Assignment for Week 3 - Naive-Bayes

Task 1 - Bayesian Classification

Important Note: This exercise is not a programming exercise, it is a math exercise to help reinforce the math behind a Bayesian Classification. You can fill free to complete this in any method you feel is appropriate (ie: pencil/paper (you will need to scan your work to submit), Excel workbook, markdown text with a jupyter notebook, etc)

Please show all your work.

1. In a study of pleas and prison sentences, it is reported that 42% of the subjects were sent to prison. Among those sent to prison, 38% plead guilty. Among those not sent to prison, 50% plead guilty.
   a) If a subject is randomly selected, what is the probability of getting a person who was not sent to prison?
   b) If a subject is randomly selected, and it is known that the subject entered a quilty plea, what is the probability that this subject was not sent to prison?
   c) If a subject is randomly selected, what is the probability of getting someone who was sent to prison?
   d) If a subject is randomly selected, and it is known that the subject entered a guilty plea, what is the probability that this person was sent to prison?

Bayes Theorem:

The probability of event A, given that event B has subsequently occurred can be defined as

$$ P(A~|~B) = \frac{P(A) \cdot P(B~|~A)}{(~P(A) \cdot P(B~|A~)~) + (~P(A') \cdot P(B~|A'~)~)} $$
Note: P(A')denotes P(not A) and the same applies for P(B').

What we know at this point

Let **A** be the event of **sent to prison**.

$$ P(A) = .42 $$$$ P(A') = .58 $$
Let **B** be the event of **entering a quilty plea**.

A) 58% of the subjects was not sent to preson. B) 50% of people who plead guilty was not sent to preson. C)x

To make it easier to calculate, I going to assume that the number f subjects is 100,000.

if the percentage of subjects sent to prison is 42%, then 0.42 *100.000 = 42,000 and if the percentage of subjects who sent to prison and plead guilty is 38%, then 0.38* 42,000 = 15,960, which is the true positive. 15.960% of the subjects who plead guilty was actually sent to prison.

The number of subject who plead not guilty and sent to prison is 42,000 - 15,960 = 26,040 = 26.04% of subjects sent to prison plead not guilty

subject was not sent to prison = 58% * 100,000 - 58,000.

50% of subject who didn't plea guilty and not sent to prison = 50% = 0.5 * 58,000 = 33,640 = 33% of all subjects (True negative) the number of subject who plead guilty and not sent to jail = 58,000 - 33,640 = 24,360 = 24.36% of all subjects (False negative)

number of subjects who plead guilty is 24,360 + 15,960 = 40,320 = 40.32% of all subjects. number of subject who plead not guilty is 100,000 - 40,320 = 59,680 = 59.68% of all subjects

A- 1-0.42 = 0.58 * 100 = 58% of the subject was not sent to jail. B- 24,360/40,320 = 60.4% probability someone who entered a guilty plea was not sent to prison. C- 42% probability of selecting someone who was sent to prison. D- 24.36% probability of someone who entered a guilty plea was sent to prison.

In [ ]:
```
#Continue defining the probabilites and compute b-d
```

1. Given the following table:

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	CO
2	M	Sports	Medium	CO
3	M	Sports	Medium	CO
4	M	Sports	Large	CO
5	M	Sports	Extra Large	CO
6	M	Sports	Extra Large	CO
7	F	Sports	Small	CO
8	F	Sports	Small	CO
9	F	Sports	Medium	CO
10	F	Luxury	Large	CO
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a) What is the value of each of the following probabilities?
- P(Gender=M | Class=C0)
- P(Gender=F | Class=C1)
- P(Car Type=Family | Class=C0)
- P(Car Type=Family | Class=C1)
- P(Shirt Size=Medium | Class=C0)
- P(Shirt Size=Medium | Class=C1)

50% of subjects are M 50% of subjects are C0 50% of subjects are F 50% of subjects are C1

A) P(Gender=M|Class=C0) = p(C0|M) *P(M)/P(C0) P(M|C0) = 6/10 = 0.6 = P(M) = 10/20 = 0.5 / P(C0) = 10/20 = 0.5 P(M|C0) = 0.6* 0.5/0.5 = 0.6 = 60% probability of P(M|C0)

B) P(Gender=F|Class=C1) = P(C1|F) *P(F)/P(C1) 6/10 = 0.6 = P(F) 10/20 = 0.5 / 0.5 0.6* 0.5/0.5= 0.6 = 60% probability of P(F|C1)

C) P(Car Type=Family | Class=C0) = P(C0 | Family) *P(Family)/P(C0) P(C0 | Family) = 1/10 = 0.1, P(C0) = 10/20 = 0.5, P(Family) = 4/20 = 0.2 P(Family | C0) = 0.1* 0.2/0.5 = 0.04 4% probability of P(C0|family)

D) P(Car Type=Family | Class=C1) = P(C1 | Family) *P(Family)/P(C1) P(C1 | Family) = 3/10 = 0.3, P(C1) = 10/20 = 0.5, P(Family) = 4/20 = 0.2 P(Family | C1) = 0.3* 0.2/0.5 = 0.12 12% of P(family|C1)

E) P(Shirt Size=Medium | Class=C0) = P(C0 | Medium) *P(Medium)/P(C0) P(C0 | Medium) = 3/10 = 0.3, P(C0) = 10/20 = 0.5, P(Medium) = 7/20 = 0.35 P(Medium | C0) = 0.3* 0.35/0.5 = 0.21 21% probability of P(medium|C0)

F) P(Shirt Size=Medium | Class=C1) = P(C1 | Medium) *P(Medium)/P(C0) P(C1 | Medium) = 4/10 = 0.4, P(C1) = 10/20 = 0.5, P(Medium) = 7/20 = 0.35 P(Medium | C1) = 0.4* 0.35/0.5 = 0.28 28% probability of getting P(Medium|C1)

b) Use Naive Bayes Classifier to find the class of P(Gender=F | Car Type=Family| Shirt Size=Medium)

P(P(C1/(F/Family/Medium)) = 0.5 *0.6* 0.3 * 0.4 = 0.036

```
0.036% of P(F|Family|Medium)
```

NOTE: Julio helped me going through this exercise through Zoom.

Task 2 - Text Classification

Data Set: spam.csv located at https://www.kaggle.com/uciml/sms-spam-collection-dataset/version/1
Note: you might want to use `encoding of latin-1` when loading this file (https://www.kaggle.com/benvozza/spam-classification)

Objective: to classify SMS message as spam or not spam (ham).

From the given data set, use Naïve Bayes to classify the SMS message. The framework for text classification is briefly summarized here:

- Transformation of your dataset(change to lower case, remove numbers, remove punctuation, stop words, white space, word stemming, etc.)
- Document-Term-Matrix creation – matrix of word counts for each individual document in the matrix (e.g. documents as rows, words as columns or vice versa)
- Text Analysis (e.g. word counts, visualizations using wordclouds)

Helpful links:
https://machinelearningmastery.com/clean-text-machine-learning-python/
http://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization
https://machinelearningmastery.com/prepare-text-data-machine-learning

Analysis Questions:

- What is the accuracy of the model? Report your finding with corresponding tables/graphs.
- Print the 5 most frequent words in each class, and their posterior probability generated by the model.
- How would you improve the model performance?
- If the data set is bigger, do you think the accuracy increases? Discuss.

Deliverables:

Upload your work from task 1 and your notebook from task 2

Important: Make sure your provide complete and thorough explanations for all of your analysis. You need to defend your thought processes and reasoning.

In [103…
```
# loading required libraries
import pandas as pd
import seaborn as sns
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import confusion_matrix , accuracy_score , classification_report
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from PIL import Image

import matplotlib.pyplot as plt
% matplotlib inline
```

```
UsageError: Line magic function `%` not found.
```

```
Downloading stop-words files from nltk package
```

In [104…
```
# nltk downloading stopwords files
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Ahmad\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Ahmad\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[104… True

In [105…
```
spam = pd.read_csv('spam.csv', encoding='latin-1')
spam.head(10)
```

Out[105…

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN
5	spam	FreeMsg Hey there darling it's been 3 week's n...	NaN	NaN	NaN
6	ham	Even my brother is not like to speak with me. ...	NaN	NaN	NaN
7	ham	As per your request 'Melle Melle (Oru Minnamin...	NaN	NaN	NaN
8	spam	WINNER!! As a valued network customer you have...	NaN	NaN	NaN
9	spam	Had your mobile 11 months or more? U R entitle...	NaN	NaN	NaN

In [106…
```
spam.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

```
Dataset has 5 columns. We only need V1 and V2 all other columns are null and not useful. So, I dropped
them and renamed V1 as Class, and V2 as text
```

In [107…
```
#drop unused columns
spam = spam.drop(["Unnamed: 2", "Unnamed: 3", "Unnamed: 4"], axis=1)
# rename columns
spam = spam.rename(columns={"v1":"class", "v2":"text"})
spam.head(10)
```

Out[107…

	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...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...

