LAB4

August 25, 2019

[30]:
```
import nltk
from nltk.corpus import stopwords
import string
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import classification_report,confusion_matrix
```

[10]:
```
msg = pd.read_csv('spam.csv',encoding='latin-1')
msg.drop(['Unnamed: 2','Unnamed: 3','Unnamed: 4'],axis=1,inplace=True)
msg = msg.rename(columns={'v1':'class','v2':'text'})
msg.head()
```

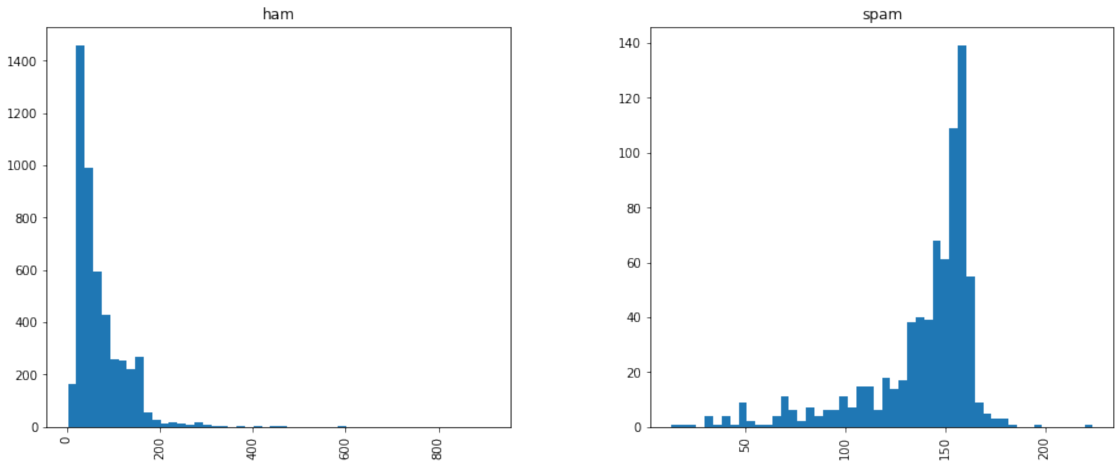
[10]:
```
  class                                               text
0   ham  Go until jurong point, crazy.. Available only ...
1   ham                      Ok lar... Joking wif u oni...
2  spam  Free entry in 2 a wkly comp to win FA Cup fina...
3   ham  U dun say so early hor... U c already then say...
4   ham  Nah I don't think he goes to usf, he lives aro...
```

[12]:
```
msg.groupby('class').describe()
```

[12]:
```
       text
      count unique                                                top freq
class
ham    4825   4516                              Sorry, I'll call later   30
spam    747    653  Please call our customer service representativ...    4
```

[15]:
```
msg['length'] = msg['text'].apply(len)
msg.hist(column='length',by='class',bins=50,figsize=(15,6))
```

[15]:
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7f3c472a3c50>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7f3c46df6898>],
      dtype=object)
```



[17]:
```
def process_text(text):
    nopunc = [char for char in text if char not in string.punctuation]
    nopunc = ''.join(nopunc)
    clean_words = [word for word in nopunc.split() if word.lower() not in␣
 ↪stopwords.words('english')]
    return clean_words
```

[20]:
```
msg['text'].apply(process_text).head()
```

[20]:
```
0    [Go, jurong, point, crazy, Available, bugis, n...
1                        [Ok, lar, Joking, wif, u, oni]
2    [Free, entry, 2, wkly, comp, win, FA, Cup, fin...
3        [U, dun, say, early, hor, U, c, already, say]
4    [Nah, dont, think, goes, usf, lives, around, t...
Name: text, dtype: object
```

[21]:
```
msg_train,msg_test,class_train,class_test =␣
 ↪train_test_split(msg['text'],msg['class'],test_size=0.2)
```

[22]:
```
pipeline = Pipeline([
    ('bow',CountVectorizer(analyzer=process_text)),
    ('tfidf',TfidfTransformer()),
    ('classifier',MultinomialNB())
])
```

[23]:
```
pipeline.fit(msg_train,class_train)
```

[23]:
```
Pipeline(memory=None,
         steps=[('bow',
                 CountVectorizer(analyzer=<function process_text at
0x7f3c46482a60>,
                                 binary=False, decode_error='strict',
                                 dtype=<class 'numpy.int64'>, encoding='utf-8',
                                 input='content', lowercase=True, max_df=1.0,
                                 max_features=None, min_df=1,
                                 ngram_range=(1, 1), preprocessor=None,
                                 stop_words=None, strip_accents=None,
                                 token_pattern='(?u)\\b\\w\\w+\\b',
                                 tokenizer=None, vocabulary=None)),
                ('tfidf',
                 TfidfTransformer(norm='l2', smooth_idf=True,
                                  sublinear_tf=False, use_idf=True)),
                ('classifier',
                 MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True))],
         verbose=False)
```

[24]:

```
predictions = pipeline.predict(msg_test)
```

[25]:

```
print(classification_report(class_test,predictions))
```

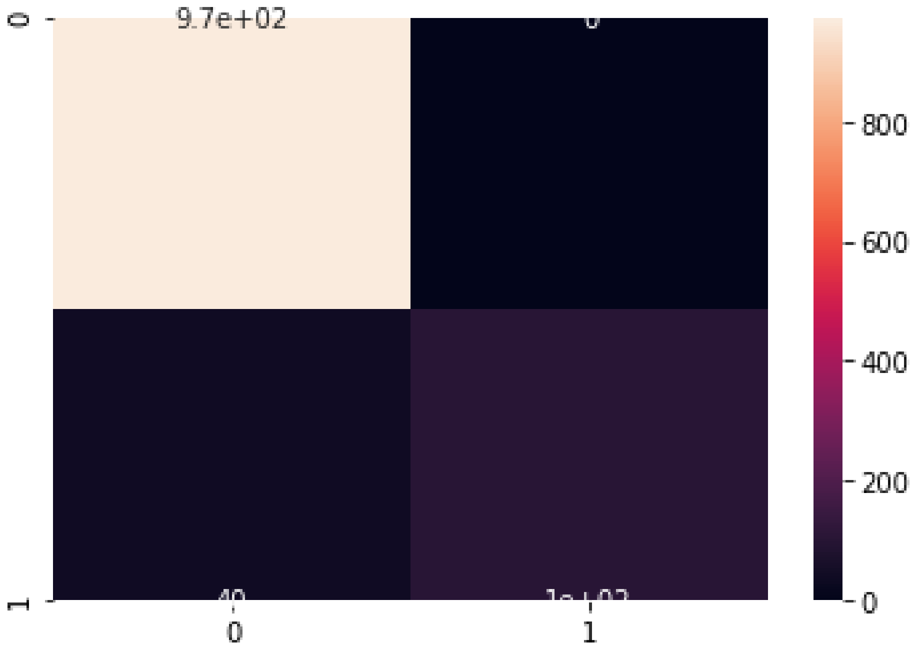
```
              precision    recall  f1-score   support

         ham       0.96      1.00      0.98       972
        spam       1.00      0.72      0.84       143

    accuracy                           0.96      1115
   macro avg       0.98      0.86      0.91      1115
weighted avg       0.97      0.96      0.96      1115
```

[26]:

```
import seaborn as sns
sns.heatmap(confusion_matrix(class_test,predictions),annot=True)
```

[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3c41cb7128>
```



[29]:
```
excr = [
    'You have won 1 bn dollar lottery',
    'Hi, I miss you.',
    'Contact customer care service for more details.',
    'Tomorrows meeting is scheduled at 1: 30 pm',
    'You can fool all the people some of the time, and you can fool some of the␣
 ↪people all the time, but you can not fool all the people all the time.',
    'Not my circus not my monkey.',
    'They say teaching is like walking in a park, what they dont say is that␣
 ↪the park is the Jurrasic park.'
]
predictions1 = pipeline.predict(excr)
predictions1
```

