In [18]:

```
import pandas as pd
import numpy as np
```

In [19]:

```
data = pd.read_csv(r"C:\Users\s323\Desktop\Gatherings\Data Science\NLP\spam.csv")
```

In [20]:

```
# First five elements in the data
data.head()
```

Out[20]:

	Category	Message
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...

In [21]:

```
# shape of data
data.shape
```

Out[21]:

```
(5572, 2)
```

In [22]:

```
# How many emails are considered as spam and not spam respectively
data.Category.value_counts()
```

Out[22]:

```
ham     4825
spam     747
Name: Category, dtype: int64
```

We can notice, it is a imbalanced dataset - but we will see how it goes

Now we will add one column in the dataset, and at place of spam we will mark as 1 and if not spam then 0

- It is good by any which of below methods

In [28]:

```
def get_spam_number(x):
    if x =="spam":
        return 1
    return 0
```

In [29]:

```
data["spam"] = data["Category"].apply(lambda x : 1 if x == "spam" else 0)
```

In [30]:

```
data.shape
```

Out[30]:

```
(5572, 3)
```

In [31]:

```
data.head()
```

Out[31]:

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

Train Test Split

In [38]:

```
X = data["Message"]
y = data["spam"]
```

In [40]:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

In [41]:

```
X_train.shape
```

Out[41]:

```
(4457,)
```

In [42]:

```
X_test.shape
```

Out[42]:

```
(1115,)
```

In [43]:

```
type(X_train)
```

Out[43]:

```
pandas.core.series.Series
```

In [44]:

```
X_train[:4]
```

Out[44]:

```
5166    Y she dun believe leh? I tot i told her it's t...
2135    If he started searching he will get job in few...
1503                               Don no da:)whats you plan?
3438                            Then what about further plan?
Name: Message, dtype: object
```

In [45]:

```
y_train[:4]
```

Out[45]:

```
5166    0
2135    0
1503    0
3438    0
Name: spam, dtype: int64
```

Create bag of words representation using CountVectorizer

In [49]:

```
from sklearn.feature_extraction.text import CountVectorizer
```

In [50]:

```
v = CountVectorizer()
```

In [51]:

```
X_train_cv = v.fit_transform(X_train.values)
X_train_cv
```

```
Out[51]: <4457x7771 sparse matrix of type '<class 'numpy.int64'>'
                 with 59207 stored elements in Compressed Sparse Row format>
```

```
In [52]: X_train_cv.toarray()[:2][0]
```

```
Out[52]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