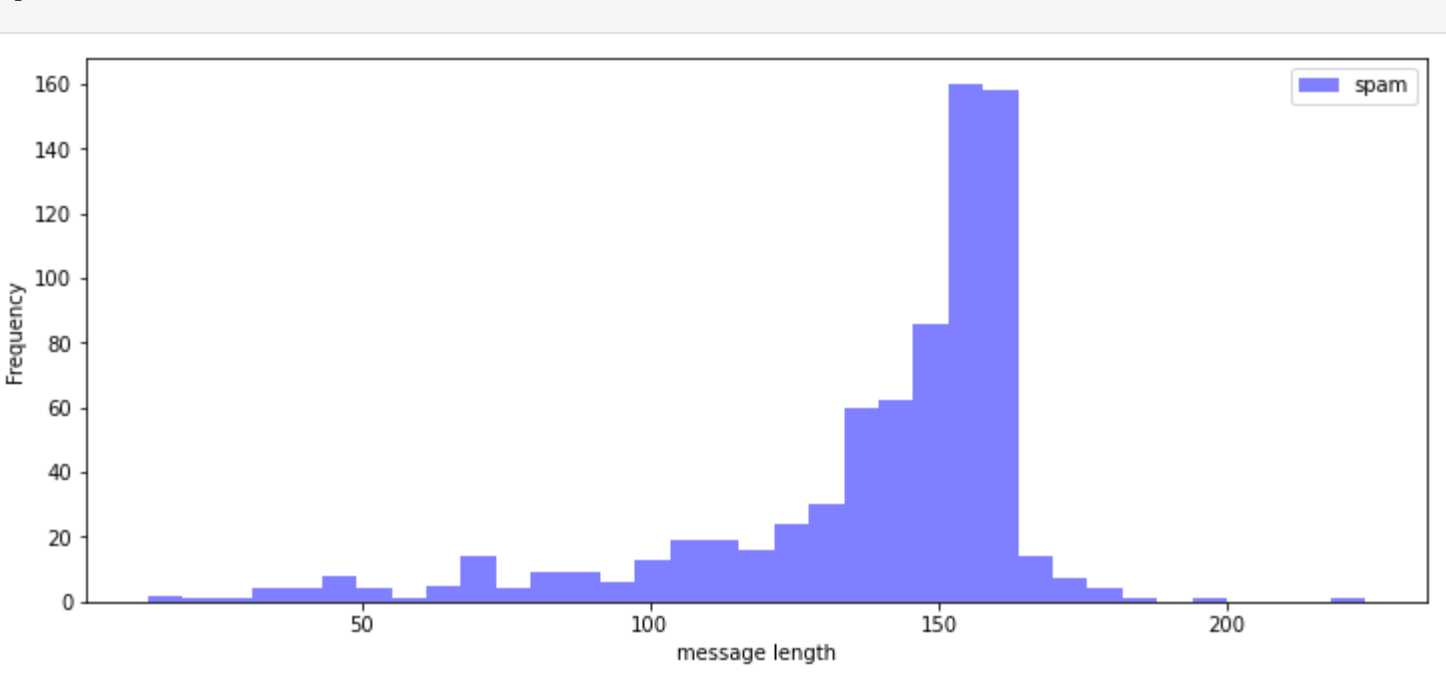
In [108…
```
# column that store text lentgh
len_text=[]
for i in spam['text']:
    len_text.append(len(i))
spam['text_length']=len_text
```

In [109…
```
spam.head()
```

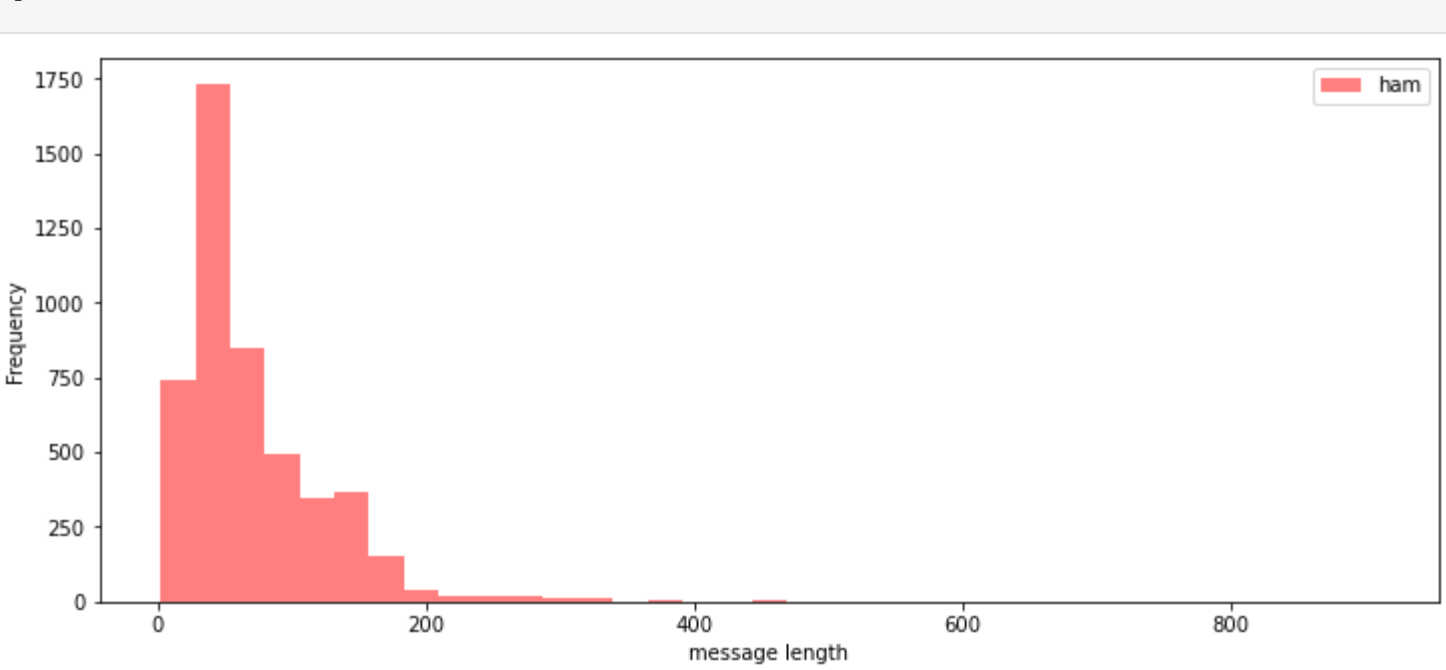
Out[109…

	class	text	text_length
0	ham	Go until jurong point, crazy.. Available only ...	111
1	ham	Ok lar... Joking wif u oni...	29
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	155
3	ham	U dun say so early hor... U c already then say...	49
4	ham	Nah I don't think he goes to usf, he lives aro...	61

In [110…
```
plt.figure(figsize=(12,5))
spam[spam['class']=='spam']['text_length'].plot(bins=35,kind='hist',color='blue',label='spam',alpha=0.5)
plt.legend()
plt.xlabel('message length')
plt.show()
```



In [111…
```
plt.figure(figsize=(12,5))
spam[spam['class']=='ham']['text_length'].plot(bins=35,kind='hist',color='red',label='ham',alpha=0.5)
plt.legend()
plt.xlabel('message length')
plt.show()
```




```
We can see that most of spam messages' length is between 100 and 150 while ham is between 40 and 70.
```

```
The function below is to remove stop-words, punctuations, and cnvert upper-case t lower-case.
```

In [112…
```
def process(sentence):
    text = re.sub('[^\w]+',' ',sentence).split()
    words = [x.lower() for x in text if x not in stopwords.words('english')]

    lemma = WordNetLemmatizer()
    word = [lemma.lemmatize(word,'v') for word in words ]
    word = ' '.join(word)
    return word
```

In [113…
```
spam['text'] = spam['text'].apply(process)
```

In [114…
```
spam.head()
```

Out[114…

	class	text	text_length
0	ham	go jurong point crazy available bugis n great ...	111
1	ham	ok lar joke wif u oni	29
2	spam	free entry 2 wkly comp win fa cup final tkts 2...	155
3	ham	u dun say early hor u c already say	49
4	ham	nah i think go usf live around though	61

```
The data now is clean. Stopwords, punctuations removed, and upper-cases converted to lower-cases.
```