[29]:
```
array(['ham', 'ham', 'ham', 'ham', 'ham', 'ham', 'ham'], dtype='<U4')
```

[32]:
```
pipeline1 = Pipeline([
    ('bow',CountVectorizer(analyzer=process_text)),
    ('tfidf',TfidfTransformer()),
    ('classifier',DecisionTreeClassifier())
])
```

[34]:
```
pipeline1.fit(msg_train,class_train)
```

[34]:
```
Pipeline(memory=None,
         steps=[('bow',
                 CountVectorizer(analyzer=<function process_text at
0x7f3c46482a60>,
                                 binary=False, decode_error='strict',
                                 dtype=<class 'numpy.int64'>, encoding='utf-8',
                                 input='content', lowercase=True, max_df=1.0,
                                 max_features=None, min_df=1,
                                 ngram_range=(1, 1), preprocessor=None,
                                 stop_words=None, strip_accents=None,
                                 token_pattern='(?u)\\b\\w\\w+\\b...
                 TfidfTransformer(norm='l2', smooth_idf=True,
                                  sublinear_tf=False, use_idf=True)),
                ('classifier',
                 DecisionTreeClassifier(class_weight=None, criterion='gini',
                                        max_depth=None, max_features=None,
                                        max_leaf_nodes=None,
                                        min_impurity_decrease=0.0,
                                        min_impurity_split=None,
                                        min_samples_leaf=1, min_samples_split=2,
                                        min_weight_fraction_leaf=0.0,
                                        presort=False, random_state=None,
                                        splitter='best'))],
         verbose=False)
```

[38]:

```
predictions2 = pipeline1.predict(msg_test)
print(classification_report(class_test,predictions2))
```

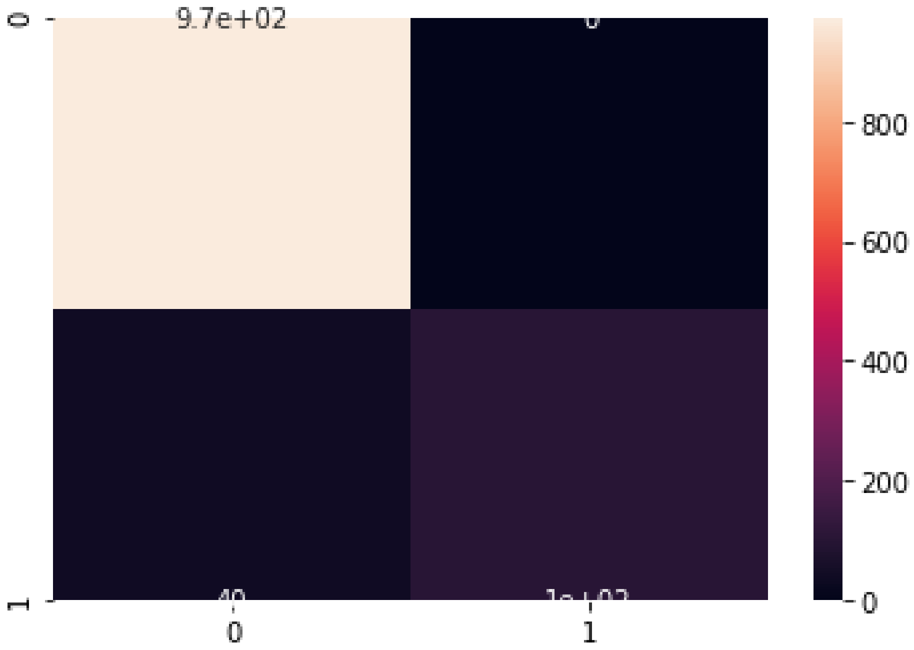
```
              precision    recall  f1-score   support

         ham       0.97      0.99      0.98       972
        spam       0.93      0.80      0.86       143

    accuracy                           0.97      1115
   macro avg       0.95      0.90      0.92      1115
weighted avg       0.97      0.97      0.97      1115
```

[40]:

```
sns.heatmap(confusion_matrix(class_test,predictions),annot=True)
```

[40]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3c303ce9e8>
```



[42]:
```
predictions3 = pipeline1.predict(excr)
predictions3
```

[42]:
```
array(['ham', 'ham', 'spam', 'ham', 'ham', 'ham', 'ham'], dtype=object)
```

[48]:
```
email = pd.read_csv('email.csv')
email = email.rename(columns={'email':'text'})
email.head()
```

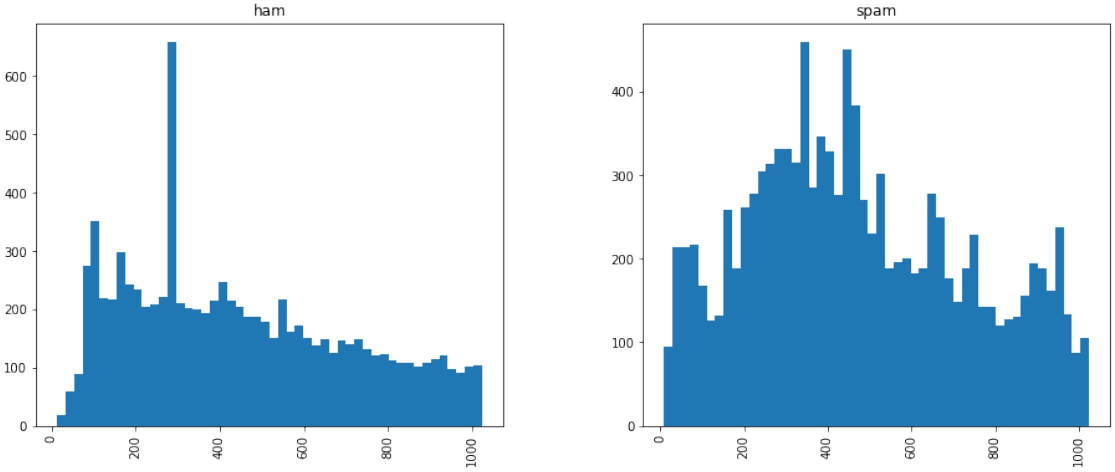
[48]:
```
                                                text class
0  Subject: what up , , your cam babe what are yo...  spam
1  Subject: want to make more money ? order confi...  spam
2  Subject: food for thoughts [ join now - take a...  spam
3  Subject: your pharmacy ta would you want cheap...  spam
4  Subject: bigger breast just from a pill image ...  spam
```

[50]:
```
email.groupby('class').describe()
```

[50]:
```
        text
       count unique                                               top freq
class
ham     8774   8336  Subject: calpine daily gas nomination > ricky ...   20
spam   11224   9494                                          Subject:     51
```

[56]:
```
email['length'] = email['text'].apply(len)
email.hist(column='length',by='class',bins=50,figsize=(15,6))
```

[56]:
```
array([<matplotlib.axes._subplots.AxesSubplot object at 0x7f3c2eff8518>,
       <matplotlib.axes._subplots.AxesSubplot object at 0x7f3c2eec32b0>],
        dtype=object)
```



[101]:

```
email['text'].apply(process_text).head()
```

[101]:

```
0    [Subject, cam, babe, looking, looking, compani...
1    [Subject, want, make, money, order, confirmati...
2    [Subject, food, thoughts, join, take, free, to...
3    [Subject, pharmacy, ta, would, want, cheap, pe...
4    [Subject, bigger, breast, pill, image, loading...
Name: text, dtype: object
```

[102]:

```
email_train,email_test,eClass_train,eClass_test =␣
 ↪train_test_split(email['text'],email['class'],test_size=0.2)
pipeline2 = Pipeline([
    ('bow',CountVectorizer(analyzer=process_text)),
    ('tfidf',TfidfTransformer()),
    ('classifier',MultinomialNB())
])
pipeline2.fit(email_train,eClass_train)
```

[102]:

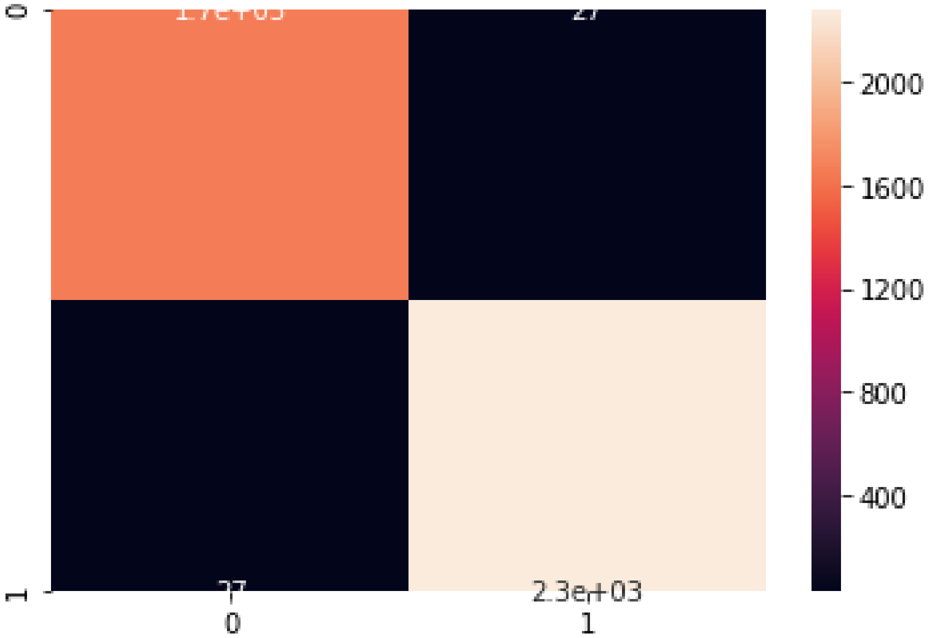
```
Pipeline(memory=None,
         steps=[('bow',
                 CountVectorizer(analyzer=<function process_text at
0x7f3c46482a60>,
                                 binary=False, decode_error='strict',
                                 dtype=<class 'numpy.int64'>, encoding='utf-8',
                                 input='content', lowercase=True, max_df=1.0,
                                 max_features=None, min_df=1,
                                 ngram_range=(1, 1), preprocessor=None,
                                 stop_words=None, strip_accents=None,
                                 token_pattern='(?u)\\b\\w\\w+\\b',
                                 tokenizer=None, vocabulary=None)),
                ('tfidf',
                 TfidfTransformer(norm='l2', smooth_idf=True,
                                  sublinear_tf=False, use_idf=True)),
                ('classifier',
                 MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True))],
         verbose=False)
```

[103]:
```
predictions4 = pipeline2.predict(email_test)
print(classification_report(eClass_test,predictions4))
```

```
              precision    recall  f1-score   support

         ham       0.98      0.98      0.98      1693
        spam       0.99      0.99      0.99      2307

    accuracy                           0.99      4000
   macro avg       0.99      0.99      0.99      4000
weighted avg       0.99      0.99      0.99      4000
```

[104]:
```
sns.heatmap(confusion_matrix(eClass_test,predictions4),annot=True)
```

[104]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3c10e066d8>
```



[106]:
```
prediction5 = pipeline2.predict(excr)
prediction5
```

[106]:
```
array(['spam', 'spam', 'spam', 'ham', 'spam', 'spam', 'spam'], dtype='<U4')
```

[132]:

```
ctest = class_test.array
mtest = msg_test.array
df = {'msg':[],'class':[],'prediction':[]}
wmsg = pd.DataFrame(df1)
for i in range(len(predictions)):
    if(predictions[i] != ctest[i]):
        new_row = {'msg':mtest[i],'class':ctest[i],'prediction':predictions[i]}
        wmsg = wmsg.append(new_row, ignore_index=True)
wmsg
```