```
In [53]: X_train_cv.shape
```

```
Out[53]: (4457, 7771)
```

```
In [55]: v.get_feature_names_out()[1771]
```

```
Out[55]: 'checked'
```

```
In [56]: v.vocabulary_
```

```
Out[56]: {'she': 6097,
          'dun': 2498,
          'believe': 1322,
          'leh': 4095,
          'tot': 7010,
          'told': 6973,
          'her': 3428,
          'it': 3772,
          'true': 7069,
          'already': 934,
          'thk': 6878,
          'muz': 4680,
          'us': 7234,
          'tog': 6967,
          'then': 6848,
          'if': 3619,
          'he': 3382,
          'started': 6473,
          'searching': 5990,
          'will': 7541,
          'get': 3134,
          'job': 3852,
          'in': 3655,
          'few': 2832,
          'days': 2198,
          'have': 3371,
          'great': 3251,
          'potential': 5365,
          'and': 970,
          'talent': 6720,
          'don': 2419,
          'no': 4815,
          'da': 2157,
          'whats': 7502,
          'you': 7730,
          'plan': 5262,
          'what': 7500,
          'about': 767,
          'further': 3065,
          'are': 1060,
          'subscribed': 6586,
          'to': 6958,
          'the': 6836,
          'best': 1340,
          'mobile': 4564,
          'content': 2001,
          'service': 6053,
          'uk': 7142,
          'for': 2952,
          'per': 5171,
          'ten': 6790,
          'until': 7202,
          'send': 6032,
          'stop': 6519,
          '83435': 674,
          'helpline': 3421,
          '08706091795': 88,
          'loosu': 4227,
          'go': 3175,
          'hospital': 3521,
          'de': 2202,
          'dont': 2423,
          'let': 4107,
          'careless': 1676,
          'boy': 1480,
          'loved': 4250,
          'gal': 3078,
          'propsd': 5495,
          'bt': 1549,
          'didnt': 2314,
          'mind': 4503,
          'gv': 3298,
          'lv': 4291,
          'lttrs': 4275,
          'frnds': 3024,
          'threw': 6896,
          'thm': 6880,
          'again': 870,
          'decided': 2217,
          'aproach': 1050,
          'dt': 2486,
          'time': 6928,
          'truck': 7068,
          'was': 7409,
          'speeding': 6399,
          'towards': 7019,
          'wn': 7589,
          'hit': 3458,
          'girl': 3155,
          'ran': 5597,
          'like': 4133,
          'hell': 3410,
          'saved': 5944,
          'asked': 1109,
          'hw': 3588,
          'cn': 1885,
          'run': 5877,
          'so': 6308,
          'fast': 2792,
          'replied': 5734,
          'boost': 1443,
          'is': 3762,
          'secret': 5997,
          'of': 4913,
          'my': 4683,
          'energy': 2603,
          'instantly': 3713,
          'shouted': 6144,
          'our': 5032,
          'thy': 6914,
          'lived': 4174,
          'happily': 3349,
          '2gthr': 392,
          'drinking': 2462,
          'evrydy': 2699,
          'moral': 4603,
          'story': 6530,
          'hv': 3586,
          'free': 2994,
          'msgs': 4643,
          'gud': 3281,
          'ni8': 4785,
          'do': 2383,
          'we': 7435,
          'any': 1004,
          'spare': 6382,
          'power': 5374,
          'supplies': 6640,
          'gonna': 3198,
          'ring': 5809,
          'this': 6877,
          'weekend': 7462,
          'or': 4997,
          'wot': 7630,
          'thanks': 6827,
          'loving': 4259,
          'me': 4425,
          'rock': 5828,
          'nt': 4869,
          'only': 4969,
          'driving': 2468,
          'even': 2673,
          'many': 4369,
          'reasons': 5642,
          'called': 1626,
          'bbd': 1268,
          'thts': 6908,
          'chikku': 1799,
          'abt': 774,
          'dvg': 2508,
          'cold': 1904,
          'heard': 3393,
          'tht': 6907,
          'vinobanagar': 7315,
          'violence': 7317,
          'condition': 1971,
          'ru': 5864,
          'problem': 5455,
          'entry': 2629,
          'weekly': 7464,
          'comp': 1939,
          'chance': 1736,
          'win': 7544,
          'an': 968,
          'ipod': 3752,
          'txt': 7114,
          'pod': 5311,
          '80182': 649,
          'std': 6489,
          'rate': 5606,
          'apply': 1034,
          '08452810073': 70,
          'details': 2287,
          '18': 319,
          'its': 3779,
          'good': 3200,
          'hear': 3392,
          'from': 3031,
          '4mths': 524,
          'half': 3319,
          'price': 5434,
          'orange': 5000,
          'line': 4147,
          'rental': 5722,
          'latest': 4051,
          'camera': 1640,
          'phones': 5219,
          'had': 3308,
          'your': 7735,
          'phone': 5215,
          '11mths': 272,
          'call': 1619,
          'mobilesdirect': 4566,
          'on': 4958,
          '08000938767': 53,
          'update': 7209,
          'now': 4862,
          'or2stoptxt': 4998,
          'not': 4847,