In [115…
```
cv = CountVectorizer()
cv.fit(spam['text'])
print(cv.vocabulary_)
vector = cv.transform(spam['text'])
print(vector.shape)
print(type(vector))
print(vector.toarray())
```

'ig11': 3632, 'oja': 4934, '08712402578': 118, 'thesmszone': 6757, 'anonymous': 1085, 'mask': 4374, 'abuse': 8
69, 'woodland': 7479, 'avenue': 1268, 'parish': 5109, 'magazine': 4315, 'billy': 1463, 'awww': 1279, 'useless':
 7131, 'loo': 4205, 'ed': 2589, 'shelf': 6034, 'glands': 3179, 'previous': 5425, 'bcaz': 1372, 'stu': 6458, 'tr
uble': 6955, 'evone': 2753, 'hat': 3379, 'mokka': 4561, 'view': 7198, 'gays': 3123, 'dual': 2538, 'hostile': 35
33, 'haircut': 3332, 'breezy': 1602, '09061744553': 232, 'polyh': 5320, '1apple': 375, '1tulsi': 391, 'leaf': 4
085, '1lemon': 382, '1cup': 377, 'bone': 1526, 'problms': 5453, 'litres': 4168, 'watr': 7308, 'diseases': 2423,
 'snd': 6212, 'lavender': 4072, 'manky': 4351, 'scouse': 5923, 'steve': 6404, 'inmind': 3713, 'recreation': 564
9, 'epsilon': 2690, 'mesages': 4455, 'judgemental': 3898, 'fridays': 3037, 'waheeda': 7264, 'bot': 1545, 'event
ually': 2733, 'tolerance': 6866, '0789xxxxxxx': 47, 'hellogorgeous': 3425, 'lst': 4249, 'nitw': 4802, 'texd': 6
705, 'hopeu': 3521, '4ward': 603, 'jaz': 3838, '09058091870': 202, 'exorcism': 2778, 'emily': 2646, 'evry': 275
6, 'emotion': 2647, 'dsn': 2536, 'prayrs': 5391, 'othrwise': 5028, 'ujhhhhhhh': 7028, 'sandiago': 5875, 'parant
ella': 5104, 'sweater': 6574, 'mango': 4349, 'massive': 4376, 'landmark': 4043, 'bob': 1517, 'barry': 1346, '83
738': 760, 'absolutly': 864, 'consent': 2054, 'tonexs': 6878, 'clubzed': 1946, 'polo': 5316, '373': 512, 'w1j':
 7248, '6hl': 681, 'mathews': 4385, 'tait': 6614, 'edwards': 2600, 'anderson': 1066, 'haunt': 3383, 'promote':
 5473, 'crowd': 2152, '8000930705': 723, 'snowboard': 6220, 'christmassy': 1898, 'recpt': 5648, 'baaaaaaaabe':
 1294, 'shola': 6061, 'academic': 872, 'sagamu': 5857, 'lautech': 4071, 'vital': 7221, 'education': 2596, 'zeal
and': 7639, 'qet': 5536, 'browser': 1632, 'subscribers': 6483, 'conversations': 2071, 'overemphasise': 5055, 'h
eadset': 3400, 'adp': 925, 'internal': 3743, 'extract': 2800, 'immed': 3654, 'skint': 6152, 'bevies': 1445, 'wa
z': 7315, 'othrs': 5027, 'watchng': 7303, 'comfey': 1995, 'least5times': 4090, 'wudn': 7531, 'coast': 1958, 'fr
equently': 3031, 'cupboard': 2186, 'route': 5809, '2mro': 459, 'floppy': 2949, 'snappy': 6210, 'grasp': 3261,
 'flavour': 2936, 'laready': 4054, 'deny': 2320, 'dom': 2475, 'ffffuuuuuuu': 2881, 'julianaland': 3903, 'oblivi
ous': 4901, 'dehydrate': 2298, 'mapquest': 4355, 'dogwood': 2465, 'archive': 1154, '08719839835': 179, 'mgs': 4
463, '89123': 804, 'lengths': 4105, 'behalf': 1413, 'stun': 6469, 'visa': 7216, 'gucci': 3299, 'sozi': 6288, 'c
uldnt': 2180, 'talkbut': 6623, 'wannatell': 7285, 'wenwecan': 7359, 'smsing': 6203, 'efficient': 2606, 'explici
t': 2790, 'secs': 5948, '02073162414': 12, '15pm': 362, 'erutupalam': 2705, 'thandiyachu': 6721, 'invention': 3
753, 'lyric': 4278, 'nevr': 4763, 'unrecognized': 7086, 'somone': 6254, 'definitly': 2294, 'undrstnd': 7058, 'g
er': 3151, 'toking': 6863, 'syd': 6591, 'khelate': 3963, 'kintu': 3979, 'opponenter': 4987, 'dhorte': 2360, 'la
ge': 4032, 'spar': 6296, 'looovvve': 4211, 'warwick': 7294, 'tmw': 6844, 'havn': 3392, 'grandma': 3256, 'parad
e': 5100, 'proze': 5493, 'norcorp': 4831, '7cfca1a': 713, 'grumble': 3291, 'linear': 4147, 'algebra': 1007, 'de
corate': 2280, '946': 824, 'roomate': 5801, 'graduate': 3249, 'adjustable': 918, 'cooperative': 2080, 'nottingh
am': 4852, '63miles': 658, '40mph': 547, 'clearly': 1925, 'mornin': 4588, 'thanku': 6726, '89938': 808, 'strin
g': 6449, '50ea': 612, 'otbox': 5023, '731': 697, 'la1': 4020, '7ws': 719, 'beside': 1434, 'brisk': 1614, 'sexi
est': 6000, 'dirtiest': 2408, '89070': 801, 'tellmiss': 6676, 'contribute': 2068, 'greatly': 3272, 'urgh': 711
7, 'coach': 1957, 'duvet': 2559, 'predictive': 5396, 'w8in': 7254, '4utxt': 601, '24th': 421, 'beverage': 1444,
 'pist': 5247, 'surrender': 6554, 'symptoms': 6595, 'rdy': 5605, 'backwards': 1306, 'abstract': 865, 'africa':
 949, 'avin': 1269, 'chit': 1884, 'logon': 4192, '8883': 797, '4217': 552, 'w1a': 7247, '6zf': 691, '118p': 31
9, 'quiteamuzing': 5553, 'thatåôscool': 6736, '1000call': 300, '09071512432': 276, '300603t': 486, 'callcost150
ppmmobilesvary': 1712, 'mnths': 4538, 'engagement': 2661, 'fixd': 2926, 'njan': 4804, 'vilikkam': 7202, 'sudn':
 6499, 'maths': 4386, 'chapter': 1819, 'chop': 1891, 'noooooooo': 4828, '08718727870150ppm': 168, 'firsg': 291
8, 'split': 6333, 'wasnt': 7297, 'heat': 3411, 'sumfing': 6518, '86888': 779, 'subscribe6gbp': 6481, '3hrs': 52
3, 'txtstop': 7010, 'hiphop': 3464, 'oxygen': 5066, 'resort': 5729, 'roller': 5796, 'recorder': 5645, 'cannam
e': 1737, 'capital': 1745, 'australia': 1254, 'mquiz': 4610, 'showr': 6080, 'upon': 7106, 'ceiling': 1791, 'pre
snts': 5414, 'bcz': 1381, 'mis': 4504, 'jeevithathile': 3846, 'irulinae': 3776, 'neekunna': 4740, 'prakasaman
u': 5382, 'sneham': 6213, 'prakasam': 5381, 'ennal': 2670, 'mns': 4536, '7634': 704, '7684': 705, 'firmware': 2
917, 'vijaykanth': 7200, 'anythiing': 1105, 'clubmoby': 1944, '08717509990': 152, 'keypad': 3959, 'btwn': 1643,
 'decades': 2271, 'goverment': 3239, 'loyal': 4244, 'customers': 2200, '09066380611': 271, 'spice': 6317, 'pras
anth': 5386, 'ettans': 2721, '08718738002': 172, '48922': 578, 'appy': 1144, 'fizz': 2928, 'contain': 2062, 'ge
nus': 3147, 'robinson': 5787, 'soz': 6287, 'imat': 3649, 'mums': 4648, 'sometext': 6244, '07099833605': 27, '92
80114': 820, 'chloe': 1886, 'wewa': 7372, '130': 336, 'iriver': 3772, '255': 425, '128': 331, 'bw': 1685, 'surl
y': 6551, '07808726822': 42, '9758': 827, 'dokey': 2470, 'mmmmmmm': 4534, 'snuggle': 6222, 'whisper': 7395, 'er
ror': 2700, 'healthy': 3404, '2bold': 438, 'scrap': 5924, 'barrel': 1345, 'misfits': 4507, 'sleepin': 6165, 'cl
earer': 1924, 'peach': 5155, 'tasts': 6638, 'rayman': 5598, 'golf': 3209, 'activ8': 902, 'termsapply': 6693, 'm
inor': 4495, 'shindig': 6046, 'phonebook': 5217, 'ash': 1194, 'shijutta': 6043, 'offense': 4916, 'dvg': 2561,
 'vinobanagar': 7206, 'ovulate': 5059, '3wks': 535, 'woah': 7468, 'orh': 5013, 'secrets': 5947, 'n8': 4678, 'ax
el': 1281, 'akon': 993, 'cashbin': 1772, 'canteen': 1741, 'stressfull': 6444, 'president': 5412, '140': 340, 'ì
ä': 7658, '180': 369, 'pleasured': 5272, 'assistance': 1210, 'whens': 7385, '1172': 318, 'watever': 7305, 'lonl
ines': 4204, 'lotz': 4224, 'memories': 4442, 'gailxx': 3095, 'complacent': 2013, 'denis': 2315, 'mina': 4483,
 'miwa': 4523, 'woulda': 7505, '09066649731from': 274, 'opposite': 4990, 'heavily': 3414, 'dolls': 2474, 'patri
ck': 5135, 'swayze': 6572, 'santa': 5879, '09077818151': 279, 'calls1': 1720, '50ppm': 619, '30s': 493, 'santac
alling': 5880, 'miserable': 4506, 'quarter': 5542, 'limp': 4144, 'aa': 842, '078498': 45, '08719180219': 174,
 'oga': 4928, 'poorly': 5332, 'punishment': 5514, 'brb': 1589, 'predicte': 5394, 'situations': 6140, 'smaller':
 6186, 'capacity': 1744, 'videos': 7195, 'shsex': 6086, 'netun': 4756, 'fgkslpopw': 2883, 'fgkslpo': 2882, '087
1277810710p': 127, 'defer': 2290, 'admission': 922, 'checkmate': 1840, 'chess': 1853, 'persian': 5186, 'phras
e': 5223, 'shah': 6013, 'maat': 4294, 'than': 6720, 'photoshop': 5221, 'manageable': 4340, '08715203652': 139,
 '42810': 554, 'shoppin': 6064, 'increase': 3684, 'north': 4837, 'carolina': 1762, 'texas': 6704, 'gre': 3269,
 'bomb': 1525, 'powerful': 5366, 'weapon': 7322, 'lovly': 4240, 'msgrcvd': 4619, 'customercare': 2199, '0871872
6270': 160, 'clas': 1915, 'loooooool': 4210, 'couch': 2098, 'wasn': 7296, 'celebration': 1794, 'swashbuckling':
 6568, 'terror': 6697, 'cruel': 2156, 'decent': 2273, 'joker': 3878, 'dip': 2401, 'gek1510': 3136, 'nuther': 48
79, 'lyricalladie': 4279, '910': 818, 'hmmross': 3482, 'happiest': 3361, 'differences': 2382, 'antibiotic': 109
7, 'abdomen': 849, 'gynae': 3320, 'infections': 3700, '6times': 689, 'chastity': 1828, 'device': 2352, 'beat':
 1387, 'gut': 3314, 'tallahassee': 6626, 'ou': 5030, 'taka': 6616, 'pobox202': 5293, 'nr31': 4862, '7zs': 720,
