. Love Pete x x
ham     Yes i have. So that's why u texted. Pshew...missing you so much
ham     No. I meant the calculation is the same. That  &lt;#&gt; units at  &lt;#&gt; . This school is really expensive. Have you started practicing your accent.
Because its important. And have you decided if you are doing 4years of dental school or if you'll just do the nmde exam.
ham     Sorry, I'll call later
ham     if you aren't here in the next  &lt;#&gt;  hours imma flip my shit
ham     Anything lor. Juz both of us lor.
ham     Get me out of this dump heap. My mom decided to come to lowes. BORING.
ham     Ok lor... Sony ericsson salesman... I ask shuhui then she say quite gd 2 use so i considering...
ham     Ard 6 like dat lor.
ham     Why don't you wait 'til at least wednesday to see if you get your .
ham     Huh y lei...
spam    REMINDER FROM O2: To get 2.50 pounds free call credit and details of great offers pls reply 2 this text with your valid name, house no and postcode
spam    This is the 2nd time we have tried 2 contact u. U have won the £750 Pound prize. 2 claim is easy, call 087187272008 NOW1! Only 10p per minute. BT-national-
rate.
ham     Will ü b going to esplanade fr home?
ham     Pity, * was in mood for that. So...any other suggestions?
ham     The guy did some bitching but I acted like i'd be interested in buying something else next week and he gave it to us for free
ham     Rofl. Its true to its name
```

Figure 1: Training data.



Dear Sir,

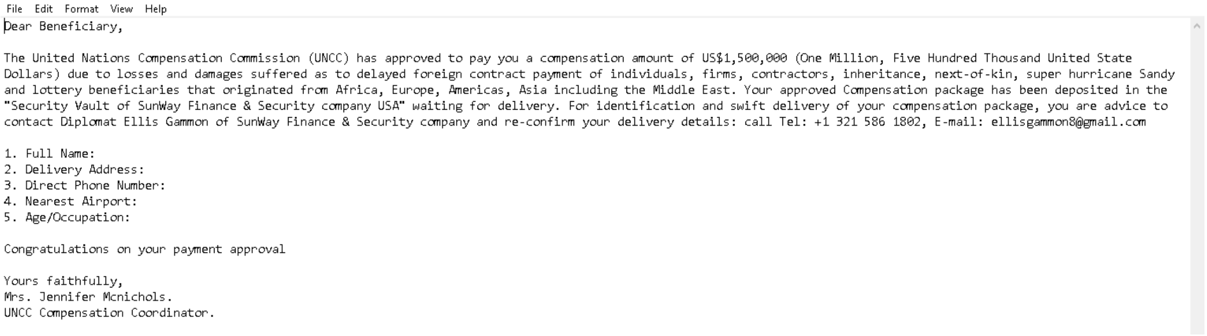
PRML quiz was the first exam we have taken among all the other courses here at IIT Madras. We were not really sure about what kind of questions to expect either.

I have received a lot of requests from my classmates to request you to increase the number of quizzes to 6 and make it best 4 out of 6. I have raised this issue in the CR meeting and they have told me to ask you.

We would be really grateful if you could consider it.

Thanks and regards

Figure 2: Testing data (email1).



Dear Beneficiary,

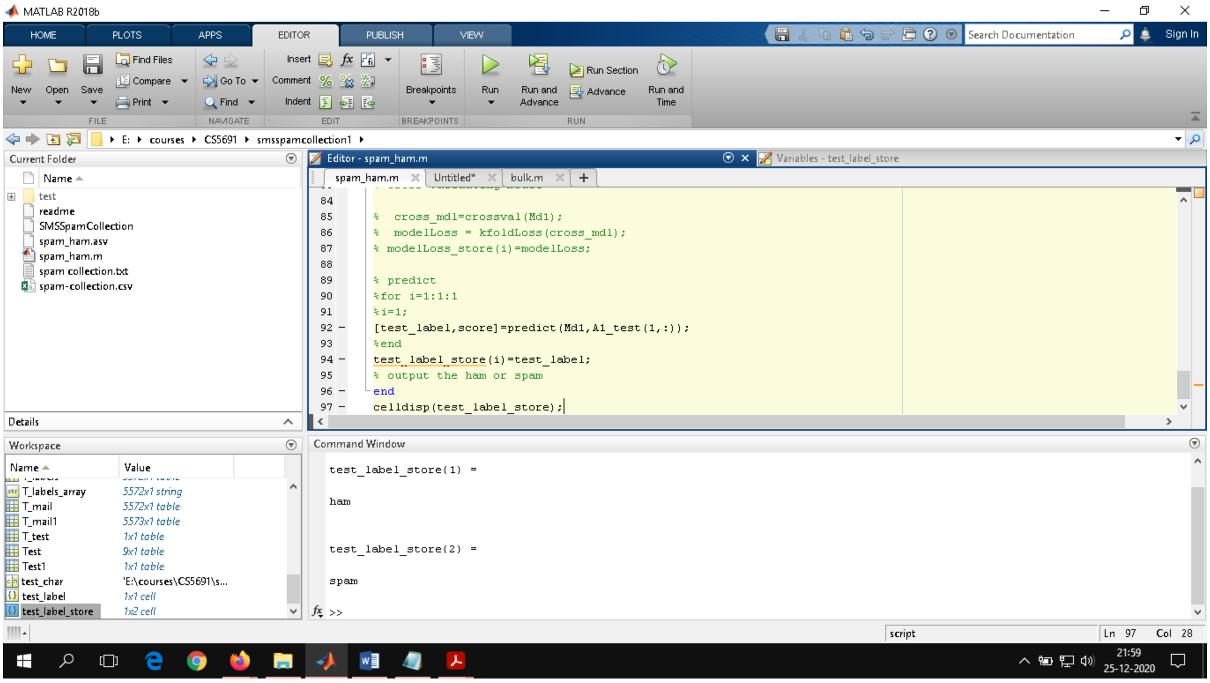
The United Nations Compensation Commission (UNCC) has approved to pay you a compensation amount of US$1,500,000 (One Million, Five Hundred Thousand United State Dollars) due to losses and damages suffered as to delayed foreign contract payment of individuals, firms, contractors, inheritance, next-of-kin, super hurricane Sandy and lottery beneficiaries that originated from Africa, Europe, Americas, Asia including the Middle East. Your approved Compensation package has been deposited in the "Security Vault of SunWay Finance & Security company USA" waiting for delivery. For identification and swift delivery of your compensation package, you are advice to contact Diplomat Ellis Gammon of SunWay Finance & Security company and re-confirm your delivery details: call Tel: +1 321 586 1802, E-mail: ellisgammon8@gmail.com

1. Full Name:
2. Delivery Address:
3. Direct Phone Number:
4. Nearest Airport:
5. Age/Occupation:

Congratulations on your payment approval

Yours faithfully,
Mrs. Jennifer Mcnichols.
UNCC Compensation Coordinator.

Figure 3: Testing data (email2)



Current Folder

- test
- readme
- SMSSpamCollection
- spam_ham.asv
- spam_ham.m
- spam collection.txt
- spam-collection.csv

```
% cross_mdl=crossval(Md1);
%  modelLoss = kfoldLoss(cross_mdl);
% modelLoss_store(i)=modelLoss;

% predict
%for i=1:1:1
%i=1;
[test_label,score]=predict(Md1,A1_test(1,:));
%end
test_label_store(i)=test_label;
% output the ham or spam
end
celldisp(test_label_store);
```

Name	Value
T_labels_array	5572x1 string
T_mail	5572x1 table
T_mail1	5573x1 table
T_test	1x1 table
Test	9x1 table
Test1	1x1 table
test_char	'E:\courses\CS5691\s...
test_label	1x1 cell
test_label_store	1x2 cell

```
test_label_store{1} =

ham


test_label_store{2} =

spam

>>
```

Figure 4: Result output from matlab for test data. (email1 and email2)