ITEC 301 - Machine Learning
Machine Problem 1
Spam Filtering using Naive Bayes Classifier

1. Introduction

Spam detection in email and messages like SMS or chat in general may be considered as a two-category classification problem. The first category represents legitimate messages or ham and the other unsolicited messages or spam. In this assignment you will be given the opportunity to make your own Naive Bayes classifier for filtering spam. For a successful implementation you need to train your classifier on some messages labeled as ham or spam and evaluate its performance on unseen test messages.

Recall that the Naive Bayes formula is given by

$$P(\omega|x_1, x_2, ..., x_d) = \frac{\prod_{i=1}^{d} P(x_i|\omega)P(\omega)}{\sum_{\omega} \prod_{i=1}^{d} P(x_i|\omega)P(\omega)}$$

where d is the vocabulary size $|V|$. For our spam filter, the class conditional likelihood for word x_i is as follows:

$$P(x_i|\omega) = \sum_{D \in D_\omega} \frac{I(x_i \in I}{|D_\omega|}$$

where $I(\cdot)$ is the indicator function, D_ω is the set of documents belonging to class ω in the training set and $|D_\omega|$ is its cardinality.

Note that when a word does not occur in a document class in the training data, it does not mean that it will never appear in any document of that class. We circumvent this problem by using Lambda Smoothing which uses a modified formula for the class conditional likelihood:

$$P(x_i|\omega) = \sum_{D \in D_\omega} \frac{I(x_i \in D) + \lambda}{|D_\omega| + \lambda|V|}$$

where $|V|$ is the vocabulary size. When $\lambda = 1$, this formula is called Laplace Smoothing.
In order to avoid loss of precision during multiplication of the likelihoods, I suggest that you instead add the logarithm of the likelihoods and exponentiate the result.

2. Naive Bayes for the SMS Spam Collection Dataset

You will design a classifier for a subset of the SMS Spam Collection Dataset which is a dataset for benchmarking spam algorithms. This dataset is provided to you in the attached file. After downloading the dataset, you should see CSV files containing a TrainingData.csv (containing 3900 data) with values label and message, and a TestData.csv (containing 1672 data) with message only, without label. Use the TrainingData.csv to build your Naive Bayes Classifier and use the TestData.csv to test the classifier as demonstrated in the class, and discussed below.

3. Classifier Construction and Evaluation

1. Parse the documents in the training set (TrainingData.csv). For simplicity, first remove all the special characters, then define a word as any sequence of alphabetic characters [a-zA-Z] delimited by a white space in front and a white space, comma, or period at the end. Each data should be delimited by a new line. Form the vocabulary V of unique words in the training data, count their statistics and report the prior probabilities for spam and ham.

2. Construct and train a Naive Bayesian Classifier from the count statistics above. Use the Laplace Smoothing $\lambda = 1$.

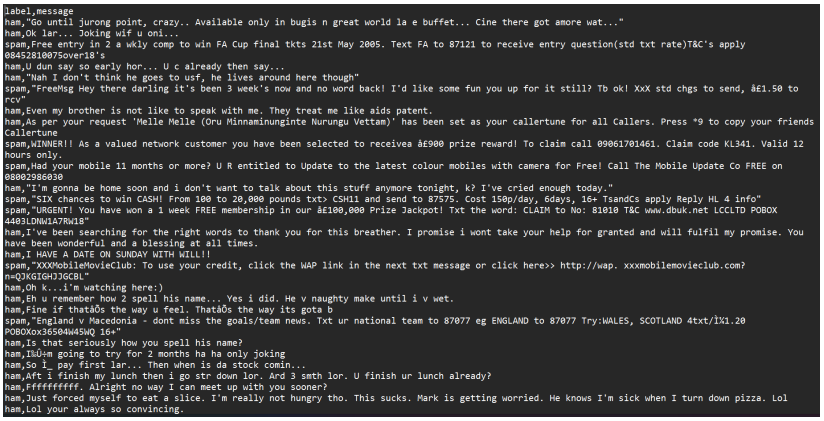
3. Implement the code for classifying an unknown message and try it on the test set (TestData.csv).

4. Output the file ResultData.csv which contains the TestData.csv but with labels like the TrainingData.csv as shown below.

The TrainingData.csv in Excel should appear as:

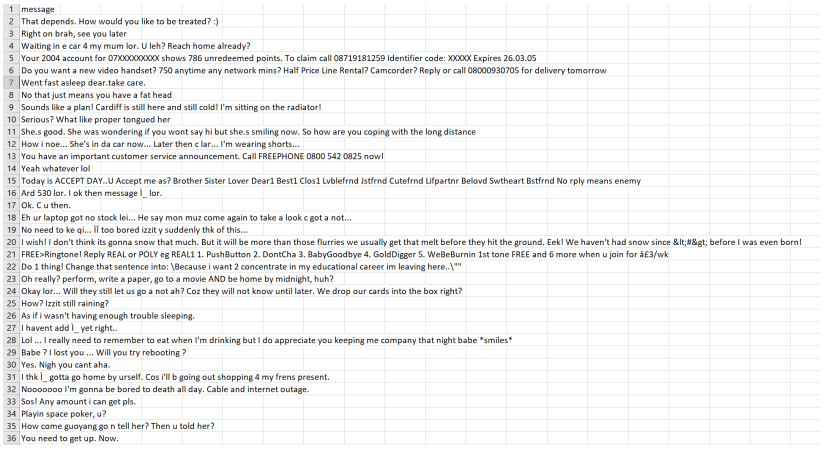
	label	message
1	label	message
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if thatåÕs the way u feel. ThatåÕs the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ì¼1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	I‰Û÷m going to try for 2 months ha ha only joking
24	ham	So ì_ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
26	ham	Ffffffffff. Alright no way I can meet up with you sooner?
27	ham	Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol
28	ham	Lol your always so convincing.
29	ham	Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ?
30	ham	I'm back & we're packing the car now, I'll let you know if there's room
31	ham	Ahhh. Work. I vaguely remember that! What does it feel like? Lol
32	ham	Wait that's still not all that clear, were you not sure about me being sarcastic or that that's why x doesn't want to live with us
33	ham	Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You?
34	ham	K tell me anything about you.
35	ham	For fear of fainting with the of all that housework you just did? Quick have a cuppa
36	spam	Thanks for your subscription to Ringtone UK your mobile will be charged å£5/month Please confirm by replying YES or NO. If you reply NO you will not be charged
37	ham	Yup... Ok i go home look at the timings then i msg ì_ again... Xuhui going to learn on 2nd may too but her lesson is at 8am

While in your IDE (Notepad++/VSCode/AndroidStudio/Sublime etc.) it will appear as:



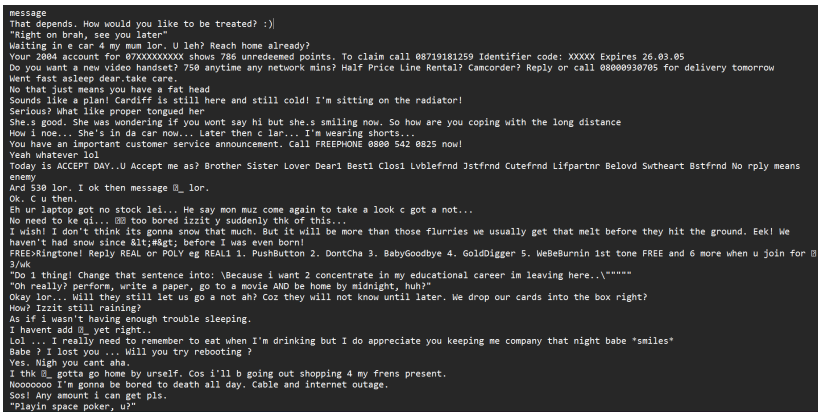
```
label,message
ham,"Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat..."
ham,Ok lar... Joking wif u oni...
spam,Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply
08452810075over18's
ham,U dun say so early hor... U c already then say...
ham,"Nah I don't think he goes to usf, he lives around here though"
spam,"FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to
rcv"
ham,Even my brother is not like to speak with me. They treat me like aids patent.
ham,As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends
Callertune
spam,WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12
hours only.
spam,Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on
08002986030
ham,"I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today."
spam,"SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info"
spam,"URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX
4403LDNW1A7RW18"
ham,I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You
have been wonderful and a blessing at all times.
ham,I HAVE A DATE ON SUNDAY WITH WILL!!
spam,"XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?
n=QJKGIGHJJGCBL"
ham,Oh k...i'm watching here:)
ham,Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham,Fine if thatâÕs the way u feel. thatâÕs the way its gota b
spam,"England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/1X1.20
POBOXox36504W45WQ 16+"
ham,Is that seriously how you spell his name?
ham,IÂÛ÷m going to try for 2 months ha ha only joking
ham,So Ì_ pay first lar... Then when is da stock comin...
ham,Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
ham,"ffffffffff. Alright no way I can meet up with you sooner?
ham,Just forced myself to eat a slice. I'm really not hungry tho. This sucks. Mark is getting worried. He knows I'm sick when I turn down pizza. Lol
ham,Lol your always so convincing.
```

The TestData.csv in Excel should appear as:



1	message
2	That depends. How would you like to be treated? :)
3	Right on brah, see you later
4	Waiting in e car 4 my mum lor. U leh? Reach home already?
5	Your 2004 account for 07XXXXXXXXX shows 786 unredeemed points. To claim call 08719181259 Identifier code: XXXXX Expires 26.03.05
6	Do you want a new video handset? 750 anytime any network mins? Half Price Line Rental? Camcorder? Reply or call 08000930705 for delivery tomorrow
7	Went fast asleep dear.take care.
8	No that just means you have a fat head
9	Sounds like a plan! Cardiff is still here and still cold! I'm sitting on the radiator!
10	Serious? What like proper tongued her
11	She.s good. She was wondering if you wont say hi but she.s smiling now. So how are you coping with the long distance
12	How i noe... She's in da car now... Later then c lar... I'm wearing shorts...
13	You have an important customer service announcement. Call FREEPHONE 0800 542 0825 now!
14	Yeah whatever lol
15	Today is ACCEPT DAY..U Accept me as? Brother Sister Lover Dear1 Best1 Clos1 Lvblefrnd Jstfrnd Cutefrnd Lifpartnr Belovd Swtheart Bstfrnd No rply means enemy
16	Ard 530 lor. I ok then message Ì_ lor.
17	Ok. C u then.
18	Eh ur laptop got no stock lei... He say mon muz come again to take a look c got a not...
19	No need to ke qi... ÌÏ too bored izzit y suddenly thk of this...
20	I wish! I don't think its gonna snow that much. But it will be more than those flurries we usually get that melt before they hit the ground. Eek! We haven't had snow since <#> before I was even born!
21	FREE>Ringtone! Reply REAL or POLY eg REAL1 1. PushButton 2. DontCha 3. BabyGoodbye 4. GoldDigger 5. WeBeBurnin 1st tone FREE and 6 more when u join for â£3/wk
22	Do 1 thing! Change that sentence into: \Because i want 2 concentrate in my educational career im leaving here..\""
23	Oh really? perform, write a paper, go to a movie AND be home by midnight, huh?
24	Okay lor... Will they still let us go a not ah? Coz they will not know until later. We drop our cards into the box right?
25	How? Izzit still raining?
26	As if i wasn't having enough trouble sleeping.
27	I havent add Ì_ yet right..
28	Lol ... I really need to remember to eat when I'm drinking but I do appreciate you keeping me company that night babe *smiles*
29	Babe ? I lost you ... Will you try rebooting ?
30	Yes. Nigh you cant aha.
31	I thk Ì_ gotta go home by urself. Cos i'll b going out shopping 4 my frens present.
32	Noooooooo I'm gonna be bored to death all day. Cable and internet outage.
33	Sos! Any amount i can get pls.
34	Playin space poker, u?
35	How come guoyang go n tell her? Then u told her?
36	You need to get up. Now.

While in your IDE (Notepad++/VSCode/AndroidStudio/Sublime etc.) it will appear as:



```
message
That depends. How would you like to be treated? :)
"Right on brah, see you later"
Waiting in e car 4 my mum lor. U leh? Reach home already?
Your 2004 account for 07XXXXXXXXX shows 786 unredeemed points. To claim call 08719181259 Identifier code: XXXXX Expires 26.03.05
Do you want a new video handset? 750 anytime any network mins? Half Price Line Rental? Camcorder? Reply or call 08000930705 for delivery tomorrow
Went fast asleep dear.take care.
No that just means you have a fat head
Sounds like a plan! Cardiff is still here and still cold! I'm sitting on the radiator!
Serious? What like proper tongued her
She.s good. She was wondering if you wont say hi but she.s smiling now. So how are you coping with the long distance
How i noe... She's in da car now... Later then c lar... I'm wearing shorts...
You have an important customer service announcement. Call FREEPHONE 0800 542 0825 now!
Yeah whatever lol
Today is ACCEPT DAY..U Accept me as? Brother Sister Lover Dear1 Best1 Clos1 Lvblefrnd Jstfrnd Cutefrnd Lifpartnr Belovd Swtheart Bstfrnd No rply means
enemy
Ard 530 lor. I ok then message Ì_ lor.
Ok. C u then.
Eh ur laptop got no stock lei... He say mon muz come again to take a look c got a not...
No need to ke qi... ÌÏ too bored izzit y suddenly thk of this...
I wish! I don't think its gonna snow that much. But it will be more than those flurries we usually get that melt before they hit the ground. Eek! We
haven't had snow since &lt;#&gt; before I was even born!
FREE>Ringtone! Reply REAL or POLY eg REAL1 1. PushButton 2. DontCha 3. BabyGoodbye 4. GoldDigger 5. WeBeBurnin 1st tone FREE and 6 more when u join for Ì
3/wk
"Do 1 thing! Change that sentence into: \Because i want 2 concentrate in my educational career im leaving here..\""""
"Oh really? perform, write a paper, go to a movie AND be home by midnight, huh?"
Okay lor... Will they still let us go a not ah? Coz they will not know until later. We drop our cards into the box right?
How? Izzit still raining?
As if i wasn't having enough trouble sleeping.
I havent add Ì_ yet right..
Lol ... I really need to remember to eat when I'm drinking but I do appreciate you keeping me company that night babe *smiles*
Babe ? I lost you ... Will you try rebooting ?
Yes. Nigh you cant aha.
I thk Ì_ gotta go home by urself. Cos i'll b going out shopping 4 my frens present.
Noooooooo I'm gonna be bored to death all day. Cable and internet outage.
Sos! Any amount i can get pls.
"Playin space poker, u?"
```

The ResultData.csv will contain the TestData.csv but with a label field like the TrainingData.csv. Focus on creating ResultData.csv for now. After this activity, you will test the precision and recall of your generated classifier in the next activity to be posted.