 '450pw': 569, 'ritten': 5781, 'fold': 2962, 'warner': 7292, '83118': 751, 'colin': 1976, 'farrell': 2840, 'swa
t': 6569, 'kiosk': 3980, 'mre': 4612, 'solihull': 6234, 'nhs': 4774, '2b': 437, 'terminate': 6692, 'inconvenien
ce': 3681, 'dentists': 2319, 'henry': 3436, 'yards': 7572, 'bergkamp': 1433, 'margin': 4361, '78': 708, 'snot':
 6217, 'unintentional': 7070, 'nonetheless': 4824, 'hooch': 3513, 'toaday': 6849, 'splat': 6331, 'graze': 3268,
 'knees': 3990, 'hearin': 3407, 'yah': 7564, 'torture': 6901, 'sisters': 6133, 'sexychat': 6004, 'lips': 4159,
 'congratulation': 2048, 'court': 2113, 'chapel': 1818, 'frontierville': 3059, 'mountain': 4601, 'deer': 2286,
 'varma': 7159, 'fart': 2841, 'ortxt': 5018, 'advisors': 936, '402': 542, 'woot': 7482, 'ahhhh': 975, 'experien
cehttp': 2784, 'vouch4me': 7236, 'etlp': 2720, 'kaila': 3924, '09058094507': 205, 'donate': 2478, 'unicef': 706
8, 'asian': 1198, 'tsunami': 6972, 'disaster': 2414, 'fund': 3077, '864233': 777, 'cme': 1950, 'hos': 3528, 'co
llapse': 1978, 'cumming': 2184, 'jade': 3819, 'paul': 5138, 'barmed': 1343, 'thinkthis': 6769, 'dangerous': 223
2, 'goldviking': 3208, '762': 703, 'coulda': 2102, 'phony': 5218, 'okday': 4937, 'hmph': 3483, 'baller': 1321,
 'punto': 5516, 'ayo': 1285, '125': 328, 'freeentry': 3021, 'xt': 7548, 'toyota': 6915, 'camry': 1728, 'olayiwo
la': 4947, 'mileage': 4475, 'clover': 1940, 'achan': 890, 'amma': 1045, 'rencontre': 5698, 'mountains': 4602,
 '08714712412': 135, 'nìâte': 4891, 'puppy': 5517, 'noise': 4812, 'sarasota': 5885, 'meg': 4428, '08715203685':
 142, '4xx26': 605, '13': 335, 'deepest': 2285, 'darkest': 2237, '09094646631': 283, 'inconvenient': 3682, 'ads
ense': 929, 'dudette': 2545, 'perumbavoor': 5196, 'stage': 6369, 'cherish': 1851, 'clarify': 1913, 'preponed':
 5406, 'natalie': 4707, '165': 364, 'natalie2k9': 4708, 'younger': 7611, 'wallpaper': 7277, '08701213186': 82,
 'liver': 4171, 'opener': 4981, 'loneliness': 4200, 'skyving': 6159, 'onwords': 4974, 'mtnl': 4631, 'mumbai': 4
645, '83039': 748, '62735': 656, '450': 567, 'accommodationvouchers': 883, 'mustprovide': 4664, '15541': 361,
 'rajitha': 5574, 'ranju': 5588, 'tscs08714740323': 6970, '1winawk': 394, '50perweeksub': 615, 'slp': 6182, 'mu
ah': 4633, '09066361921': 261, 'disagreeable': 2411, 'afterwards': 956, 'uawake': 7018, 'feellikw': 2868, 'just
found': 3914, 'aletter': 1003, 'thatmum': 6733, 'gotmarried': 3235, '4thnov': 598, 'ourbacks': 5033, 'fuckinnic
e': 3066, 'rearrange': 5623, 'dormitory': 2489, 'astronomer': 1219, 'starer': 6378, 'election': 2626, 'recoun
t': 5646, 'hitler': 3469, 'eleven': 2630, 'worm': 7494, 'dysentry': 2563, 'andre': 1067, 'virgil': 7212, 'gokil
a': 3204, 'shanil': 6023, 'exchange': 2767, 'uncut': 7053, 'diamond': 2364, 'dino': 2399, 'kotees': 4005, 'pant
her': 5092, 'sugababes': 6503, 'zebra': 7640, 'animation': 1074, 'badass': 1308, 'hoody': 3514, 'resent': 5720,
 'query': 5545, 'customersqueries': 2201, 'netvision': 4757, 'hassle': 3378, 'andres': 1068, 'haughaighgtujhygu
j': 3381, 'fassyole': 2842, 'blacko': 1482, 'londn': 4198, '4eva': 583, 'responsibilities': 5735, '0871520527
3': 144, 'vco': 7166, 'humanities': 3574, 'reassurance': 5626, 'aslamalaikkum': 1203, 'tohar': 6859, 'beeen': 1
401, 'muht': 4640, 'albi': 997, 'mufti': 4638, 'mahfuuz': 4323, '078': 38, 'enufcredeit': 2683, 'tocall': 6852,
 'ileave': 3641, 'okors': 4943, 'ibored': 3606, 'zero': 7642, 'goigng': 3201, 'perfume': 5176, 'sday': 5934, 'g
rocers': 3283, 'pubs': 5508, 'frankie': 3011, 'bennys': 1432, 'diapers': 2366, 'unlike': 7081, 'patients': 513
4, 'turkeys': 6986, 'helens': 3421, 'princes': 5433, 'unintentionally': 7071, 'garden': 3110, 'bulbs': 1656, 's
eed': 5953, 'scotsman': 5922, 'go2': 3193, 'notxt': 4853, 'wenever': 7358, 'stability': 6365, 'tranquility': 69
25, 'vibrant': 7185, 'colourful': 1986, 'bawl': 1364, 'failure': 2820, 'velusamy': 7175, 'facilities': 2811, 'k
arnan': 3938, 'bluray': 1512, 'salt': 5868, 'wound': 7507, 'geoenvironmental': 3148, 'implications': 3660, 'fuu
uuck': 3087, 'salmon': 5866, 'wrkin': 7516, 'ree': 5655, 'compensation': 2011, 'awkward': 1277, 'splash': 6329,
 'leg': 4096, 'musta': 4662, 'overdo': 5053, 'telediscount': 6671, 'gastroenteritis': 3116, 'reduce': 5654, 'il
lness': 3643, 'foned': 2970, 'chuck': 1899, 'port': 5338, 'juswoke': 3916, 'boatin': 1516, 'spinout': 6322, '08
715203656': 140, '42049': 551, 'uworld': 7144, 'qbank': 5535, 'assessment': 1208, 'someonone': 6241, '090640153
07': 244, 'tke': 6832, 'temales': 6680, 'finishd': 2911, '09061213237': 220, '177': 367, 'm227xy': 4282, 'dul
l': 2548, 'anyones': 1103, 'treadmill': 6937, 'craigslist': 2126, 'absolutely': 863, 'swan': 6565, 'lamp': 403
7, 'foward': 3003, '09061790126': 236, 'misundrstud': 4519, '2u2': 474, 'genes': 3140, 'com1win150ppmx3age16sub
scription': 1989, 'reapply': 5622, 'treatin': 6940, 'treacle': 6936, 'mumhas': 4646, 'beendropping': 1403, 'the
place': 6751, 'adress': 926, 'favorite': 2854, 'oyster': 5068, 'sashimi': 5891, 'rumble': 5832, 'marandratha':
 4357, 'correctly': 2089, 'alaikkum': 995, 'pride': 5429, 'respectful': 5731, 'heaven': 3413, 'pisces': 5245,
 'aquarius': 1149, '2yrs': 481, 'steyn': 6406, 'wicket': 7408, 'sterm': 6403, 'resolve': 5728, 'hannaford': 335
4, 'wheat': 7379, 'chex': 1855, 'grownup': 3289, 'jerk': 3853, 'stink': 6413, 'subsequent': 6488, 'upcharge': 7
098, 'guai': 3297, 'bottle': 1548, 'amuse': 1058, 'astrology': 1218, 'slack': 6161, 'mentor': 4449, 'percent':
 5170, '09095350301': 285, 'erotic': 2698, 'ecstacy': 2587, 'dept': 2328, '08717507382': 151, 'coincidence': 19
72, 'sane': 5876, '151': 359, 'pause': 5139, 'gr8prizes': 3246, '8800': 790, 'psp': 5498, 'spacebucks': 6291,
 '083': 69, '6089': 648, 'squeeze': 6357, 'dread': 2513, 'thou': 6783, 'suggestion': 6508, 'forgt': 2990, 'ajit
h': 990, 'ooooooh': 4978, 'yoville': 7622, 'mega': 4429, 'asda': 1193, 'officer': 4919, 'bffs': 1449, 'carly':
 1761, 'seperated': 5982, 'ó_': 7660, 'ud': 7021, 'brolly': 1622, 'franxx': 3012, 'prometazine': 5469, 'syrup':
 6598, '5mls': 636, 'shu': 6087, 'singapore': 6124, 'victoria': 7189, 'island': 3781, 'pocay': 5301, 'wocay': 7
469, '2morrowxxxx': 458, 'broth': 1624, 'ramen': 5581, 'fowler': 3004, '09058094583': 207, 'attention': 1235,
 'tix': 6830, 'fne': 2960, 'youdoing': 7608, 'worc': 7485, 'foregate': 2982, 'shrub': 6085, 'get4an18th': 3154,
 '32000': 500, 'legitimat': 4098, 'efreefone': 2607, 'pendent': 5164, 'toilet': 6860, 'hu': 3563, 'navigate': 4
719, 'guidance': 3306, 'chick': 1861, 'boob': 1530, 'sparkle': 6299, '45': 566, '0121': 5, '2025050': 406, 'sho
rtbreaks': 6069, 'org': 5010, 'gyno': 3321, 'gamb': 3101, 'treasure': 6938, '820554ad0a1705572711': 742, 'trueå
ác': 6959, 'ringtoneåå': 5775, '09050000332': 186, 'positive': 5346, 'negative': 4742, 'hmmmm': 3481, 'comman
d': 1999, 'stressful': 6443, 'holby': 3492, '09064017295': 245, 'li': 4119, 'lecturer': 4094, 'yeovil': 7587,
 'motor': 4599, 'max': 4394, 'rhode': 5760, 'bong': 1527, 'ofcourse': 4912, '2p': 466, '08448350055': 71, 'plan
ettalkinstant': 5259, 'marvel': 4371, 'ultimate': 7032, 'spider': 6318, '83338': 754, '8ball': 810, 'tamilnad
u': 6628, '07808247860': 41, '08719899229': 181, '40411': 544, 'identification': 3621, 'boundaries': 1551, 'end
less': 2654, 'reassure': 5627, 'young': 7610, 'referin': 5658, 'saibaba': 5858, 'colany': 1974, 'chic': 1860,
 'declare': 2279, '49557': 580, 'disappointment': 2413, 'irritation': 3775, 'tantrum': 6631, 'compliment': 202
0, 'chief': 1864, 'gsex': 3293, '2667': 428, 'wc1n': 7316, '3xx': 536, '3mobile': 528, 'chatlines': 1831, 'incl
u': 3676, 'servs': 5992, 'l8er': 4015, 'bailiff': 1314, 'mouse': 4603, 'desk': 2338, 'childporn': 1869, 'jumper
s': 3906, 'belt': 1426, 'cribbs': 2144, 'spiritual': 6325, 'sudden': 6497, 'influx': 3702, 'kane': 3933, 'shu
d': 6088, 'pshew': 5497, 'units': 7074, 'accent': 874, '4years': 606, 'dental': 2317, 'nmde': 4805, 'dump': 255
0, 'heap': 3405, 'salesman': 5865, '087187272008': 165, 'now1': 4858, 'pity': 5248, 'suggestions': 6509}

```
(5572, 7670)
<class 'scipy.sparse.csr.csr_matrix'>
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

In [116…