[132]:

```
                                                  msg class prediction
0   XCLUSIVE@CLUBSAISAI 2MOROW 28/5 SOIREE SPECIAL...  spam        ham
1   Free msg: Single? Find a partner in your area!...  spam        ham
2   Got what it takes 2 take part in the WRC Rally...  spam        ham
3   Hi, the SEXYCHAT girls are waiting for you to ...  spam        ham
4   Orange brings you ringtones from all time Char...  spam        ham
5   Free-message: Jamster!Get the crazy frog sound...  spam        ham
6     You have 1 new message. Please call 08715205273  spam        ham
7   thesmszone.com lets you send free anonymous an...  spam        ham
8   I don't know u and u don't know me. Send CHAT ...  spam        ham
9   ROMCAPspam Everyone around should be respondin...  spam        ham
10  Hi babe its Chloe, how r u? I was smashed on s...  spam        ham
11  SMS. ac Sptv: The New Jersey Devils and the De...  spam        ham
12    3. You have received your mobile content. Enjoy  spam        ham
13  Sorry! U can not unsubscribe yet. THE MOB offe...  spam        ham
14  Money i have won wining number 946 wot do i do...  spam        ham
15  Romantic Paris. 2 nights, 2 flights from åč79 ...  spam        ham
16  Fantasy Football is back on your TV. Go to Sky...  spam        ham
17  Talk sexy!! Make new friends or fall in love i...  spam        ham
18  Free Msg: Ringtone!From: http://tms. widelive...  spam        ham
19  Bored of speed dating? Try SPEEDCHAT, txt SPEE...  spam        ham
20  SMS. ac JSco: Energy is high, but u may not kn...  spam        ham
21  Customer service announcement. We recently tri...  spam        ham
22  Ur balance is now åč500. Ur next question is: ...  spam        ham
23  FREE2DAY sexy St George's Day pic of Jordan!Tx...  spam        ham
24  You have 1 new voicemail. Please call 08719181503  spam        ham
25  We currently have a message awaiting your coll...  spam        ham
26  Dear Voucher Holder 2 claim your 1st class air...  spam        ham
27  I don't know u and u don't know me. Send CHAT ...  spam        ham
28  You can donate åč2.50 to UNICEF's Asian Tsunam...  spam        ham
29  Hi, this is Mandy Sullivan calling from HOTMIX...  spam        ham
30  Missed call alert. These numbers called but le...  spam        ham
31  Hello. We need some posh birds and chaps to us...  spam        ham
32  We know someone who you know that fancies you...  spam        ham
33  Reminder: You have not downloaded the content ...  spam        ham
34  accordingly. I repeat, just text the word ok o...  spam        ham
35  Here is your discount code RP176781. To stop f...  spam        ham
```

```
36  I'd like to tell you my deepest darkest fantas...  spam        ham
37  FreeMsg Hey there darling it's been 3 week's n...  spam        ham
38  Fantasy Football is back on your TV. Go to Sky...  spam        ham
39  Your B4U voucher w/c 27/03 is MARSMS. Log onto...  spam        ham
```

[130]:

```
ctest1 = eClass_test.array
mtest1 = email_test.array
df1 = {'msg':[],'class':[],'prediction':[]}
wmsg1 = pd.DataFrame(df1)
for i in range(len(predictions4)):
    if(predictions4[i] != ctest1[i]):
        new_row = {'msg':mtest1[i],'class':ctest1[i],'prediction':
 ↪predictions4[i]}
        wmsg1 = wmsg1.append(new_row, ignore_index=True)
wmsg1
```

