In [1]:
```
from urllib.request import urlretrieve
import pandas as pd
import numpy as np
```

1. Downlod and Load the Data

In [2]:
```
urlretrieve("https://raw.githubusercontent.com/codebasics/nlp-tutorials/refs/heads/
```

Out[2]:
```
('./dataset/emails.csv', <http.client.HTTPMessage at 0x205c5937d40>)
```

In [3]:
```
df = pd.read_csv('./dataset/emails.csv')
```

2. Analyze the dataset

In [4]:
```
df.head()
```

Out[4]:

	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 [5]:
```
df.shape
```

Out[5]:
```
(5572, 2)
```

In [6]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Category  5572 non-null   object
 1   Message   5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

In [7]:
```
df.groupby('Category').count()
```

Out[7]:

Category	Message
ham	4825
spam	747

3. Pre-Processing the categorical data

In [8]:
```
df['spam'] = df.Category.apply(lambda x: 1 if x == 'spam' else 0)
```

In [9]:
```
df
```

Out[9]:

	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
...	...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...	1
5568	ham	Will ü b going to esplanade fr home?	0
5569	ham	Pity, * was in mood for that. So...any other s...	0
5570	ham	The guy did some bitching but I acted like i'd...	0
5571	ham	Rofl. Its true to its name	0

5572 rows × 3 columns

In [10]:
```
x = df['Message']
y = df['spam']
```

In [11]:
```
y
```

Out[11]:
```
0       0
1       0
2       1
3       0
4       0
       ..
5567    1
5568    0
5569    0
5570    0
5571    0
Name: spam, Length: 5572, dtype: int64
```

4. Split the dataset into training and testing portions

In [12]:
```
from sklearn.model_selection import train_test_split
```

```
C:\Anaconda3\Lib\site-packages\sklearn\utils\_param_validation.py:14: UserWarning: A
NumPy version >=1.22.4 and <2.3.0 is required for this version of SciPy (detected ve
rsion 2.3.2)
  from scipy.sparse import csr_matrix, issparse
```

In [13]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2)
```

In [14]:
```
x_train.shape
```

Out[14]:
```
(4457,)
```

In [15]:
```
y_train.shape
```

Out[15]:
```
(4457,)
```

In [16]:
```
y_test.shape
```

Out[16]:
```
(1115,)
```

In [17]:
```
x_test.shape
```

Out[17]:
```
(1115,)
```

In [18]:
```
type(x_test)
```

Out[18]:
```
pandas.core.series.Series
```

In [19]:
```
x_train[0:5]
```

Out[19]:
```
2965    Do you ever notice that when you're driving, a...
2948                 Leave it. U will always be ignorant.
2810    Oh yeah I forgot. U can only take 2 out shoppi...
4172    Pls what's the full name of joke's school cos ...
2124                                            #ERROR!
Name: Message, dtype: object
```

In [20]:
```
x_train
```

Out[20]:
```
2965    Do you ever notice that when you're driving, a...
2948                 Leave it. U will always be ignorant.
2810    Oh yeah I forgot. U can only take 2 out shoppi...
4172    Pls what's the full name of joke's school cos ...
2124                                            #ERROR!
                            ...
2762    I am not sure about night menu. . . I know onl...
2800    I've told him that i've returned it. That shou...
5202             WOT STUDENT DISCOUNT CAN U GET ON BOOKS?
296     T-Mobile customer you may now claim your FREE ...
1430    For sale - arsenal dartboard. Good condition b...
Name: Message, Length: 4457, dtype: object
```

4. Pre-Processing the message data using Bag Of Words (BOF) method

In [21]:
```
from sklearn.feature_extraction.text import CountVectorizer
```

In [22]:
```
v = CountVectorizer()
x_train_cv = v.fit_transform(x_train.values)
x_train_cv
```

Out[22]:
```
<4457x7805 sparse matrix of type '<class 'numpy.int64'>'
        with 59513 stored elements in Compressed Sparse Row format>
```

In [23]:
```
x_train_cv.toarray()
```

Out[23]:
```
array([[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]], shape=(4457, 7805))
```

In [24]:
```
x_train_cv.shape
```

Out[24]:
```
(4457, 7805)
```

In [25]:
```
v.get_feature_names_out()[1000:1100]
```

Out[25]:
```
array(['ans', 'ansr', 'answer', 'answered', 'answerin', 'answering',
       'answers', 'answr', 'antha', 'anthony', 'anti', 'any', 'anybody',
       'anyhow', 'anymore', 'anyone', 'anyones', 'anyplaces', 'anythiing',
       'anythin', 'anything', 'anytime', 'anyway', 'anyways', 'anywhere',
       'aom', 'apart', 'apartment', 'apes', 'apeshit', 'aphex', 'apnt',
       'apo', 'apologetic', 'apologise', 'apologize', 'apology', 'app',
       'apparently', 'appear', 'applausestore', 'applebees', 'apples',
       'application', 'apply', 'applyed', 'applying', 'appointment',
       'appointments', 'appreciate', 'appreciated', 'approaches',
       'approaching', 'appropriate', 'approve', 'approx', 'apps', 'appt',
       'appy', 'april', 'aproach', 'apt', 'aptitude', 'aquarius', 'ar',
       'arab', 'arabian', 'arcade', 'archive', 'ard', 'are', 'area',
       'aren', 'arent', 'arestaurant', 'areyouunique', 'argentina',
       'argh', 'argue', 'arguing', 'argument', 'arguments', 'aries',
       'arise', 'arises', 'arm', 'armand', 'armenia', 'arms', 'arng',
       'arngd', 'arnt', 'around', 'aroundn', 'arr', 'arrange', 'arrested',
       'arrival', 'arrive', 'arrived'], dtype=object)
```

In [26]:
```
v.get_feature_names_out().shape
```

Out[26]:
```
(7805,)
```

In [ ]:
```
dir(v)
```

In [28]:
```
v.vocabulary_
```

Out[28]:

```
{'do': 2405,
 'you': 7768,
 'ever': 2697,
 'notice': 4883,
 'that': 6873,
 'when': 7545,
 're': 5649,
 'driving': 2485,
 'anyone': 1015,
 'going': 3219,
 'slower': 6303,
 'than': 6861,
 'is': 3810,
 'an': 976,
 'idiot': 3652,
 'and': 980,
 'everyone': 2703,
 'faster': 2816,
 'maniac': 4401,
 'leave': 4117,
 'it': 3822,
 'will': 7584,
 'always': 949,
 'be': 1297,
 'ignorant': 3661,
 'oh': 4965,
 'yeah': 7740,
 'forgot': 2990,
 'can': 1658,
 'only': 5001,
 'take': 6750,
 'out': 5070,
 'shopping': 6179,
 'at': 1139,
 'once': 4993,
 'pls': 5317,
 'what': 7539,
 'the': 6877,
 'full': 3082,
 'name': 4738,
 'of': 4943,
 'joke': 3902,
 'school': 6008,
 'cos': 2042,
 'fees': 2845,
 'in': 3703,
 'university': 7228,
 'florida': 2943,
 'seem': 6060,
 'to': 6995,
 'actually': 825,
 'lt': 4301,
 'gt': 3305,
 'holla': 3520,
 'back': 1215,
 'error': 2662,
```

```
'total': 7057,
'disappointment': 2371,
'texted': 6853,
'was': 7454,
'craziest': 2084,
'shit': 6164,
'got': 3245,
'ill': 3671,
'down': 2456,
'soon': 6384,
'dude': 2508,
'we': 7479,
'should': 6188,
'go': 3207,
'sup': 6672,
'again': 876,
'ambrith': 955,
'madurai': 4358,
'met': 4518,
'arun': 1107,
'dha': 2315,
'marrge': 4417,
'remembr': 5753,
'short': 6181,
'but': 1604,
'cute': 2153,
'good': 3230,
'person': 5224,
'dont': 2437,
'try': 7121,
'prove': 5527,
'gud': 3311,
'noon': 4862,
'nothing': 4882,
'just': 3935,
'getting': 3172,
'msgs': 4676,
'by': 1619,
'dis': 2369,
'wit': 7612,
'different': 2345,
'no': 4845,
'urgent': 7266,
'are': 1070,
'trying': 7123,
'contact': 2002,
'todays': 7003,
'draw': 2469,
'shows': 6202,
'have': 3407,
'won': 7641,
'800': 654,
'prize': 5479,
'guaranteed': 3309,
'call': 1633,
'09050003091': 175,
```