```
vectorizer = TfidfVectorizer()
vectorizer.fit(spam['text'])
print(vectorizer.vocabulary_)
print(vectorizer.idf_)
# encode document
vector = vectorizer.transform(spam['text'])
# summarize encoded vector
print(vector.shape)
print(vector.toarray())
```

```
 'ig11': 3632, 'oja': 4934, '08712402578': 118, 'thesmszone': 6757, 'anonymous': 1085, 'mask': 4374, 'abuse': 8
69, 'woodland': 7479, 'avenue': 1268, 'parish': 5109, 'magazine': 4315, 'billy': 1463, 'awww': 1279, 'useless':
 7131, 'loo': 4205, 'ed': 2589, 'shelf': 6034, 'glands': 3179, 'previous': 5425, 'bcaz': 1372, 'stu': 6458, 'tr
uble': 6955, 'evone': 2753, 'hat': 3379, 'mokka': 4561, 'view': 7198, 'gays': 3123, 'dual': 2538, 'hostile': 35
33, 'haircut': 3332, 'breezy': 1602, '09061744553': 232, 'polyh': 5320, '1apple': 375, '1tulsi': 391, 'leaf': 4
085, '1lemon': 382, '1cup': 377, 'bone': 1526, 'problms': 5453, 'litres': 4168, 'watr': 7308, 'diseases': 2423,
 'snd': 6212, 'lavender': 4072, 'manky': 4351, 'scouse': 5923, 'steve': 6404, 'inmind': 3713, 'recreation': 564
9, 'epsilon': 2690, 'mesages': 4455, 'judgemental': 3898, 'fridays': 3037, 'waheeda': 7264, 'bot': 1545, 'event
ually': 2733, 'tolerance': 6866, '0789xxxxxxx': 47, 'hellogorgeous': 3425, 'lst': 4249, 'nitw': 4802, 'texd': 6
705, 'hopeu': 3521, '4ward': 603, 'jaz': 3838, '09058091870': 202, 'exorcism': 2778, 'emily': 2646, 'evry': 275
6, 'emotion': 2647, 'dsn': 2536, 'prayrs': 5391, 'othrwise': 5028, 'ujhhhhhhh': 7028, 'sandiago': 5875, 'parant
ella': 5104, 'sweater': 6574, 'mango': 4349, 'massive': 4376, 'landmark': 4043, 'bob': 1517, 'barry': 1346, '83
738': 760, 'absolutly': 864, 'consent': 2054, 'tonexs': 6878, 'clubzed': 1946, 'polo': 5316, '373': 512, 'w1j':
 7248, '6hl': 681, 'mathews': 4385, 'tait': 6614, 'edwards': 2600, 'anderson': 1066, 'haunt': 3383, 'promote':
 5473, 'crowd': 2152, '8000930705': 723, 'snowboard': 6220, 'christmassy': 1898, 'recpt': 5648, 'baaaaaaaabe':
 1294, 'shola': 6061, 'academic': 872, 'sagamu': 5857, 'lautech': 4071, 'vital': 7221, 'education': 2596, 'zeal
and': 7639, 'qet': 5536, 'browser': 1632, 'subscribers': 6483, 'conversations': 2071, 'overemphasise': 5055, 'h
eadset': 3400, 'adp': 925, 'internal': 3743, 'extract': 2800, 'immed': 3654, 'skint': 6152, 'bevies': 1445, 'wa
z': 7315, 'othrs': 5027, 'watchng': 7303, 'comfey': 1995, 'least5times': 4090, 'wudn': 7531, 'coast': 1958, 'fr
equently': 3031, 'cupboard': 2186, 'route': 5809, '2mro': 459, 'floppy': 2949, 'snappy': 6210, 'grasp': 3261,
 'flavour': 2936, 'laready': 4054, 'deny': 2320, 'dom': 2475, 'ffffuuuuuuu': 2881, 'julianaland': 3903, 'oblivi
ous': 4901, 'dehydrate': 2298, 'mapquest': 4355, 'dogwood': 2465, 'archive': 1154, '08719839835': 179, 'mgs': 4
463, '89123': 804, 'lengths': 4105, 'behalf': 1413, 'stun': 6469, 'visa': 7216, 'gucci': 3299, 'sozi': 6288, 'c
uldnt': 2180, 'talkbut': 6623, 'wannatell': 7285, 'wenwecan': 7359, 'smsing': 6203, 'efficient': 2606, 'explici
t': 2790, 'secs': 5948, '02073162414': 12, '15pm': 362, 'erutupalam': 2705, 'thandiyachu': 6721, 'invention': 3
753, 'lyric': 4278, 'nevr': 4763, 'unrecognized': 7086, 'somone': 6254, 'definitly': 2294, 'undrstnd': 7058, 'g
er': 3151, 'toking': 6863, 'syd': 6591, 'khelate': 3963, 'kintu': 3979, 'opponenter': 4987, 'dhorte': 2360, 'la
ge': 4032, 'spar': 6296, 'looovvve': 4211, 'warwick': 7294, 'tmw': 6844, 'havn': 3392, 'grandma': 3256, 'parad
e': 5100, 'proze': 5493, 'norcorp': 4831, '7cfca1a': 713, 'grumble': 3291, 'linear': 4147, 'algebra': 1007, 'de
corate': 2280, '946': 824, 'roomate': 5801, 'graduate': 3249, 'adjustable': 918, 'cooperative': 2080, 'nottingh
am': 4852, '63miles': 658, '40mph': 547, 'clearly': 1925, 'mornin': 4588, 'thanku': 6726, '89938': 808, 'strin
g': 6449, '50ea': 612, 'otbox': 5023, '731': 697, 'la1': 4020, '7ws': 719, 'beside': 1434, 'brisk': 1614, 'sexi
est': 6000, 'dirtiest': 2408, '89070': 801, 'tellmiss': 6676, 'contribute': 2068, 'greatly': 3272, 'urgh': 711
7, 'coach': 1957, 'duvet': 2559, 'predictive': 5396, 'w8in': 7254, '4utxt': 601, '24th': 421, 'beverage': 1444,
 'pist': 5247, 'surrender': 6554, 'symptoms': 6595, 'rdy': 5605, 'backwards': 1306, 'abstract': 865, 'africa':
 949, 'avin': 1269, 'chit': 1884, 'logon': 4192, '8883': 797, '4217': 552, 'w1a': 7247, '6zf': 691, '118p': 31
9, 'quiteamuzing': 5553, 'thatåõscool': 6736, '1000call': 300, '09071512432': 276, '300603t': 486, 'callcost150
ppmmobilesvary': 1712, 'mnths': 4538, 'engagement': 2661, 'fixd': 2926, 'njan': 4804, 'vilikkam': 7202, 'sudn':
 6499, 'maths': 4386, 'chapter': 1819, 'chop': 1891, 'noooooooo': 4828, '08718727870150ppm': 168, 'firsg': 291
8, 'split': 6333, 'wasnt': 7297, 'heat': 3411, 'sumfing': 6518, '86888': 779, 'subscribe6gbp': 6481, '3hrs': 52
3, 'txtstop': 7010, 'hiphop': 3464, 'oxygen': 5066, 'resort': 5729, 'roller': 5796, 'recorder': 5645, 'cannam
e': 1737, 'capital': 1745, 'australia': 1254, 'mquiz': 4610, 'showr': 6080, 'upon': 7106, 'ceiling': 1791, 'pre
snts': 5414, 'bcz': 1381, 'mis': 4504, 'jeevithathile': 3846, 'irulinae': 3776, 'neekunna': 4740, 'prakasaman
u': 5382, 'sneham': 6213, 'prakasam': 5381, 'ennal': 2670, 'mns': 4536, '7634': 704, '7684': 705, 'firmware': 2
917, 'vijaykanth': 7200, 'anythiing': 1105, 'clubmoby': 1944, '08717509990': 152, 'keypad': 3959, 'btwn': 1643,
 'decades': 2271, 'goverment': 3239, 'loyal': 4244, 'customers': 2200, '09066380611': 271, 'spice': 6317, 'pras
anth': 5386, 'ettans': 2721, '08718738002': 172, '48922': 578, 'appy': 1144, 'fizz': 2928, 'contain': 2062, 'ge