          'that': 6833,
          'know': 3982,
          'most': 4612,
          'people': 5170,
          'up': 7204,
          'here': 3429,
          'still': 6507,
          'out': 5035,
          'town': 7020,
          'today': 6964,
          'am': 946,
          'going': 3187,
          'college': 1913,
          'able': 765,
          'atten': 1138,
          'class': 1847,
          'morning': 4607,
          'plz': 5292,
          'sir': 6206,
          'remain': 5705,
          'unconvinced': 7162,
          'isn': 3768,
          'elaborate': 2567,
          'test': 6807,
          'willpower': 7543,
          'also': 937,
          'track': 7023,
          'down': 2438,
          'lighters': 4131,
          'can': 1644,
          'find': 2866,
          'give': 3161,
          'back': 1203,
          'id': 3608,
          'proof': 5489,
          'lt': 4272,
          'gt': 3275,
          'rs': 5859,
          'wont': 7606,
          'allow': 927,
          'work': 7614,
          'come': 1924,
          'home': 3490,
          'within': 7573,
          'juz': 3902,
          'google': 3211,
          'search': 5989,
          'qet': 5549,
          'likely': 4135,
          'mittelschmertz': 4544,
          'paracetamol': 5101,
          'worry': 7624,
          'means': 4431,
          'got': 3217,
          'epi': 2635,
          'fine': 2869,
          'haven': 3373,
          'eaten': 2527,
          'all': 923,
          'day': 2197,
          'sitting': 6216,
          'staring': 6468,
          'at': 1127,
          'juicy': 3883,
          'pizza': 5255,
          'eat': 2526,
          'these': 6856,
          'meds': 4441,
          'ruining': 5871,
          'life': 4123,
          'wonders': 7605,
          'world': 7620,
          '7th': 638,
          '6th': 612,
          'ur': 7220,
          'style': 6572,
          '5th': 566,
          'smile': 6278,
          '4th': 528,
          'personality': 5192,
          '3rd': 472,
          'nature': 4727,
          '2nd': 406,
          'sms': 6289,
          '1st': 335,
          'lovely': 4252,
          'friendship': 3018,
          'dear': 2207,
          'todays': 6965,
          'voda': 7336,
          'numbers': 4879,
          'ending': 2596,
          'with': 7570,
          '7634': 626,
          'selected': 6019,
          'receive': 5650,
          '350': 449,
          'reward': 5793,
          'match': 4397,
          'please': 5279,
          '08712300220': 102,
          'quoting': 5572,
          'claim': 1839,
          'code': 1896,
          '7684': 627,
          'standard': 6462,
          'rates': 5607,
          'yes': 7715,
          'tv': 7104,
          'always': 944,
          'available': 1164,
          'place': 5257,
          'damn': 2167,
          'make': 4344,
          'tonight': 6989,
          'want': 7396,
          'just': 3896,
          'wait': 7373,
          'til': 6926,
          'tomorrow': 6982,
          'notice': 4853,
          'looking': 4220,
          'shit': 6116,
          'mirror': 4522,
          'youre': 7736,
          'turning': 7101,
          'into': 3731,
          'right': 5804,
          'freak': 2991,
          'hope': 3506,
          're': 5620,
          'having': 3377,
          'too': 6993,
          'much': 4655,
          'fun': 3055,
          'without': 7574,
          'see': 6006,
          'love': 4249,
          'jess': 3839,
          'ok': 4937,
          'feel': 2815,
          'john': 3858,
          'lennon': 4099,
          'being': 1321,
          'ripped': 5816,
          'off': 4915,
          'www': 7664,
          'clubmoby': 1877,
          'com': 1918,
          '08717509990': 131,
          'poly': 5322,
          'pix': 5253,
          'ringtones': 5813,
          'games': 3084,
          'six': 6220,
          'downloads': 2441,
          'haha': 3310,
          'think': 6868,
          'did': 2312,
          'thank': 6826,
          'calling': 1632,
          'forgot': 2964,
          'say': 5948,
          'happy': 3351,
          'onam': 4959,
          'sirji': 6207,
          'remembered': 5709,
          'when': 7507,
          'met': 4482,
          'insurance': 3717,
          'person': 5189,
          'meet': 4443,
          'qatar': 5547,
          'insha': 3704,
          'allah': 924,
          'rakhesh': 5593,
          'ex': 2701,
          'tata': 6740,
          'aig': 891,
          'who': 7518,
          'joined': 3860,
          'tissco': 6939,
          'tayseer': 6751,
          'lazy': 4066,
          'type': 7126,
          'lect': 4086,
          'saw': 5947,
          'pouch': 5367,
          'but': 1587,
          'nice': 4787,
          'as': 1098,
          'request': 5743,
          'melle': 4454,
          'oru': 5020,
          'minnaminunginte': 4513,
          'nurungu': 4881,
          'vettam': 7297,
          'has': 3360,
          'been': 1304,
          'set': 6057,
          'callertune': 1629,
          'callers': 1628,
          'press': 5422,