```
'from': 3063,
'land': 4060,
'line': 4177,
'claim': 1846,
'c52': 1622,
'valid12hrs': 7307,
'think': 6904,
'could': 2052,
'stop': 6554,
'like': 4167,
'hour': 3566,
'or': 5034,
'so': 6348,
'my': 4720,
'roommate': 5885,
'looking': 4254,
'stock': 6546,
'up': 7248,
'for': 2978,
'trip': 7103,
'thought': 6924,
'lemme': 4130,
'know': 4018,
'if': 3657,
'anything': 1020,
'goin': 3217,
'on': 4989,
'later': 4086,
'morning': 4641,
'princess': 5470,
'happy': 3382,
'new': 4811,
'year': 7741,
'beerage': 1319,
'wylie': 7708,
'update': 7252,
'weed': 7500,
'dealer': 2214,
'carlos': 1693,
'went': 7522,
'freedom': 3027,
'had': 3337,
'class': 1853,
'with': 7613,
'lunsford': 4315,
'ok': 4970,
'lor': 4260,
'wan': 7436,
'me': 4460,
'look': 4251,
'get': 3164,
'monthly': 4631,
'password': 5166,
'wap': 7446,
'mobsi': 4604,
'com': 1925,
```

```
'391784': 469,
'use': 7280,
'your': 7773,
'phone': 5249,
'not': 4878,
'pc': 5190,
'please': 5311,
'09066612661': 239,
'landline': 4062,
'5000': 558,
'cash': 1702,
'luxury': 4321,
'canary': 1661,
'islands': 3816,
'holiday': 3519,
'await': 1186,
'collection': 1918,
'cs': 2114,
'sae': 5939,
'award': 1189,
'20m12aq': 365,
'150ppm': 316,
'16': 324,
'horrible': 3553,
'gal': 3107,
'knew': 4015,
'dat': 2195,
'wif': 7574,
'him': 3485,
'yest': 7753,
'still': 6543,
'come': 1931,
'ask': 1116,
'congrats': 1987,
'nokia': 4853,
'3650': 462,
'video': 7346,
'camera': 1654,
'09066382422': 238,
'calls': 1648,
'cost': 2044,
'ave': 1177,
'3mins': 478,
'vary': 7316,
'mobiles': 4598,
'close': 1871,
'300603': 441,
'post': 5384,
'bcm4284': 1290,
'ldn': 4105,
'wc1n3xx': 7478,
'much': 4687,
'better': 1352,
'now': 4890,
'thanks': 6865,
'lol': 4242,
```

```
'taxt': 6785,
'massage': 4425,
'tie': 6952,
'pos': 5372,
'argh': 1077,
'lool': 4256,
'regret': 5730,
'inform': 3736,
'nhs': 4819,
'has': 3392,
'made': 4354,
'mistake': 4574,
'were': 7525,
'never': 4808,
'born': 1465,
'report': 5781,
'yor': 7766,
'local': 4226,
'hospital': 3556,
'2b': 391,
'terminated': 6834,
'sorry': 6392,
'inconvenience': 3715,
'most': 4645,
'beautiful': 1303,
'girl': 3186,
'ive': 3835,
'seen': 6063,
'baby': 1210,
'common': 1942,
'room': 5883,
'waaaat': 7406,
'lololo': 4243,
'next': 4818,
'time': 6963,
'then': 6888,
'hi': 3475,
'darlin': 2188,
'really': 5668,
'nice': 4822,
'night': 4829,
'lift': 4160,
'see': 6055,
'tomorrow': 7021,
'xxx': 7718,
'remind': 5755,
'how': 3574,
'there': 6892,
'shall': 6128,
'very': 7335,
'babe': 1207,
'woo': 7649,
'hoo': 3534,
'party': 5157,
'work': 7657,
'where': 7549,
```

```
'home': 3524,
'calicut': 1631,
'its': 3830,
'poking': 5346,
'man': 4389,
'everyday': 2702,
'they': 6898,
'teach': 6795,
'canada': 1659,
'abi': 772,
'saying': 5997,
'wat': 7460,
'makes': 4380,
'thk': 6913,
'll': 4216,
'fall': 2794,
'quite': 5594,
'prone': 5515,
'falls': 2797,
'lucky': 4307,
'dad': 2167,
'fetch': 2853,
'already': 939,
'honestly': 3528,
've': 7322,
'lovely': 4282,
'cup': 2136,
'tea': 6794,
'promptly': 5513,
'dropped': 2488,
'keys': 3979,
'burnt': 1596,
'fingers': 2894,
'them': 6884,
'docs': 2409,
'appointments': 1048,
'week': 7501,
'tired': 6972,
'shoving': 6193,
'stuff': 6601,
'ugh': 7180,
'why': 7569,
'couldn': 2054,
'normal': 4871,
'body': 1439,
'life': 4155,
'been': 1317,
'this': 6912,
'fun': 3085,
'great': 3278,
'until': 7246,
'came': 1653,
'truly': 7116,
'special': 6424,
'forget': 2986,
'enjoy': 2629,
```

```
'one': 4995,
'gbp': 3137,
'sms': 6330,
'outside': 5079,
'way': 7475,
'office': 4954,
'da': 2165,
'heard': 3427,
'want': 7441,
'tel': 6812,
'thing': 6902,
'message': 4511,
'sent': 6091,
'spook': 6462,
'mob': 4595,
'halloween': 3353,
'logo': 4238,
'pic': 5262,
'plus': 5321,
'free': 3025,
'eerie': 2565,
'tone': 7023,
'txt': 7160,
'card': 1677,
'8007': 657,
'zed': 7798,
'08701417012150p': 76,
'per': 5207,
'err': 2661,
'cud': 2127,
'8pm': 734,
'haven': 3409,
'yo': 7763,
'few': 2856,
'friends': 3046,
'asking': 1120,
'about': 777,
'working': 7661,
'all': 930,
'weekend': 7503,
'gentle': 3155,
'taking': 6755,
'inches': 3707,
'deep': 2238,
'inside': 3756,
'tight': 6955,
'pussy': 5567,
'ya': 7728,
'even': 2691,
'those': 6920,
'cookies': 2026,
'jelly': 3874,
'thank': 6864,
'selflessness': 6073,
'love': 4279,
'plenty': 5315,
```

```
'also': 942,
'andros': 984,
'ice': 3638,
'etc': 2680,
'sis': 6250,
'catching': 1711,
'show': 6194,
'afternoon': 871,
'watching': 7465,
'her': 3460,
'watch': 7461,
'today': 7002,
'tmr': 6989,
'tried': 7102,
'our': 5067,
'offer': 4950,
'750': 636,
'anytime': 1021,
'any': 1011,
'network': 4803,
'mins': 4550,
'half': 3348,
'price': 5464,
'rental': 5767,
'camcorder': 1652,
'08000930705': 50,
'reply': 5778,
'delivery': 2262,
'wed': 7493,
'remember': 5750,
'alex': 921,
'his': 3490,
'pizza': 5289,
'am': 951,
'escape': 2670,
'theatre': 6879,
'kavalan': 3962,
'minutes': 4554,
'face': 2775,
'book': 1451,
'status': 6516,
'frequently': 3038,
'save': 5990,
'yourself': 7778,
'stress': 6575,
'dorm': 2445,
'account': 802,
'send': 6081,
'details': 2302,
'money': 4624,
'vote': 7385,
'sing': 6241,
'along': 938,
'stars': 6503,
'karaoke': 3957,
'mobile': 4597,
```

```
'link': 4184,
'anyway': 1022,
'own': 5097,
'done': 2435,
'yet': 7755,
'dun': 2513,
'disturb': 2390,
'liao': 4147,
'yar': 7734,
'noe': 4848,
'used': 7281,
'route': 5893,
'too': 7035,
'creep': 2094,
'possibility': 5381,
'being': 1332,
'pub': 5544,
'who': 7561,
'say': 5994,
'drugdealer': 2493,
'honey': 3529,
'boo': 1449,
'missing': 4567,
'askd': 1117,
'question': 5587,
'some': 6361,
'hours': 3568,
'before': 1321,
'answer': 1002,
'150': 310,
'text': 6849,
'five': 2913,
'pounds': 5396,
'08000776320': 48,
'fast': 2815,
'approaching': 1052,
'wish': 7606,
'sankranti': 5965,
'republic': 5784,
'day': 2204,
'valentines': 7305,
'shivratri': 6169,
'ugadi': 7179,
'fools': 2971,
'may': 4451,
'independence': 3722,
'friendship': 3048,
'mother': 4647,
'father': 2819,
'teachers': 6797,
'childrens': 1812,
'amp': 968,
'birthday': 1387,
'ganesh': 3115,
'festival': 2852,
'dasara': 2194,
```