nus': 3147, 'robinson': 5787, 'soz': 6287, 'imat': 3649, 'mums': 4648, 'sometext': 6244, '07099833605': 27, '92
80114': 820, 'chloe': 1886, 'wewa': 7372, '130': 336, 'iriver': 3772, '255': 425, '128': 331, 'bw': 1685, 'surl
y': 6551, '07808726822': 42, '9758': 827, 'dokey': 2470, 'mmmmmmm': 4534, 'snuggle': 6222, 'whisper': 7395, 'er
ror': 2700, 'healthy': 3404, '2bold': 438, 'scrap': 5924, 'barrel': 1345, 'misfits': 4507, 'sleepin': 6165, 'cl
earer': 1924, 'peach': 5155, 'tasts': 6638, 'rayman': 5598, 'golf': 3209, 'activ8': 902, 'termsapply': 6693, 'm
inor': 4495, 'shindig': 6046, 'phonebook': 5217, 'ash': 1194, 'shijutta': 6043, 'offense': 4916, 'dvg': 2561,
 'vinobanagar': 7206, 'ovulate': 5059, '3wks': 535, 'woah': 7468, 'orh': 5013, 'secrets': 5947, 'n8': 4678, 'ax
el': 1281, 'akon': 993, 'cashbin': 1772, 'canteen': 1741, 'stressfull': 6444, 'president': 5412, '140': 340, 'i
ä': 7658, '180': 369, 'pleasured': 5272, 'assistance': 1210, 'whens': 7385, '1172': 318, 'watever': 7305, 'lonl
ines': 4204, 'lotz': 4224, 'memories': 4442, 'gailxx': 3095, 'complacent': 2013, 'denis': 2315, 'mina': 4483,
 'miwa': 4523, 'woulda': 7505, '09066649731from': 274, 'opposite': 4990, 'heavily': 3414, 'dolls': 2474, 'patri
ck': 5135, 'swayze': 6572, 'santa': 5879, '09077818151': 279, 'calls1': 1720, '50ppm': 619, '30s': 493, 'santac
alling': 5880, 'miserable': 4506, 'quarter': 5542, 'limp': 4144, 'aa': 842, '078498': 45, '08719180219': 174,
 'oga': 4928, 'poorly': 5332, 'punishment': 5514, 'brb': 1589, 'predicte': 5394, 'situations': 6140, 'smaller':
 6186, 'capacity': 1744, 'videos': 7195, 'shsex': 6086, 'netun': 4756, 'fgkslpopw': 2883, 'fgkslpo': 2882, '087
1277810710p': 127, 'defer': 2290, 'admission': 922, 'checkmate': 1840, 'chess': 1853, 'persian': 5186, 'phras
e': 5223, 'shah': 6013, 'maat': 4294, 'than': 6720, 'photoshop': 5221, 'manageable': 4340, '08715203652': 139,
 '42810': 554, 'shoppin': 6064, 'increase': 3684, 'north': 4837, 'carolina': 1762, 'texas': 6704, 'gre': 3269,
 'bomb': 1525, 'powerful': 5366, 'weapon': 7322, 'lovly': 4240, 'msgrcvd': 4619, 'customercare': 2199, '0871872
6270': 160, 'clas': 1915, 'loooooool': 4210, 'couch': 2098, 'wasn': 7296, 'celebration': 1794, 'swashbuckling':
 6568, 'terror': 6697, 'cruel': 2156, 'decent': 2273, 'joker': 3878, 'dip': 2401, 'gek1510': 3136, 'nuther': 48
79, 'lyricalladie': 4279, '910': 818, 'hmmross': 3482, 'happiest': 3361, 'differences': 2382, 'antibiotic': 109
7, 'abdomen': 849, 'gynae': 3320, 'infections': 3700, '6times': 689, 'chastity': 1828, 'device': 2352, 'beat':
 1387, 'gut': 3314, 'tallahassee': 6626, 'ou': 5030, 'taka': 6616, 'pobox202': 5293, 'nr31': 4862, '7zs': 720,
 '450pw': 569, 'ritten': 5781, 'fold': 2962, 'warner': 7292, '83118': 751, 'colin': 1976, 'farrell': 2840, 'swa
t': 6569, 'kiosk': 3980, 'mre': 4612, 'solihull': 6234, 'nhs': 4774, '2b': 437, 'terminate': 6692, 'inconvenien
ce': 3681, 'dentists': 2319, 'henry': 3436, 'yards': 7572, 'bergkamp': 1433, 'margin': 4361, '78': 708, 'snot':
 6217, 'unintentional': 7070, 'nonetheless': 4824, 'hooch': 3513, 'toaday': 6849, 'splat': 6331, 'graze': 3268,
 'knees': 3990, 'hearin': 3407, 'yah': 7564, 'torture': 6901, 'sisters': 6133, 'sexychat': 6004, 'lips': 4159,
 'congratulation': 2048, 'court': 2113, 'chapel': 1818, 'frontierville': 3059, 'mountain': 4601, 'deer': 2286,
 'varma': 7159, 'fart': 2841, 'ortxt': 5018, 'advisors': 936, '402': 542, 'woot': 7482, 'ahhhh': 975, 'experien
cehttp': 2784, 'vouch4me': 7236, 'etlp': 2720, 'kaila': 3924, '09058094507': 205, 'donate': 2478, 'unicef': 706
8, 'asian': 1198, 'tsunami': 6972, 'disaster': 2414, 'fund': 3077, '864233': 777, 'cme': 1950, 'hos': 3528, 'co
llapse': 1978, 'cumming': 2184, 'jade': 3819, 'paul': 5138, 'barmed': 1343, 'thinkthis': 6769, 'dangerous': 223
2, 'goldviking': 3208, '762': 703, 'coulda': 2102, 'phony': 5218, 'okday': 4937, 'hmph': 3483, 'baller': 1321,
 'punto': 5516, 'ayo': 1285, '125': 328, 'freeentry': 3021, 'xt': 7548, 'toyota': 6915, 'camry': 1728, 'olayiwo
la': 4947, 'mileage': 4475, 'clover': 1940, 'achan': 890, 'amma': 1045, 'rencontre': 5698, 'mountains': 4602,
 '08714712412': 135, 'nìâte': 4891, 'puppy': 5517, 'noise': 4812, 'sarasota': 5885, 'meg': 4428, '08715203685':
 142, '4xx26': 605, '13': 335, 'deepest': 2285, 'darkest': 2237, '09094646631': 283, 'inconvenient': 3682, 'ads
ense': 929, 'dudette': 2545, 'perumbavoor': 5196, 'stage': 6369, 'cherish': 1851, 'clarify': 1913, 'preponed':
 5406, 'natalie': 4707, '165': 364, 'natalie2k9': 4708, 'younger': 7611, 'wallpaper': 7277, '08701213186': 82,
 'liver': 4171, 'opener': 4981, 'loneliness': 4200, 'skyving': 6159, 'onwords': 4974, 'mtnl': 4631, 'mumbai': 4
645, '83039': 748, '62735': 656, '450': 567, 'accommodationvouchers': 883, 'mustprovide': 4664, '15541': 361,
 'rajitha': 5574, 'ranju': 5588, 'tscs08714740323': 6970, '1winawk': 394, '50perweeksub': 615, 'slp': 6182, 'mu
ah': 4633, '09066361921': 261, 'disagreeable': 2411, 'afterwards': 956, 'uawake': 7018, 'feellikw': 2868, 'just
found': 3914, 'aletter': 1003, 'thatmum': 6733, 'gotmarried': 3235, '4thnov': 598, 'ourbacks': 5033, 'fuckinnic
e': 3066, 'rearrange': 5623, 'dormitory': 2489, 'astronomer': 1219, 'starer': 6378, 'election': 2626, 'recoun
t': 5646, 'hitler': 3469, 'eleven': 2630, 'worm': 7494, 'dysentry': 2563, 'andre': 1067, 'virgil': 7212, 'gokil
a': 3204, 'shanil': 6023, 'exchange': 2767, 'uncut': 7053, 'diamond': 2364, 'dino': 2399, 'kotees': 4005, 'pant
her': 5092, 'sugababes': 6503, 'zebra': 7640, 'animation': 1074, 'badass': 1308, 'hoody': 3514, 'resent': 5720,