[130]:

```
                                                  msg class prediction
0   Subject: referred by , james hi , i found foll...  spam        ham
1   Subject: important video announcement i have a...   ham       spam
2   Subject: new schedule councilwomen aloft ringe...  spam        ham
3   Subject: re : billing question thank you for y...   ham       spam
4   Subject: yummy frappachino hey let ' s go get ...   ham       spam
5   Subject: [ blacken ] 84 % - off vicodin . puri...  spam        ham
6   Subject: peter g [ tour dates ] tour dates fri...  spam        ham
7   Subject: re [ 13 ] : dr . dree ricky martin in...  spam        ham
8   Subject: rodrigo lamas - best wishes i would l...   ham       spam
9   Subject: help chinesse new year i ' ll take it...  spam        ham
10  Subject: free latex go to http : / / www . win...   ham       spam
11  Subject: hey ! guess it was hard to get back ,...   ham       spam
12  Subject: re : buzzwords also please search for...   ham       spam
13  Subject: re : boat i believe the boat is 18 to...   ham       spam
14  Subject: no risk kiosk pg lnbcer hey guy , yom...  spam        ham
15  Subject: fw : rock bottom hey guys - i know th...   ham       spam
16  Subject: re [ 16 ] yes , it ' s great . wait f...  spam        ham
17  Subject: jake hey ! what did the doctor say ab...   ham       spam
18  Subject: let ' s get this settled hey , what '...  spam        ham
19  Subject: kate ' s birthday party ! i will be 4...   ham       spam
20  Subject: one more time questioned buzzer sheer...  spam        ham
21  Subject: new research tool - too cool ! ! ! th...   ham       spam
22  Subject: fortune here ' s the fortune link . . .    ham       spam
23  Subject: update your account information we ar...  spam        ham
24  Subject: skilling ranked # 2 in the top ceo li...   ham       spam
25  Subject: join focus groups to earn money a la ...  spam        ham
26  Subject: income tax hey ! tonya said to staple...   ham       spam
27  Subject: re : booty hey dude , how about booty...   ham       spam
28  Subject: life in general good god - - - - wher...   ham       spam
29  Subject: your confirmation is needed please re...   ham       spam
30  Subject: new pictures for faster viewing , i w...   ham       spam
```

```
31  Subject: re [ 23 ] in 1986 to deal with you ho...  spam         ham
32  Subject: here is $ 10 for you . please use it ...  spam         ham
33  Subject: ? ? ? ? 13 ? ? ? ? ? ? ? * ? * ? * ? ...  spam         ham
34                        Subject: get a date tonight  spam         ham
35                labSubject: cyprus hilarion exhibition  spam         ham
36  Subject: platts energy trader free trial pleas...   ham        spam
37                                      Subject: note  spam         ham
38  Subject: please strictly confidential . attn :...  spam         ham
39  Subject: bike ride this weekend ! ! ! 3 rd ann...  spam         ham
40  Subject: re : your document your document is a...  spam         ham
41  Subject: re : thank you vince , you were a mos...   ham        spam
42  Subject: out of office autoreply : just to her...  spam         ham
43  Subject: re : woohoo that is so rad bill . i '...   ham        spam
44  Subject: clickathome is coming soon ! want a n...   ham        spam
45  Subject: http : / / 208 . 246 . 87 . 65 / info...   ham        spam
46  Subject: re [ 3 ] come on ! as follows let ' s...  spam         ham
47  Subject: information request received we are i...  spam         ham
48  Subject: * information only * work on the floo...   ham        spam
49  Subject: re : site license for power world i c...   ham        spam
50                         Subject: anderson jeromy  spam         ham
51  Subject: delivery failure : user antonio _ lam...  spam         ham
52  Subject: enrondirectfinance . com usernames an...   ham        spam
53  Subject: join focus groups to earn money a la ...  spam         ham
```

[ ]:

```
'''
The given UTC SMS data set is biased towards the ham messages because it␣
 ↪contains
4516 distinct ham messages and only 653 spam messages.
So the Naive Bayes model predicts ham correctly but not spams.
Ex.: Recall score for spam is 0.72 only and for ham it is 1.0

By using Decision Tree classifier this does't improve that much.
Ex. Recall for ham is 0.99 and for spam it is 0.80 only

To improve spam detection performance we need to use a fair balanced dataset␣
 ↪between ham
and spam messages.
The second used data set has 8774 ham messages and 11224 spam messages.
We can see that the recall for that dataset is:
0.98 for ham and 0.99 for spam.
'''
```