```
'diwali': 2396,
'christmas': 1835,
'mornings': 4642,
'afternoons': 872,
'evenings': 2693,
'nights': 4830,
'rememberi': 5752,
'first': 2908,
'wishing': 7609,
'these': 6894,
'raj': 5613,
'journey': 3914,
'let': 4140,
'need': 4781,
'receipts': 5684,
'tell': 6817,
'pendent': 5202,
'derek': 2284,
'taylor': 6786,
'walmart': 7433,
'mouse': 4655,
'desk': 2292,
'priscilla': 5476,
'ready': 5659,
'mum': 4696,
'buy': 1608,
'food': 2968,
'wanna': 7439,
'movie': 4660,
'friend': 3045,
'around': 1092,
'unless': 7232,
'guys': 3327,
'sooner': 6385,
'make': 4379,
'lasagna': 4079,
'vodka': 7379,
'pansy': 5126,
'living': 4213,
'jungle': 3932,
'two': 7158,
'years': 7742,
'more': 4637,
'worried': 7667,
'called': 1640,
'while': 7556,
'hoping': 3548,
'l8r': 4040,
'malaria': 4384,
'miss': 4563,
'bani': 1242,
'big': 1369,
'give': 3192,
'especially': 2674,
'probably': 5485,
'couple': 2060,
```

```
'tops': 7048,
'meeting': 4479,
'ge': 3143,
'nite': 4838,
'she': 6143,
'doesnt': 2418,
'test': 6845,
'boytoy': 1499,
'feeling': 2841,
'hope': 3540,
'boy': 1494,
'obedient': 4926,
'slave': 6283,
'queen': 5584,
'win': 7587,
'200': 356,
'spree': 6473,
'every': 2698,
'starting': 6507,
'play': 5303,
'store': 6563,
'88039': 717,
'skilgme': 6268,
'tscs08714740323': 7127,
'1winawk': 352,
'age16': 880,
'50perweeksub': 563,
'18': 329,
'days': 2205,
'euro2004': 2684,
'kickoff': 3987,
'kept': 3973,
'informed': 3738,
'latest': 4087,
'news': 4816,
'results': 5821,
'daily': 2171,
'unsubscribe': 7243,
'euro': 2683,
'83222': 682,
'promises': 5510,
'though': 6923,
'gotten': 3252,
'dinner': 2359,
'os': 5055,
'ubandu': 7178,
'which': 7555,
'run': 5919,
'without': 7616,
'installing': 3761,
'hard': 3383,
'disk': 2383,
'copy': 2034,
'important': 3691,
'files': 2874,
'system': 6734,
```

```
'repair': 5770,
'shop': 6177,
'449050000301': 513,
'000': 1,
'09050000301': 167,
'wanted': 7443,
'score': 6014,
'might': 4529,
'relax': 5738,
'motivating': 4650,
'sharing': 6138,
'dear': 2217,
'urgh': 7268,
'coach': 1893,
'hot': 3561,
'smells': 6316,
'chip': 1820,
'fat': 2818,
'duvet': 2522,
'predictive': 5432,
'word': 7655,
'moji': 4614,
'saved': 5991,
'lives': 4212,
'happen': 3372,
'truth': 7120,
'did': 2331,
'yes': 7752,
'cheesy': 1791,
'songs': 6380,
'frosty': 3068,
'snowman': 6346,
'oops': 5012,
'somerset': 6366,
'bit': 1389,
'far': 2811,
'tomo': 7019,
'tot': 7056,
'group': 3296,
'mate': 4434,
'havent': 3410,
'double': 2449,
'eviction': 2709,
'spiral': 6444,
'michael': 4524,
'riddance': 5843,
'haha': 3340,
'sounds': 6402,
'crazy': 2085,
'dunno': 2515,
'tahan': 6745,
'anot': 998,
'sad': 5938,
'story': 6567,
'last': 4080,
'wife': 7575,
```