 'query': 5545, 'customersqueries': 2201, 'netvision': 4757, 'hassle': 3378, 'andres': 1068, 'haughaighgtujhygu
j': 3381, 'fassyole': 2842, 'blacko': 1482, 'londn': 4198, '4eva': 583, 'responsibilities': 5735, '0871520527
3': 144, 'vco': 7166, 'humanities': 3574, 'reassurance': 5626, 'aslamalaikkum': 1203, 'tohar': 6859, 'beeen': 1
401, 'muht': 4640, 'albi': 997, 'mufti': 4638, 'mahfuuz': 4323, '078': 38, 'enufcredeit': 2683, 'tocall': 6852,
 'ileave': 3641, 'okors': 4943, 'ibored': 3606, 'zero': 7642, 'goigng': 3201, 'perfume': 5176, 'sday': 5934, 'g
rocers': 3283, 'pubs': 5508, 'frankie': 3011, 'bennys': 1432, 'diapers': 2366, 'unlike': 7081, 'patients': 513
4, 'turkeys': 6986, 'helens': 3421, 'princes': 5433, 'unintentionally': 7071, 'garden': 3110, 'bulbs': 1656, 's
eed': 5953, 'scotsman': 5922, 'go2': 3193, 'notxt': 4853, 'wenever': 7358, 'stability': 6365, 'tranquility': 69
25, 'vibrant': 7185, 'colourful': 1986, 'bawl': 1364, 'failure': 2820, 'velusamy': 7175, 'facilities': 2811, 'k
arnan': 3938, 'bluray': 1512, 'salt': 5868, 'wound': 7507, 'geoenvironmental': 3148, 'implications': 3660, 'fuu
uuck': 3087, 'salmon': 5866, 'wrkin': 7516, 'ree': 5655, 'compensation': 2011, 'awkward': 1277, 'splash': 6329,
 'leg': 4096, 'musta': 4662, 'overdo': 5053, 'telediscount': 6671, 'gastroenteritis': 3116, 'reduce': 5654, 'il
lness': 3643, 'foned': 2970, 'chuck': 1899, 'port': 5338, 'juswoke': 3916, 'boatin': 1516, 'spinout': 6322, '08
715203656': 140, '42049': 551, 'uworld': 7144, 'qbank': 5535, 'assessment': 1208, 'someonone': 6241, '090640153
07': 244, 'tke': 6832, 'temales': 6680, 'finishd': 2911, '09061213237': 220, '177': 367, 'm227xy': 4282, 'dul
l': 2548, 'anyones': 1103, 'treadmill': 6937, 'craigslist': 2126, 'absolutely': 863, 'swan': 6565, 'lamp': 403
7, 'foward': 3003, '09061790126': 236, 'misundrstud': 4519, '2u2': 474, 'genes': 3140, 'com1win150ppmx3age16sub
scription': 1989, 'reapply': 5622, 'treatin': 6940, 'treacle': 6936, 'mumhas': 4646, 'beendropping': 1403, 'the
place': 6751, 'adress': 926, 'favorite': 2854, 'oyster': 5068, 'sashimi': 5891, 'rumble': 5832, 'marandratha':
 4357, 'correctly': 2089, 'alaikkum': 995, 'pride': 5429, 'respectful': 5731, 'heaven': 3413, 'pisces': 5245,
 'aquarius': 1149, '2yrs': 481, 'steyn': 6406, 'wicket': 7408, 'sterm': 6403, 'resolve': 5728, 'hannaford': 335
4, 'wheat': 7379, 'chex': 1855, 'grownup': 3289, 'jerk': 3853, 'stink': 6413, 'subsequent': 6488, 'upcharge': 7
098, 'guai': 3297, 'bottle': 1548, 'amuse': 1058, 'astrology': 1218, 'slack': 6161, 'mentor': 4449, 'percent':
 5170, '09095350301': 285, 'erotic': 2698, 'ecstacy': 2587, 'dept': 2328, '08717507382': 151, 'coincidence': 19
72, 'sane': 5876, '151': 359, 'pause': 5139, 'gr8prizes': 3246, '8800': 790, 'psp': 5498, 'spacebucks': 6291,
 '083': 69, '6089': 648, 'squeeze': 6357, 'dread': 2513, 'thou': 6783, 'suggestion': 6508, 'forgt': 2990, 'ajit
h': 990, 'oooooh': 4978, 'yoville': 7622, 'mega': 4429, 'asda': 1193, 'officer': 4919, 'bffs': 1449, 'carly':
 1761, 'seperated': 5982, 'ó_': 7660, 'ud': 7021, 'brolly': 1622, 'franxx': 3012, 'prometazine': 5469, 'syrup':
 6598, '5mls': 636, 'shu': 6087, 'singapore': 6124, 'victoria': 7189, 'island': 3781, 'pocay': 5301, 'wocay': 7
469, '2morrowxxxx': 458, 'broth': 1624, 'ramen': 5581, 'fowler': 3004, '09058094583': 207, 'attention': 1235,
 'tix': 6830, 'fne': 2960, 'youdoing': 7608, 'worc': 7485, 'foregate': 2982, 'shrub': 6085, 'get4an18th': 3154,
 '32000': 500, 'legitimat': 4098, 'efreefone': 2607, 'pendent': 5164, 'toilet': 6860, 'hu': 3563, 'navigate': 4
719, 'guidance': 3306, 'chick': 1861, 'boob': 1530, 'sparkle': 6299, '45': 566, '0121': 5, '2025050': 406, 'sho
rtbreaks': 6069, 'org': 5010, 'gyno': 3321, 'gamb': 3101, 'treasure': 6938, '820554ad0a1705572711': 742, 'trueå
ác': 6959, 'ringtoneåá': 5775, '09050000332': 186, 'positive': 5346, 'negative': 4742, 'hmmmm': 3481, 'comman
d': 1999, 'stressful': 6443, 'holby': 3492, '09064017295': 245, 'li': 4119, 'lecturer': 4094, 'yeovil': 7587,
 'motor': 4599, 'max': 4394, 'rhode': 5760, 'bong': 1527, 'ofcourse': 4912, '2p': 466, '08448350055': 71, 'plan
ettalkinstant': 5259, 'marvel': 4371, 'ultimate': 7032, 'spider': 6318, '83338': 754, '8ball': 810, 'tamilnad
u': 6628, '07808247860': 41, '08719899229': 181, '40411': 544, 'identification': 3621, 'boundaries': 1551, 'end
less': 2654, 'reassure': 5627, 'young': 7610, 'referin': 5658, 'saibaba': 5858, 'colany': 1974, 'chic': 1860,
 'declare': 2279, '49557': 580, 'disappointment': 2413, 'irritation': 3775, 'tantrum': 6631, 'compliment': 202
0, 'chief': 1864, 'gsex': 3293, '2667': 428, 'wcln': 7316, '3xx': 536, '3mobile': 528, 'chatlines': 1831, 'incl
u': 3676, 'servs': 5992, 'l8er': 4015, 'bailiff': 1314, 'mouse': 4603, 'desk': 2338, 'childporn': 1869, 'jumper
s': 3906, 'belt': 1426, 'cribbs': 2144, 'spiritual': 6325, 'sudden': 6497, 'influx': 3702, 'kane': 3933, 'shu
d': 6088, 'pshew': 5497, 'units': 7074, 'accent': 874, '4years': 606, 'dental': 2317, 'nmde': 4805, 'dump': 255
0, 'heap': 3405, 'salesman': 5865, '087187272008': 165, 'now1': 4858, 'pity': 5248, 'suggestions': 6509}
[7.22779351 6.29348428 8.93254161 ... 8.93254161 7.67977864 8.93254161]
(5572, 7670)
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]
```

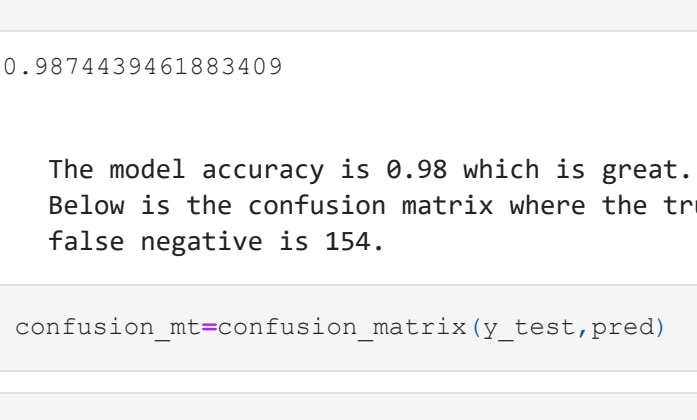

The Two code above are counting the words frequency in all documents (rows). they present same result, but I don't know how they work differently.

In [117…