          'copy': 2029,
          'friends': 3016,
          'new': 4776,
          'car': 1664,
          'house': 3535,
          'parents': 5110,
          'hand': 3327,
          'ditto': 2373,
          'won': 7599,
          'saying': 5951,
          'anything': 1013,
          'anymore': 1007,
          'said': 5901,
          'last': 4044,
          'night': 4795,
          'whatever': 7501,
          'll': 4181,
          'same': 5914,
          'peace': 5157,
          'hmm': 3469,
          'bad': 1206,
          'news': 4780,
          'hype': 3590,
          'park': 5113,
          'plaza': 5277,
          '700': 615,
          'studio': 6563,
          'taken': 6716,
          'left': 4089,
          'bedrm': 1300,
          '900': 727,
          'alrite': 936,
          'jod': 3855,
          'hows': 3544,
          'revision': 5792,
          'goin': 3185,
          'keris': 3939,
          'bin': 1369,
          'doin': 2407,
          'smidgin': 6277,
          'way': 7430,
          'wanna': 7394,
          'cum': 2130,
          'over': 5053,
          'after': 863,
          'xx': 7678,
          'mom': 4583,
          'wants': 7399,
          'where': 7509,
          'money': 4589,
          'wining': 7552,
          'number': 4878,
          '946': 734,
          'next': 4783,
          'everyone': 2685,
          'might': 4491,
          'cancer': 1651,
          'throat': 6898,
          'hurts': 3581,
          'talk': 6722,
          'be': 1282,
          'answering': 996,
          'everyones': 2686,
          'calls': 1634,
          'one': 4964,
          'more': 4604,
          'babysitting': 1201,
          'monday': 4588,
          'sorry': 6354,
          'hurt': 3579,
          'cant': 1655,
          'anyone': 1008,
          'room': 5840,
          'top': 6999,
          'head': 3383,
          'luck': 4276,
          'draw': 2452,
          'takes': 6717,
          '28th': 378,
          'feb': 2810,
          '06': 20,
          'removal': 5716,
          '87239': 704,
          'customer': 2145,
          'services': 6054,
          '08708034412': 93,
          'maybe': 4418,
          'leave': 4083,
          'credit': 2088,
          'card': 1665,
          'gas': 3096,
          'explicit': 2737,
          'sex': 6066,
          '30': 425,
          'secs': 6001,
          '02073162414': 11,
          'costs': 2044,
          '20p': 355,
          'min': 4500,
          'accept': 783,
          'brother': 1536,
          'sister': 6209,
          'lover': 4254,
          'dear1': 2208,
          'best1': 1341,
          'clos1': 1863,
          'lvblefrnd': 4292,
          'jstfrnd': 3879,
          'cutefrnd': 2150,
          'lifpartnr': 4126,
          'belovd': 1330,
          'swtheart': 6690,
          'bstfrnd': 1548,
          'rply': 5857,
          'enemy': 2602,
          '88066': 708,
          'lost': 4235,
          '3pound': 470,
          'help': 3415,
          'pick': 5230,
          'something': 6330,
          'important': 3649,
          'there': 6854,
          'tell': 6781,
          'hrishi': 3549,
          'wan': 7391,
          'combine': 1923,
          'parts': 5124,
          'how': 3539,
          'rest': 5768,
          'project': 5475,
          'texted': 6817,
          'finished': 2874,
          'long': 4212,
          'ago': 881,
          'showered': 6150,
          'er': 2637,
          'ything': 7747,
          'working': 7618,
          'except': 2707,
          'saturday': 5939,
          'sunday': 6626,
          'cashbin': 1691,
          'co': 1887,
          'lots': 4238,
          'cash': 1690,
          'welcome': 7475,
          'biggest': 1360,
          'ever': 2679,
          'away': 1180,
          'cup': 2133,
          'coffee': 1897,
          'animation': 977,
          'could': 2050,
          'read': 5625,
          'answered': 994,
          'im': 3636,
          'gonnamissu': 3199,
          'would': 7634,
          'il': 3631,
          'postcard': 5359,
          'buttheres': 1589,
          'aboutas': 768,
          'merememberin': 4472,
          'asthere': 1123,
          'ofsi': 4927,
          'breakin': 1502,
          'his': 3456,
          'contract': 2007,
          'luv': 4286,
          'yaxx': 7700,
          'neshanth': 4763,
          'tel': 6776,
          'aiyar': 901,
          'hard': 3352,
          'later': 4050,
          'scold': 5966,
          'month': 4598,
          'kotees': 3995,
          'birthday': 1376,
          'announcement': 985,
          'recently': 5655,
          'tried': 7058,
          'delivery': 2253,
          'were': 7486,
          'unable': 7153,
          '07090298926': 25,
          'schedule': 5962,
          'ref': 5674,
          '9307622': 732,
          'sleep': 6241,
          'whole': 7520,
          'appreciated': 1039,
          'two': 7112,
          'dad': 2161,
          'map': 4370,
          'reading': 5628,
          'semi': 6029,
          'argument': 1070,