```
'nt': 4896,
'parents': 5142,
'kids': 3991,
'colleagues': 1914,
'as': 1108,
'entered': 2638,
'cabin': 1624,
'pa': 5104,
'said': 5946,
'boss': 1467,
'felt': 2849,
'lunch': 4313,
'after': 869,
'invited': 3787,
'apartment': 1027,
'mind': 4539,
'into': 3780,
'bedroom': 1315,
'minute': 4553,
'sed': 6054,
'sexy': 6116,
'mood': 4634,
'minuts': 4555,
'latr': 4088,
'wid': 7572,
'cake': 1626,
'kidz': 3992,
'screaming': 6026,
'surprise': 6695,
'waiting': 7418,
'sofa': 6351,
'naked': 4735,
'heart': 3429,
'empty': 2609,
'wisdom': 7604,
'eyes': 2770,
'dreams': 2472,
'frnds': 3055,
'alwys': 950,
'touch': 7060,
'sweet': 6713,
'lar': 4074,
'ba': 1203,
'dao': 2182,
'pm': 5323,
'ah': 888,
'would': 7678,
'fri': 3042,
'he': 3417,
'hmm': 3502,
'thinking': 6906,
'ac': 786,
'blind': 1413,
'date': 2197,
'4u': 549,
'rodds1': 5873,
```

```
'21': 367,
'aberdeen': 771,
'united': 7226,
'kingdom': 4002,
'check': 1780,
'http': 3590,
'img': 3682,
'icmb3cktz8r7': 3642,
'dates': 2199,
'hide': 3477,
'4mths': 541,
'orange': 5037,
'phones': 5252,
'11mths': 284,
'mobilesdirect': 4599,
'08000938767': 51,
'or2stoptxt': 5036,
'don': 2433,
'gonna': 3228,
'snow': 6343,
'flurries': 2948,
'usually': 7291,
'melt': 4491,
'hit': 3492,
'ground': 3295,
'eek': 2564,
'since': 6239,
'plz': 5322,
'ans': 1000,
'bslvyl': 1563,
'via': 7338,
'fullonsms': 3083,
'la': 4043,
'wana': 7438,
'pei': 5199,
'bf': 1360,
'oso': 5057,
'rite': 5863,
'other': 5059,
'den': 2267,
'fps': 3011,
'reminder': 5756,
'o2': 4924,
'50': 556,
'credit': 2091,
'offers': 4953,
'valid': 7306,
'house': 3569,
'postcode': 5386,
'smile': 6319,
'pleasure': 5313,
'pain': 5114,
'trouble': 7109,
'pours': 5398,
'rain': 5609,
'sum1': 6654,
```

```
'hurts': 3615,
'becoz': 1310,
'someone': 6364,
'loves': 4287,
'smiling': 6323,
'water': 7467,
'logging': 4236,
'desert': 2288,
'geoenvironmental': 3160,
'implications': 3690,
'wanting': 7444,
'allday': 932,
'didnt': 2334,
'piss': 5283,
'off': 4945,
'right': 5846,
'pdate_now': 5192,
'1000': 262,
'txts': 7166,
'tariffs': 6773,
'motorola': 4653,
'sonyericsson': 6382,
'bluetooth': 1430,
'mobileupd8': 4601,
'08000839402': 49,
'call2optout': 1635,
'yhl': 7758,
'bognor': 1441,
'splendid': 6453,
'away': 1191,
'shattered': 6139,
'stay': 6517,
'sir': 6248,
'mail': 4371,
'tunji': 7142,
'doing': 2426,
'abiola': 774,
'ned': 4780,
'convince': 2021,
'tht': 6941,
'possible': 5382,
'witot': 7618,
'hurting': 3614,
'main': 4375,
'wondering': 7646,
'wont': 7648,
'coping': 2033,
'long': 4248,
'distance': 2388,
'ago': 886,
'bank': 1243,
'honeybee': 3530,
'sweetest': 6714,
'world': 7664,
'god': 3213,
'laughed': 4090,
```

```
'wait': 7415,
'havnt': 3414,
'reading': 5658,
'msg': 4671,
'moral': 4636,
'crack': 2076,
'jokes': 3904,
'gm': 3203,
'gn': 3204,
'super': 6673,
'replacement': 5774,
'murali': 4703,
'having': 3412,
'number': 4905,
'keep': 3967,
'ten': 6828,
'rs': 5901,
'shelf': 6147,
'egg': 2572,
'confirm': 1981,
'collect': 1915,
'cheque': 1795,
'picked': 5264,
'flower': 2945,
'dippeditinadew': 2363,
'lovingly': 4290,
'touched': 7061,
'itwhichturnedinto': 3832,
'gifted': 3180,
'tomeandsaid': 7018,
'late': 4083,
'power': 5400,
'frndship': 3056,
'thangam': 6863,
'held': 3442,
'prasad': 5419,
'ringtone': 5854,
'order': 5040,
'reference': 5712,
'x49': 7709,
'charged': 1761,
'arrive': 1098,
'customer': 2148,
'services': 6104,
'09065989182': 223,
'de': 2211,
'fill': 2875,
'ur': 7262,
'visa': 7367,
'coming': 1937,
'buying': 1611,
'gucci': 3310,
'bags': 1223,
'sister': 6251,
'things': 6903,
'easy': 2543,
```

```
'uncle': 7200,
'john': 3898,
'bills': 1376,
'sha': 6121,
'huh': 3600,
'lesson': 4138,
'lei': 4129,
'thinkin': 6905,
'sch': 6006,
'earlier': 2531,
'parkin': 5147,
'kent': 3972,
'vale': 7303,
'chance': 1749,
'250': 378,
'wkly': 7626,
'80878': 667,
'www': 7707,
'custcare': 2146,
'08715705022': 130,
'1x150p': 353,
'wk': 7622,
'fuck': 3072,
'family': 2801,
'rhode': 5839,
'island': 3815,
'wherever': 7551,
'leaving': 4119,
'alone': 937,
'bong': 1447,
'maybe': 4453,
'press': 5453,
'buttons': 1607,
'doesn': 2417,
'side': 6218,
'linerental': 4180,
'49': 527,
'month': 4630,
'cross': 2102,
'ntwk': 4900,
'bundle': 1587,
'deals': 2216,
'avble': 1176,
'08001950382': 52,
'mf': 4520,
'tonight': 7030,
'dentist': 2272,
'leh': 4128,
'mayb': 4452,
'bag': 1222,
'goigng': 3216,
'small': 6308,
'jus': 3934,
'except': 2727,
'perfume': 5214,
'smth': 6334,
```

```
'bless': 1408,
'sleep': 6284,
'pray': 5423,
'finish': 2895,
'clas': 1852,
'registered': 5728,
'optin': 5031,
'subscriber': 6622,
'100': 261,
'gift': 3179,
'voucher': 7388,
'receipt': 5683,
'correct': 2038,
'80062': 656,
'whats': 7541,
'no1': 4846,
'bbc': 1281,
'charts': 1767,
'hey': 3471,
'buns': 1589,
'told': 7010,
'adore': 842,
'loverboy': 4285,
'law': 4096,
'meatballs': 4471,
'grins': 3288,
'future': 3098,
'planned': 5300,
'result': 5820,
'best': 1348,
'present': 5448,
'successful': 6630,
'appointment': 1047,
'timin': 6966,
'same': 5960,
'information': 3737,
'02': 7,
'user': 7284,
'find': 2889,
'log': 4234,
'onto': 5004,
'urawinner': 7263,
'fantastic': 2809,
'awaiting': 1187,
'gd': 3141,
'cream': 2088,
'lap': 4071,
'thats': 6876,
...}
```

In [29]:
```
v.get_feature_names_out()[3617]
```

Out[29]:
```
'hussey'
```

In [30]:
```
x_train_np = x_train_cv.toarray()
```

In [31]:
```
x_train_np[0]
```

Out[31]:
```
array([0, 0, 0, ..., 0, 0, 0], shape=(7805,))
```

In [32]:
```
np.where(x_train_np[0] !=0)
```

Out[32]:
```
(array([ 976,  980, 1015, 2405, 2485, 2697, 2703, 2816, 3219, 3652, 3810,
        4401, 4883, 5649, 6303, 6861, 6873, 7545, 7768]),)
```

In [33]:
```
for k in np.where(x_train_np[0] !=0):
    print(v.get_feature_names_out()[k])
```

```
['an' 'and' 'anyone' 'do' 'driving' 'ever' 'everyone' 'faster' 'going'
 'idiot' 'is' 'maniac' 'notice' 're' 'slower' 'than' 'that' 'when' 'you']
```

In [34]:
```
x_train[:4]
```

Out[34]:
```
2965    Do you ever notice that when you're driving, a...
2948                  Leave it. U will always be ignorant.
2810    Oh yeah I forgot. U can only take 2 out shoppi...
4172    Pls what's the full name of joke's school cos ...
Name: Message, dtype: object
```

5. Create the first model using Naive Bayes algorithm

In [35]:
```
from sklearn.naive_bayes import MultinomialNB
```

In [65]:
```
mn = MultinomialNB()
mn.fit(x_train_cv,y_train)
```

Out[65]:
```
MultinomialNB()
```

In [37]:
```
x_test_cv = v.transform(x_test)
```

In [38]:
```
x_test_cv.shape
```

Out[38]:
```
(1115, 7805)
```

5.1 Evaluate the model

In [39]:
```
mn.score(x_train_cv,y_train)
```

Out[39]:
```
0.9923715503702042
```

In [40]:
```
y_pred = mn.predict(x_test_cv)
```

In [41]:
```
mn.score(x_test_cv, y_test)
```

Out[41]:
```
0.9847533632286996
```

In [42]:
```
from sklearn.metrics import confusion_matrix, classification_report
```

In [43]:
```
confusion_matrix(y_pred,y_test)
```

Out[43]:
```
array([[974,  14],
       [  3, 124]])
```

In [44]:
```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      1.00      0.99       977
           1       0.98      0.90      0.94       138

    accuracy                           0.98      1115
   macro avg       0.98      0.95      0.96      1115
weighted avg       0.98      0.98      0.98      1115
```

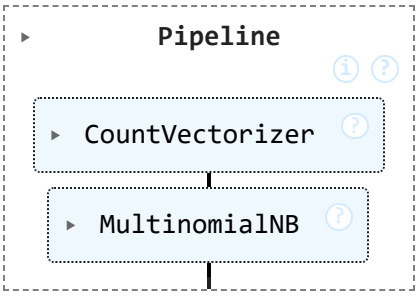
5.2 Creating an NLP Pipe Line

In [45]:
```
from sklearn.pipeline import Pipeline
```

In [46]:
```
model2 = Pipeline(
    [
        ("vectorizer", CountVectorizer()),
        ("MultinomialNB", MultinomialNB())
    ]
)
```

In [47]:
```
model2.fit(x_train,y_train)
```

Out[47]:



In [49]:
```
x_test
```

Out[49]:
```
2038                              Oh sorry please its over
5358    Hmm. Shall i bring a bottle of wine to keep us...
2485    Only if you promise your getting out as SOON a...
249     It didnt work again oh. Ok goodnight then. I.l...
5450                               Sac needs to carry on:)
                              ...
2329        That day you asked about anand number. Why:-)
1132                               Sorry, I'll call later
4607    Oh... Haha... Den we shld had went today too.....
4030               […] anyway, many good evenings to u! s
1137    Dont forget you can place as many FREE Request...
Name: Message, Length: 1115, dtype: object
```

In [50]:
```
pred2 = model2.predict(x_test)
```

In [51]:
```
print(classification_report(y_pred, y_test))
```

```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       988
           1       0.90      0.98      0.94       127

    accuracy                           0.98      1115
   macro avg       0.95      0.98      0.96      1115
weighted avg       0.99      0.98      0.99      1115
```

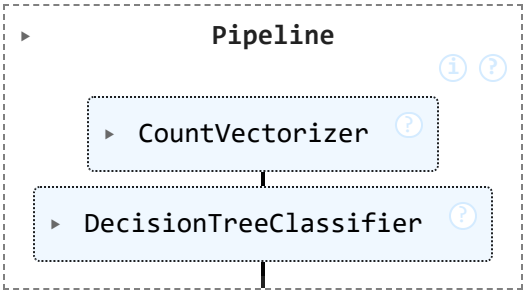
6. Create a model using Decision Tree Classifier

In [57]:
```
from sklearn.tree import DecisionTreeClassifier
```

In [58]:
```
model3 = Pipeline(
    [
        ("vectorizer", CountVectorizer()),
        (" DecisionTreeClassifier",  DecisionTreeClassifier())
    ]
)
```

In [59]:
```
model3.fit(x_train,y_train)
```

Out[59]:



6.1 Evaluate the model

In [60]:
```
pred3 = model3.predict(x_test)
```

In [62]:
```
model3.score(x_train,y_train)
```

Out[62]: 1.0

In [63]:
```
model3.score(x_test, y_test)
```

Out[63]: 0.9650224215246637

In [64]:
```
print(classification_report(pred3, y_test))
```

```
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       984
           1       0.83      0.88      0.86       131

    accuracy                           0.97      1115
   macro avg       0.91      0.93      0.92      1115
weighted avg       0.97      0.97      0.97      1115
```

In [ ]:
```

```