```
text = spam.text[0]

# Create and generate a word cloud image:
wordcloud = WordCloud().generate(text)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```



The wordcloud above shows the most frequent word in the text.

Below is the modeling part.

we split the data into train and test wit ratio of 0.8 for train and 0.2 for test.
and we use Multinomial Naive Bayes model.

In [99]:

In [118…

```
# setting Target and features
y=spam['class']
x=spam['text']
```

In [119…

```
# splitting data
x_train,x_test,y_train,y_test=train_test_split(x,y,random_state=0,test_size=0.2)
```

In [120…

```
x_train=cv.fit_transform(x_train)
```

In [121…

```
nb=MultinomialNB()
nb.fit(x_train,y_train)
pred=nb.predict(cv.transform(x_test))
score=accuracy_score(y_test,pred)
score
```

Out[121… 0.9874439461883409

The model accuracy is 0.98 which is great.
Below is the confusion matrix where the true positive is 947, false positive is 2, true negative 12, and false negative is 154.

In [123…

```
confusion_mt=confusion_matrix(y_test,pred)
```

In [124…

```
confusion_mt
```

Out[124…

```
array([[947,   2],
       [ 12, 154]], dtype=int64)
```

Conclusion

To conclude, Naive Bayes is proved to be effecient in term of detection of spam messages. An accuracy of 98% is satisifying. However, if we can improve our model, we need to increase informative features.

References

https://www.kaggle.com/sachinsharma1123/spam-classifier-for-freshers-acc-98

https://www.kaggle.com/uciml/sms-spam-collection-dataset/tasks?taskId=1188

https://www.kaggle.com/aryantiwari123/nlp-sms-spam-detection-text-processing

https://medium.com/@cristhianboujon/how-to-list-the-most-common-words-from-text-corpus-using-scikit-learn-dad4d0cab41d

https://machinelearningmastery.com/clean-text-machine-learning-python/

https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn/

https://textminingonline.com/dive-into-nltk-part-iv-stemming-and-lemmatization

https://www.datacamp.com/community/tutorials/wordcloud-python

https://www.quora.com/Does-increasing-the-number-of-features-always-improve-a-naive-Bayes-classifier