          'apart': 1019,
          'things': 6867,
          'message': 4475,
          'using': 7246,
          'auction': 1148,
          'subscription': 6589,
          '150p': 302,
          'msgrcvd': 4640,
          'skip': 6231,
          'unsubscribe': 7199,
          'customercare': 2146,
          '08718726270': 139,
          'little': 4172,
          'child': 1800,
          'afraid': 860,
          'dark': 2177,
          'become': 1294,
          'teenager': 6773,
          'stay': 6484,
          'cmon': 1884,
          'babe': 1195,
          'horny': 3516,
          'turn': 7099,
          'fantasy': 2786,
          'hot': 3526,
          'sticky': 6504,
          'need': 4747,
          'replies': 5735,
          'cost': 2041,
          '50': 538,
          'cancel': 1648,
          'sch': 5961,
          'mon': 4587,
          'sis': 6208,
          'take': 6714,
          'smth': 6293,
          'goal': 3177,
          'arsenal': 1091,
          'henry': 3426,
          'liverpool': 4176,
          'scores': 5970,
          'simple': 6189,
          'shot': 6140,
          'yards': 7698,
          'pass': 5128,
          'by': 1602,
          'bergkamp': 1338,
          'margin': 4376,
          '78': 630,
          'mins': 4515,
          'hey': 3437,
          'boys': 1484,
          'xxx': 7680,
          'pics': 5234,
          'sent': 6040,
          'direct': 2344,
          'porn': 5340,
          '69855': 597,
          '24hrs': 366,
          '50p': 544,
          'text': 6812,
          'stopbcm': 6522,
          'sf': 6073,
          'wc1n3xx': 7434,
          'stupid': 6571,
          'sending': 6034,
          'those': 6885,
          'pandy': 5089,
          'mental': 4466,
          'try': 7077,
          'week': 7460,
          'end': 2594,
          'course': 2062,
          'coimbatore': 1899,
          'midnight': 4489,
          'earliest': 2515,
          'alright': 935,
          'tyler': 7125,
          'minor': 4514,
          'crisis': 2098,
          'sooner': 6345,
          'than': 6823,
          'thought': 6889,
          'asap': 1100,
          'which': 7513,
          'rooms': 5844,
          'cheap': 1765,
          'hi': 3440,
          'msg': 4637,
          'office': 4921,
          'check': 1770,
          'befor': 1309,
          'activities': 813,
          'together': 6968,
          'should': 6141,
          'thinkin': 6870,
          'him': 3450,
          'loan': 4189,
          'purpose': 5535,
          '500': 539,
          '75': 621,
          '000': 1,
          'homeowners': 3491,
          'tenants': 6791,
          'previously': 5432,
          'refused': 5684,
          '0800': 48,
          '1956669': 325,
          'sleeping': 6243,
          'feeling': 2817,
          'well': 7477,
          'christmas': 1828,
          'occasion': 4905,
          'celebrated': 1716,
          'reflection': 5679,
          'values': 7269,
          'desires': 2279,
          'affections': 858,
          'amp': 961,
          'traditions': 7026,
          'ideal': 3611,
          'awesome': 1181,
          'bit': 1378,
          'first': 2884,
          'ask': 1107,
          'hello': 3412,
          'didn': 2313,
          'quite': 5565,
          'limping': 4145,
          'slowly': 6263,
          'followed': 2935,
          'aa': 751,
          'exhaust': 2717,
          'hanging': 3337,
          'gym': 3300,
          'goodmorning': 3204,
          'late': 4047,
          '1hr': 330,
          'thnx': 6883,
          'dude': 2491,
          'guys': 3297,
          '2nite': 409,
          'watch': 7417,
          'infernal': 3686,
          'affair': 854,
          'food': 2945,
          'ringtone': 5811,
          'sub': 6577,
          'subpoly': 6581,
          '81618': 658,
          '08718727870': 145,
          'getting': 3141,
          'thinking': 6871,
          'staying': 6487,
          'mcr': 4424,
          'seems': 6013,
          'unnecessarily': 7191,
          'hostile': 3525,
          'exam': 2704,
          'march': 4373,
          'daddy': 2162,
          'care': 1669,
          'interflora': 3725,
          'order': 5004,
          'flowers': 2922,
          '505060': 541,
          'before': 1310,
          'ass': 1116,
          'enjoy': 2610,
          'doggy': 2404,
          'beauty': 1290,
          'pimples': 5246,
          'winner': 7553,
          'specially': 6392,
          '1000': 253,
          'holiday': 3485,
          'flights': 2908,
          'inc': 3657,
          'speak': 6386,
          'live': 4173,
          'operator': 4985,
          '0871277810810': 114,
          'lets': 4108,
          'use': 7238,
          'princess': 5440,
          'ah': 883,
          'poor': 5335,
          'baby': 1198,
          'urfeeling': 7223,
          'bettersn': 1346,
          'probthat': 5462,
          'overdose': 5056,
          'careful': 1674,
          'spk': 6417,
          'sn': 6294,
          'lovejen': 4251,
          'prabha': 5381,
          'soryda': 6361,
          'realy': 5637,
          'frm': 3021,
          'heart': 3396,
          'sory': 6360,
          'joy': 3876,
          'father': 2797,
          '____': 748,
          'ans': 991,
          'ths': 6906,
          'hav': 3367,
          'iq': 3753,
          'tis': 6938,
          'ias': 3594,
          'question': 5558,
          'answer': 993,
          'reply': 5736,
          'age': 873,
          'gender': 3119,
          'begin': 1315,
          '24m': 367,
          '2p': 411,
          'germany': 3133,
          '08448350055': 66,
          'planettalkinstant': 5265,
          'info': 3688,
          'opt': 4991,
          'urgent': 7224,
          'awarded': 1179,
          '2000': 346,
          'prize': 5448,
          'guaranteed': 3279,
          '09058094455': 175,
          'land': 4026,
          '3030': 431,
          'valid': 7264,
          '12hrs': 284,
          'lor': 4228,
          'matter': 4408,
          'break': 1499,
          'howz': 3546,
          'pain': 5081,
          'sounds': 6366,
          'crazy': 2081,
          'dunno': 2500,
          'tahan': 6711,
          'anot': 989,
          'wat': 7416,
          'doing': 2409,
          'coughing': 2049,
          'nothing': 4852,
          'coming': 1930,
          'address': 825,
          'voucher': 7348,
          'holder': 3482,
          'weeks': 7465,
          'offer': 4917,
          'pc': 5155,
          'http': 3553,
          'wtlp': 7658,
          'ts': 7080,
          'cs': 2112,
          'ya': 7690,
          'cookies': 2020,
          'jelly': 3829,
          'them': 6844,
          'tunji': 7097,
          'queen': 5556,
          'wishing': 7566,
          'abiola': 763,
          'hr': 3548,
          'b4': 1189,
          'prepare': 5409,
          'lol': 4205,
          'trust': 7074,
          'luton': 4285,
          '0125698789': 8,
          'around': 1082,
          'movie': 4627,
          'theatre': 6839,
          'unlimited': 7189,
          'movies': 4628,
          'pay': 5145,
          'once': 4962,
          'dnt': 2382,
          'wnt': 7591,
          'tlk': 6947,
          'wid': 7531,
          'okey': 4941,
          'doke': 2411,
          'dressed': 2459,
          'cos': 2039,
          'laying': 4064,
          'ill': 3632,
          'bout': 1461,
          'times': 6929,
          'stuff': 6567,
          'mum': 4661,
          'workin': 7617,
          'probably': 5454,
          'pop': 5337,
          'miss': 4529,
          'special': 6388,
          'sight': 6174,
          'very': 7296,
          'till': 6927,
          'maintain': 4342,
          'ends': 2600,
          'sh': 6075,
          'jas': 3814,
          'habit': 3306,
          'nan': 4710,
          'bari': 1239,
          'hudgi': 3557,
          'yorge': 7729,
          'pataistha': 5136,
          'ertini': 2646,
          'kano': 3917,
          'glad': 3165,
          'talking': 6726,
          'hmmm': 3470,
          'guess': 3284,
          'kb': 3928,
          'yoga': 7726,
          'lo': 4186,
          'oso': 5023,
          'liao': 4114,
          'loud': 4244,
          'scream': 5979,
          'minutes': 4519,
          'cause': 1702,
          'gyno': 3303,
          'shoving': 6147,
          'belong': 1328,
          'raji': 5589,
          'pls': 5286,
          'favour': 2805,
          'convey': 2013,
          'wishes': 7564,
          'nimya': 4804,
          'ugh': 7136,
          'fuck': 3041,
          'resubbing': 5774,
          'eve': 2671,
          'someone': 6325,
          'dating': 2192,
          'contact': 1997,
          '09058091854': 172,
          'revealed': 5788,
          'po': 5295,
          'box385': 1472,
          'm6': 4302,
          '6wu': 614,
          'trek': 7055,
          'family': 2777,
          'understanding': 7167,
          've': 7282,
          'trying': 7079,
          'sura': 6648,
          'shop': 6128,
          'drop': 2471,
          'either': 2564,
          '10k': 261,
          '5k': 561,
          '100': 252,
          'travel': 7043,
          '09064011000': 204,
          'ntt': 4872,
          'box': 1464,
          'cr01327bt': 2069,
          'fixedline': 2894,
          '150ppm': 306,
          'vary': 7275,
          'sad': 5895,
          'thing': 6866,
          'past': 5135,
          'fancy': 2782,
          'shag': 6079,
          'interested': 3723,
          'sextextuk': 6068,
          'xxuk': 7679,
          'suzy': 6663,
          '69876': 599,
          'txts': 7121,
          'tncs': 6957,
          'website': 7450,
          'oh': 4932,
          'really': 5636,
          'perform': 5176,
          'write': 7641,
          'paper': 5098,
          'huh': 3563,
          'private': 5445,
          '2003': 347,
          'account': 795,
          'statement': 6480,
          'fone': 2941,
          'shows': 6155,
          '800': 641,
          'un': 7152,
          'redeemed': 5671,
          'points': 5315,
          ...}
```

In [57]:

```
X_train_np = X_train_cv.toarray()
X_train_np[0]
```

Out[57]:

```
array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

In [58]:

```
np.where(X_train_np[0]!=0)
```

Out[58]:

```
(array([ 934, 1322, 2498, 3428, 3772, 4095, 4680, 6097, 6848, 6878, 6967,
        6973, 7010, 7069, 7234], dtype=int64),)
```

Train the naive bayes model

In [64]:

```
from sklearn.naive_bayes import MultinomialNB

model = MultinomialNB()
model.fit(X_train_cv, y_train)
```

Out[64]:

```
MultinomialNB()
```

In [65]:
```
X_test_cv = v.transform(X_test)
```

Evaluate Performance

In [67]:
```
from sklearn.metrics import classification_report
```

In [69]:
```
from sklearn.metrics import classification_report

y_pred = model.predict(X_test_cv)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       975
           1       0.97      0.96      0.96       140

    accuracy                           0.99      1115
   macro avg       0.98      0.98      0.98      1115
weighted avg       0.99      0.99      0.99      1115
```

In [70]:
```
emails = [
    'Hey mohan, can we get together to watch footbal game tomorrow?',
    'Upto 20% discount on parking, exclusive offer just for you. Dont miss this re
]

emails_count = v.transform(emails)
model.predict(emails_count)
```

Out[70]:
```
array([0, 1], dtype=int64)
```

Train the model using sklearn pipeline and reduce number of lines of code

In [71]:
```
from sklearn.pipeline import Pipeline

clf = Pipeline([
    ('vectorizer', CountVectorizer()),
    ('nb', MultinomialNB())
])
```

In [72]:
```
clf.fit(X_train, y_train)
```

Out[72]:
```
Pipeline(steps=[('vectorizer', CountVectorizer()), ('nb', MultinomialNB())])
```

In [73]:
```
y_pred = clf.predict(X_test)

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       975
           1       0.97      0.96      0.96       140

    accuracy                           0.99      1115
   macro avg       0.98      0.98      0.98      1115
weighted avg       0.99      0.99      0.99      1115
```

In [ ]: