SMS Classifier_Multinomial Naive Bayes

October 6, 2020

0.1 SMS Spam Classifier: Multinomial Naive Bayes

The notebook is divided into the following sections: 1. Importing and preprocessing data 2. Building the model: Multinomial Naive Bayes - Model building - Model evaluation

0.1.1 1. Importing and Preprocessing Data

In [110]:
```
import pandas as pd

# reading the training data
docs = pd.read_table('SMSSpamCollection', header=None, names=['Class', 'sms'])
docs.head()
```

Out[110]:

	Class	sms
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 [111]:
```
# number of SMSes / documents
len(docs)
```

Out[111]:
```
5572
```

In [112]:
```
# counting spam and ham instances
ham_spam = docs.Class.value_counts()
ham_spam
```

Out[112]:
```
ham     4825
spam     747
Name: Class, dtype: int64
```

In [113]:
```
print("spam rate is about {0}%".format(
    round((ham_spam[1]/float(ham_spam[0]+ham_spam[1]))*100), 2))
```

```
spam rate is about 13.0%
```

In [114]:
```
# mapping labels to 0 and 1
docs['label'] = docs.Class.map({'ham':0, 'spam':1})
```

In [115]:
```
docs.head()
```

Out[115]:

	Class	sms	label
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

In [116]:
```
# we can now drop the column 'Class'
docs = docs.drop('Class', axis=1)
docs.head()
```

Out[116]:

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

In [117]:
```
# convert to X and y
X = docs.sms
y = docs.label
print(X.shape)
print(y.shape)
```

```
(5572,)
(5572,)
```

In [118]:
```
# splitting into test and train
from sklearn.model_selection  import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)
```

In [119]:
```
X_train.head()
```

Out[119]:
```
710     4mths half price Orange line rental & latest c...
3740                          Did you stitch his trouser
2711    Hope you enjoyed your new content. text stop t...
3155    Not heard from U4 a while. Call 4 rude chat pr...
3748    Ü neva tell me how i noe... I'm not at home in...
Name: sms, dtype: object
```

In [120]:
```
y_train.head()
```

```
Out[120]: 710     1
          3740    0
          2711    1
          3155    1
          3748    0
          Name: label, dtype: int64

In [121]: # vectorizing the sentences; removing stop words
          from sklearn.feature_extraction.text import CountVectorizer
          vect = CountVectorizer(stop_words='english')

In [122]: vect.fit(X_train)

Out[122]: CountVectorizer(analyzer='word', binary=False, decode_error='strict',
                  dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
                  lowercase=True, max_df=1.0, max_features=None, min_df=1,
                  ngram_range=(1, 1), preprocessor=None, stop_words='english',
                  strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
                  tokenizer=None, vocabulary=None)

In [123]: # printing the vocabulary
          vect.vocabulary_

Out[123]: {'hey': 3198,
           'satisfied': 5504,
           'figuring': 2651,
           'missing': 4217,
           'idea': 3351,
           '2gthr': 384,
           '09099726395': 237,
           'fffffffff': 2635,
           'thm': 6377,
           'nimbomsons': 4466,
           '09050002311': 159,
           'realize': 5216,
           'wake': 6838,
           'lect': 3796,
           'daytime': 2056,
           'polyphonic': 4922,
           'takin': 6231,
           'wudn': 7106,
           'sacrifice': 5469,
           'gin': 2927,
           'min': 4187,
           '13': 278,
           'cw25wx': 2012,
           'title': 6433,
           'westlife': 6950,
           'hopefully': 3262,
```

```
'goodnite': 2979,
'song': 5863,
'tuth': 6581,
'ctargg': 1977,
'eve': 2486,
'sick': 5716,
'pixels': 4859,
'mila': 4180,
'cyclists': 2013,
'sar': 5495,
'w1j6hl': 6817,
'provided': 5089,
'engaged': 2431,
'jules': 3612,
'bootydelious': 1343,
'wheat': 6956,
'btw': 1445,
'07821230901': 37,
'wisdom': 7004,
'britney': 1420,
'weed': 6917,
'gei': 2890,
'txtx': 6601,
'kanagu': 3637,
'unconscious': 6630,
'director': 2199,
'3x': 460,
'makiing': 4044,
'4utxt': 515,
'wishlist': 7011,
'drivin': 2315,
'sonyericsson': 5866,
'2moro': 389,
'snap': 5823,
'09061221066': 183,
'idps': 3359,
'luxury': 3991,
'ache': 764,
'neglect': 4422,
'nigh': 4459,
'opinion': 4613,
'530': 539,
'daddy': 2019,
'handsomes': 3104,
'deliver': 2106,
'onwords': 4601,
'barmed': 1154,
'brother': 1429,
```

```
'otbox': 4646,
'beverage': 1245,
'wasnt': 6872,
'ansr': 927,
'14': 282,
'78': 607,
'pobox1': 4900,
'grumpy': 3041,
'intention': 3459,
'regretted': 5279,
'help08700621170150p': 3181,
'laugh': 3766,
'hopeing': 3263,
'career': 1554,
'flea': 2696,
'height': 3173,
'powerful': 4970,
'tenerife': 6302,
'eighth': 2399,
'arrive': 1016,
'spoken': 5948,
'macleran': 4019,
'das': 2043,
'stable': 5973,
'babyjontet': 1123,
'spk': 5938,
'sumthin': 6141,
'tkts': 6437,
'mis': 4207,
'aom': 948,
'mk17': 4231,
'aroundn': 1010,
'braindance': 1386,
'spree': 5961,
'2mrw': 395,
'accept': 749,
'ibored': 3340,
'erm': 2460,
'saves': 5513,
'sleepy': 5778,
'maturity': 4105,
'coincidence': 1774,
'6pm': 590,
'beggar': 1213,
'clos1': 1740,
'09050003091': 160,
'mobcudb': 4245,
'safe': 5473,
```

```
'features': 2615,
'cares': 1559,
'61610': 563,
'upgrading': 6676,
'gravel': 3015,
'tried': 6537,
'rudi': 5444,
'text82228': 6322,
'books': 1339,
'wherre': 6963,
'logon': 3908,
'permissions': 4789,
'08712400603': 99,
'sweets': 6195,
'hurricanes': 3318,
'attention': 1065,
'weighed': 6925,
'arestaurant': 994,
'07732584351': 25,
'motivate': 4295,
'type': 6603,
'llc': 3888,
'sore': 5874,
'wrench': 7086,
'bedbut': 1204,
'goodies': 2975,
'versus': 6755,
'jacuzzi': 3531,
'salmon': 5484,
'hypertension': 3331,
'm8': 4006,
'pop': 4933,
'occupied': 4552,
'shola': 5677,
'fri': 2795,
'sms': 5817,
'meet': 4130,
'anot': 925,
'nver': 4534,
'unmits': 6657,
'cold': 1777,
'balloon': 1141,
'rushing': 5458,
'gently': 2900,
'spatula': 5907,
'unless': 6654,
'bluff': 1318,
'dusk': 2348,
```

```
'ma': 4009,
'springs': 5963,
'functions': 2833,
'brownies': 1433,
'messaged': 4164,
'shampain': 5637,
'documents': 2244,
'solve': 5846,
'ditto': 2225,
'rcvd': 5198,
'dysentry': 2353,
'leonardo': 3810,
'magazine': 4028,
'market': 4077,
'09058094507': 170,
'worse': 7072,
'singles': 5739,
'07808': 33,
'anythin': 944,
'africa': 824,
'nitro': 4471,
'txttowin': 6600,
'returned': 5365,
'pack': 4682,
'monday': 4269,
'appointment': 971,
'fuuuuck': 2843,
'pongal': 4924,
'propose': 5081,
'loads': 3893,
'blood': 1308,
'stuffing': 6087,
'inever': 3420,
'08712402902': 103,
'oga': 4571,
'fresh': 2793,
'voted': 6806,
'claim': 1718,
'm95': 4008,
'bad': 1128,
'purse': 5119,
'ltdhelpdesk': 3975,
'jontin': 3596,
'tui': 6570,
'1tulsi': 328,
'tight': 6411,
'shinco': 5659,
'sagamu': 5474,
```

```
'coulda': 1910,
'hvae': 3327,
'picking': 4835,
'blessings': 1298,
'chinnu': 1692,
'ranjith': 5185,
'alwys': 891,
'jstfrnd': 3606,
'problems': 5049,
'bbd': 1177,
'wknd': 7025,
'151': 299,
'request': 5326,
'operate': 4611,
'fromm': 2815,
'fuck': 2821,
'crash': 1938,
'shortbreaks': 5685,
'epi': 2454,
'usps': 6710,
'complementary': 1819,
'sit': 5747,
'aren': 992,
'0870241182716': 79,
'matches': 4095,
'level': 3819,
'carpark': 1567,
'silent': 5726,
'happen': 3111,
'perpetual': 4790,
'mns': 4239,
'nmde': 4475,
'appt': 979,
'200': 335,
'inconvenience': 3407,
'inlude': 3435,
'bringing': 1415,
'msging': 4317,
'movietrivia': 4306,
'mails': 4037,
'lighters': 3837,
'description': 2131,
'risk': 5394,
'wenwecan': 6941,
'miwa': 4228,
'aunts': 1078,
'smear': 5802,
'notixiquating': 4512,
```

```
'09064018838': 205,
'massages': 4089,
'careless': 1558,
'08712317606': 97,
'swalpa': 6181,
'offer': 4560,
'evrydy': 2506,
'bailiff': 1135,
'donno': 2268,
'lifting': 3835,
'inclu': 3400,
'flippin': 2703,
'3lp': 449,
'paining': 4691,
'js': 3603,
'gimmi': 2926,
'onam': 4594,
'terry': 6313,
'barkleys': 1153,
'unsubscribe': 6665,
'booty': 1342,
'arcade': 988,
'self': 5581,
'oso': 4645,
'pobox45w2tg150p': 4905,
'sib': 5715,
'transaction': 6515,
'soz': 5896,
'apology': 959,
'greetings': 3025,
'hill': 3208,
'restocked': 5354,
'vouchers': 6808,
'1winaweek': 330,
'mathews': 4100,
'stupid': 6089,
'shoppin': 5680,
'katexxx': 3644,
'die': 2173,
'raji': 5170,
'pansy': 4701,
'pls': 4888,
'evn': 2499,
'loves': 3961,
'gets': 2910,
'theater': 6348,
'bras': 1390,
'returning': 5366,
```

```
'innocent': 3439,
'unbreakable': 6625,
'hp': 3291,
'animation': 917,
'romantic': 5415,
'shitin': 5669,
'community': 1807,
'responsible': 5350,
'abj': 733,
'pl': 4861,
'fucks': 2825,
'kane': 3638,
'cd': 1596,
'fold': 2719,
'eaten': 2370,
'err': 2463,
'plz': 4895,
'wamma': 6851,
'suitemates': 6134,
'smartcall': 5799,
'isn': 3506,
'dificult': 2183,
'ask': 1032,
'chosen': 1705,
'falling': 2578,
'prescription': 5010,
'risks': 5395,
'08714712394': 115,
'lovly': 3965,
'suffer': 6123,
'mornin': 4288,
'hardcore': 3122,
'ponnungale': 4925,
'befor': 1211,
'09099725823': 236,
'squid': 5969,
'mjzgroup': 4230,
'killing': 3678,
'stayed': 6002,
'drunk': 2326,
'rtm': 5440,
'publish': 5103,
'prayers': 4993,
'acl03530150pm': 767,
'restrict': 5355,
'auto': 1082,
'deepest': 2088,
'screaming': 5544,
```

```
'lasting': 3759,
'salad': 5477,
'450p': 487,
's89': 5464,
'efreefone': 2393,
'web': 6905,
'favorite': 2609,
'cliff': 1737,
'thriller': 6391,
'fridays': 2797,
'advise': 811,
'puppy': 5113,
'pubs': 5104,
'hamster': 3097,
'sept': 5603,
'outsider': 4662,
'tiwary': 6435,
'10k': 253,
'lambda': 3737,
'ans': 926,
'08700621170150p': 72,
'arm': 1005,
'mob': 4244,
'lookin': 3923,
'cheaper': 1649,
'ibm': 3338,
'fills': 2655,
'4mths': 509,
'antelope': 934,
'kegger': 3653,
'guilty': 3061,
'njan': 4474,
'hugs': 3305,
'5ish': 544,
'50gbp': 525,
'ring': 5387,
'50rcvd': 531,
'boobs': 1334,
'fast': 2597,
'wt': 7099,
'wire3': 7003,
'england': 2433,
'08717205546': 125,
'mycalls': 4356,
'reallyneed': 5220,
'accessible': 751,
'ham': 3095,
'sucks': 6119,
```

```
'dwn': 2352,
'neighbour': 4424,
'wetherspoons': 6953,
'09050000460': 153,
'buy': 1483,
'm26': 4002,
'flood': 2707,
'fizz': 2687,
'pattern': 4747,
'ideal': 3352,
'cinema': 1713,
'set': 5611,
'changed': 1624,
'crisis': 1960,
'bright': 1410,
'mah': 4031,
'fake': 2572,
'choose': 1700,
'nahi': 4369,
'864233': 663,
'tot': 6496,
'witout': 7017,
'green': 3022,
'ha': 3075,
'12': 266,
'mess': 4162,
'walking': 6844,
'c52': 1496,
'aproach': 982,
'textbuddy': 6324,
'aaniye': 722,
'downon': 2292,
'clocks': 1739,
'hourish': 3281,
'victoria': 6761,
'08714342399': 112,
'events': 2491,
'nyc': 4539,
'shining': 5661,
'promise': 5068,
'doggy': 2254,
'shaping': 5640,
'paul': 4749,
'casualty': 1581,
'xxxx': 7129,
'collect': 1782,
'signal': 5722,
'2optout': 399,
```

```
'anyones': 941,
'phil': 4814,
'help': 3180,
'fone': 2726,
'1pm': 323,
'hitler': 3217,
'shitinnit': 5670,
'09094646631': 233,
'yorge': 7171,
'hppnss': 3293,
'visa': 6787,
'secondary': 5558,
'losers': 3935,
'tram': 6514,
'enemies': 2427,
'monoc': 4276,
'owned': 4677,
'aah': 721,
'teeth': 6284,
'console': 1860,
'customersqueries': 2004,
'44': 478,
'bomb': 1330,
'enjoyed': 2436,
'invitation': 3476,
'checkmate': 1657,
'mon': 4268,
'3510i': 433,
'goto': 2991,
'1x150p': 332,
'dammit': 2023,
'truffles': 6550,
'400': 463,
'break': 1396,
'8p': 695,
'babe': 1119,
'keralacircle': 3658,
'textoperator': 6329,
'arguing': 999,
'wavering': 6889,
'refunded': 5268,
'snappy': 5824,
'order': 4630,
'advising': 812,
'smokes': 5813,
'2stoptxt': 403,
'argue': 998,
'guai': 3046,
```

```
'accident': 753,
'iyo': 3522,
'division': 2227,
'dippeditinadew': 2196,
'lifetime': 3832,
'breaking': 1400,
'gotbabes': 2990,
'5000': 522,
'vibrate': 6758,
'princess': 5034,
'accomodations': 757,
'wc1n3xx': 6895,
'pataistha': 4740,
'listening': 3873,
'forwarded': 2762,
'5min': 546,
'swollen': 6201,
'calculated': 1503,
'87131': 671,
'struggling': 6071,
'batchlor': 1168,
'nagar': 4367,
'di': 2159,
'get4an18th': 2907,
'seing': 5577,
'textcomp': 6325,
'counts': 1915,
'including': 3403,
'sapna': 5494,
'cutest': 2008,
'shracomorsglsuplt': 5702,
'cafe': 1500,
'said': 5475,
'sw73ss': 6180,
'outgoing': 4657,
'blokes': 1305,
'irritated': 3496,
'machan': 4015,
'w4': 6820,
'lou': 3947,
'08714712412': 116,
'1st4terms': 325,
'spiral': 5934,
'londn': 3913,
'chillin': 1686,
'lesser': 3812,
'2morow': 390,
'thanku': 6340,
```

```
'buzy': 1488,
'fuckin': 2823,
'okey': 4581,
'hold': 3236,
'searching': 5553,
'tigress': 6413,
'complaining': 1817,
'okies': 4583,
'gokila': 2960,
'squatting': 5968,
'tons': 6478,
'totally': 6498,
'manage': 4053,
'dear1': 2067,
'ass': 1040,
'maybe': 4111,
'gives': 2934,
'ukp': 6615,
'gek1510': 2891,
'beers': 1210,
'patty': 4748,
'bathroom': 1172,
'judgemental': 3609,
'railway': 5164,
'tongued': 6474,
'prods': 5054,
'happened': 3113,
'ahhh': 845,
'number': 4526,
'avatar': 1088,
'bcum': 1187,
'working': 7062,
'suggest': 6129,
'lacs': 3729,
'saeed': 5472,
'express': 2546,
'axel': 1108,
'neck': 4410,
'fall': 2576,
'til': 6416,
'114': 261,
'mac': 4013,
'eng': 2430,
'path': 4742,
'25': 358,
'wtf': 7101,
'hun': 3311,
'office': 4564,
```

```
'10p': 254,
'limit': 3847,
'craving': 1942,
'avent': 1091,
'enemy': 2428,
'ms': 4314,
'matthew': 4103,
'ex': 2508,
'guide': 3059,
'mk45': 4232,
'maga': 4027,
'sisters': 5746,
'fraction': 2770,
'outage': 4651,
'exterminator': 2550,
'cartoon': 1571,
'vid': 6763,
'texted': 6326,
'prometazine': 5066,
'answered': 929,
'j89': 3526,
'iter': 3514,
'fightng': 2647,
'thesis': 6360,
'physics': 4830,
'goal': 2951,
'08715203694': 121,
'grins': 3027,
'films': 2657,
'150pw': 298,
'33': 430,
'exam': 2511,
'outstanding': 4664,
'bought': 1354,
'anniversary': 920,
'strong': 6068,
'agents': 836,
'hostel': 3274,
'spice': 5928,
'theplace': 6357,
'approve': 975,
'txtstop': 6599,
'lick': 3825,
'89545': 688,
'winning': 6998,
'sherawat': 5652,
'vava': 6742,
'virgil': 6783,
```

```
'cheetos': 1665,
'warned': 6866,
'response': 5347,
'large': 3754,
'0578': 18,
'isnt': 3507,
'83049': 640,
'coca': 1764,
'dined': 2191,
'page': 4685,
'hmph': 3229,
'100': 245,
'skint': 5763,
'whr': 6970,
'hurts': 3323,
'siva': 5753,
'friend': 2800,
'sian': 5714,
'steed': 6012,
'actual': 780,
'praying': 4994,
'tape': 6248,
'eveb': 2487,
'murdered': 4346,
'witot': 7016,
'activate': 776,
'sis': 5744,
'euro2004': 2482,
'intend': 3458,
'canlove': 1536,
'queen': 5139,
'shock': 5674,
'bet': 1239,
'5digital': 542,
'citizen': 1714,
'change': 1623,
'lucy': 3982,
'mall': 4050,
'sickness': 5717,
'sweet': 6191,
'horo': 3267,
'yunny': 7186,
'holder': 3237,
'clue': 1753,
'reaction': 5205,
'noe': 4477,
'desparately': 2138,
'ela': 2402,
```

```
'adress': 801,
'everytime': 2496,
'rpl': 5434,
'mindset': 4192,
'city': 1715,
'gym': 3071,
'140': 283,
'opportunity': 4616,
'floppy': 2709,
'rofl': 5409,
'cps': 1929,
'kg': 3664,
'affection': 818,
'panties': 4703,
'shipping': 5665,
'comes': 1796,
'usb': 6699,
'management': 4056,
'imat': 3377,
'cust': 1999,
'stunning': 6088,
'misscall': 4214,
'fgkslpopw': 2638,
'showers': 5698,
'hiya': 3220,
'frying': 2819,
'read': 5206,
'stamped': 5981,
'1st': 324,
'max': 4106,
'starwars3': 5997,
'limits': 3850,
'efficient': 2392,
'snowboarding': 5833,
'drop': 2319,
'disturbance': 2223,
'royal': 5432,
'zoe': 7198,
'dvg': 2351,
'christ': 1706,
'received': 5239,
'forwarding': 2763,
'pending': 4773,
'maaaan': 4010,
'mapquest': 4068,
'2channel': 375,
'faber': 2558,
'instead': 3453,
```

```
'bmw': 1321,
'itxt': 3517,
'unredeemed': 6661,
'fine': 2667,
'questions': 5144,
'urgent': 6688,
'harish': 3126,
'raksha': 5174,
'elvis': 2410,
'jolt': 3594,
'forum': 2759,
'yellow': 7155,
'renewal': 5307,
'jazz': 3551,
'08718726971': 135,
'muz': 4354,
'w1': 6813,
'tonight': 6475,
'checkup': 1658,
'heehee': 3171,
'dlf': 2234,
'iwas': 3521,
'9758': 709,
'euro': 2481,
'nowadays': 4518,
'participate': 4724,
'called': 1513,
'meat': 4123,
'09065989180': 213,
'wrking': 7092,
'honesty': 3249,
'superb': 6152,
'planned': 4870,
'yer': 7159,
'fringe': 2806,
'transfer': 6517,
'madam': 4024,
'bf': 1248,
'ringtoneking': 5390,
'year': 7150,
'2667': 364,
'winds': 6992,
'smsco': 5818,
'collected': 1783,
'taking': 6232,
'offc': 4559,
'1win150ppmx3': 329,
'bribe': 1408,
```

```
'allah': 881,
'korche': 3708,
'tok': 6458,
'thts': 6399,
'hello': 3178,
'ethnicity': 2480,
'posts': 4959,
'blog': 1301,
'member': 4143,
'lionp': 3863,
'wildlife': 6984,
'wondar': 7042,
'decisions': 2080,
'joy': 3600,
'fgkslpo': 2637,
'woodland': 7050,
'match': 4094,
'subpoly': 6100,
'landlines': 3743,
'speaking': 5909,
'vomitin': 6803,
'freeentry': 2783,
'workage': 7059,
'820554ad0a1705572711': 635,
'sub': 6095,
'tree': 6533,
'ip4': 3484,
'game': 2858,
'weddin': 6911,
'skinny': 5762,
'compass': 1814,
'hotels': 3278,
'happily': 3118,
'realising': 5214,
'scotland': 5536,
'varma': 6736,
'hesitate': 3195,
'bein': 1220,
'jo': 3578,
'asp': 1039,
'toledo': 6461,
'def': 2090,
'rgent': 5376,
'matured': 4104,
'settle': 5614,
'yards': 7143,
'eastenders': 2366,
'mtmsg': 4325,
```

```
'hmmross': 3228,
'registration': 5277,
'flies': 2698,
'swimming': 6197,
'arise': 1003,
'09066380611': 224,
'meg': 4135,
'prabha': 4977,
'birla': 1271,
'buffet': 1455,
'proove': 5077,
'mentionned': 4154,
'pract': 4979,
'pull': 5106,
'blackberry': 1285,
'helpline': 3185,
'dollar': 2261,
'length': 3806,
'trained': 6511,
'ec2a': 2374,
'torch': 6490,
'swatch': 6186,
'unbelievable': 6624,
'lunsford': 3985,
'juz': 3626,
'txting': 6595,
'lolnice': 3911,
'box42wr29c': 1372,
'wounds': 7082,
'khelate': 3665,
'tissco': 6432,
'password': 4737,
'09050000878': 155,
'050703': 17,
'basic': 1162,
'ji': 3573,
'kept': 3656,
'voda': 6796,
'carolina': 1565,
'08709501522': 89,
'happiness': 3119,
'able': 734,
'ovulation': 4673,
've': 6744,
'aco': 769,
'love': 3953,
'fired': 2675,
'randomlly': 5180,
```

```
'night': 4460,
'a21': 719,
'missunderstding': 4219,
'txt': 6590,
'starting': 5995,
'clothes': 1746,
'delete': 2102,
'dehydrated': 2098,
'quality': 5136,
'èn': 7202,
'lautech': 3772,
'reset': 5337,
'rental': 5310,
'truro': 6552,
'calling': 1519,
'little': 3879,
'habit': 3077,
'sambar': 5489,
'cal': 1502,
'supposed': 6161,
'adrian': 802,
'ignorant': 3361,
'loxahatchee': 3968,
'slower': 5792,
'attractive': 1068,
'midnight': 4176,
'spl': 5939,
'09064012103': 201,
'lord': 3933,
'clark': 1722,
'woman': 7038,
'nails': 4370,
'hubby': 3300,
'missin': 4216,
'infernal': 3423,
'09066364589': 222,
'orange': 4626,
'individual': 3418,
'wet': 6952,
'valentine': 6724,
'clover': 1748,
'word': 7056,
'establish': 2478,
'txts': 6597,
'0870': 69,
'irulinae': 3500,
'24': 354,
'130': 279,
```

```
'fromwrk': 2816,
'sf': 5623,
'single': 5738,
'brandy': 1389,
'otside': 4648,
'mind': 4190,
'09090204448': 231,
'lovers': 3960,
'7732584351': 606,
'sweater': 6189,
'arab': 986,
'cm': 1754,
'brolly': 1425,
'janinexx': 3540,
'soft': 5838,
'letters': 3818,
'cds': 1598,
'teacher': 6271,
'delivery': 2109,
'buen': 1453,
'priest': 5030,
'babes': 1120,
'happier': 3116,
'shit': 5668,
'09099726553': 240,
'ar': 985,
'40mph': 470,
'main': 4038,
'appear': 963,
'ba': 1117,
'restaurant': 5352,
'user': 6705,
'natural': 4393,
'08712405020': 106,
'studying': 6083,
'chatting': 1647,
'partner': 4726,
'lower': 3967,
'egg': 2395,
'running': 5454,
'nz': 4543,
'nokia6650': 4485,
'pressies': 5018,
'landline': 3741,
'mi': 4172,
'85': 655,
'chasing': 1641,
'buns': 1465,
```

```
 '08718730666': 141,
 'nigpun': 4463,
 'prefer': 4998,
 'petey': 4805,
 'training': 6512,
 'pobox36504w45wq': 4903,
 'services': 5609,
 'goverment': 2995,
 'eightish': 2400,
 'glasgow': 2938,
 'march': 4072,
 'necesity': 4407,
 'verifying': 6753,
 'social': 5836,
 ...}
```

In [124]:
```
# vocab size
len(vect.vocabulary_.keys())
```

Out[124]: 7204

In [125]:
```
# transforming the train and test datasets
X_train_transformed = vect.transform(X_train)
X_test_transformed = vect.transform(X_test)
```

In [126]:
```
# note that the type is transformed (sparse) matrix
print(type(X_train_transformed))
print(X_train_transformed)
```

```
<class 'scipy.sparse.csr.csr_matrix'>
  (0, 50)        1
  (0, 264)        1
  (0, 509)        1
  (0, 1527)        1
  (0, 1971)        1
  (0, 2780)        2
  (0, 3089)        1
  (0, 3763)        1
  (0, 3852)        1
  (0, 4248)        1
  (0, 4624)        1
  (0, 4626)        1
  (0, 4818)        1
  (0, 4822)        1
  (0, 5027)        1
  (0, 5310)        1
  (0, 6673)        1
  (1, 2169)        1
  (1, 6028)        1
```

```
  (1, 6545)        1
  (2, 98)        1
  (2, 563)        1
  (2, 1867)        1
  (2, 2436)        1
  (2, 3180)        1
  :        :
  (4176, 3879)        1
  (4176, 4417)        1
  (4176, 5229)        1
  (4176, 6191)        1
  (4176, 7134)        1
  (4177, 254)        1
  (4177, 307)        1
  (4177, 358)        1
  (4177, 831)        1
  (4177, 2046)        1
  (4177, 2704)        1
  (4177, 3585)        1
  (4177, 3623)        1
  (4177, 4130)        1
  (4177, 4315)        1
  (4177, 4771)        1
  (4177, 5234)        1
  (4177, 5321)        1
  (4177, 5487)        1
  (4177, 5620)        1
  (4177, 6321)        1
  (4177, 6374)        1
  (4177, 6453)        1
  (4178, 1643)        1
  (4178, 5817)        1
```

0.1.2 2. Building and Evaluating the Model

```
In [127]: # training the NB model and making predictions
          from sklearn.naive_bayes import MultinomialNB
          mnb = MultinomialNB()

          # fit
          mnb.fit(X_train_transformed,y_train)

          # predict class
          y_pred_class = mnb.predict(X_test_transformed)

          # predict probabilities
          y_pred_proba = mnb.predict_proba(X_test_transformed)
```

```
In [143]: # note that alpha=1 is used by default for smoothing
          mnb

Out[143]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

0.1.3 Model Evaluation

```
In [129]: # printing the overall accuracy
          from sklearn import metrics
          metrics.accuracy_score(y_test, y_pred_class)

Out[129]: 0.9877961234745154

In [145]: # confusion matrix
          metrics.confusion_matrix(y_test, y_pred_class)
          # help(metrics.confusion_matrix)

Out[145]: array([[1201,    7],
                 [  10,  175]])

In [131]: confusion = metrics.confusion_matrix(y_test, y_pred_class)
          print(confusion)
          TN = confusion[0, 0]
          FP = confusion[0, 1]
          FN = confusion[1, 0]
          TP = confusion[1, 1]

[[1201    7]
 [  10  175]]


In [132]: sensitivity = TP / float(FN + TP)
          print("sensitivity",sensitivity)

sensitivity 0.9459459459459459


In [133]: specificity = TN / float(TN + FP)
          print("specificity",specificity)

specificity 0.9942052980132451


In [134]: precision = TP / float(TP + FP)
          print("precision",precision)
          print(metrics.precision_score(y_test, y_pred_class))

precision 0.9615384615384616
0.9615384615384616
```

```
In [135]: print("precision",precision)
          print("PRECISION SCORE :",metrics.precision_score(y_test, y_pred_class))
          print("RECALL SCORE :", metrics.recall_score(y_test, y_pred_class))
          print("F1 SCORE :",metrics.f1_score(y_test, y_pred_class))

precision 0.9615384615384616
PRECISION SCORE : 0.9615384615384616
RECALL SCORE : 0.9459459459459459
F1 SCORE : 0.9536784741144414


In [136]: y_pred_class

Out[136]: array([0, 0, 0, ..., 0, 1, 0])

In [137]: y_pred_proba

Out[137]: array([[9.95239557e-01, 4.76044325e-03],
                 [9.99852357e-01, 1.47642544e-04],
                 [9.27878579e-01, 7.21214213e-02],
                 ...,
                 [9.99999671e-01, 3.28799076e-07],
                 [3.72703622e-09, 9.99999996e-01],
                 [9.99999985e-01, 1.46852511e-08]])

In [138]: # creating an ROC curve
          from sklearn.metrics import confusion_matrix as sk_confusion_matrix
          from sklearn.metrics import roc_curve, auc
          import matplotlib.pyplot as plt

          false_positive_rate, true_positive_rate, thresholds = roc_curve(y_test, y_pred_proba
          roc_auc = auc(false_positive_rate, true_positive_rate)

In [139]: # area under the curve
          print (roc_auc)

0.9921872203329157


In [140]: # matrix of thresholds, tpr, fpr
          pd.DataFrame({'Threshold': thresholds,
                        'TPR': true_positive_rate,
                        'FPR':false_positive_rate
                       })

Out[140]:          FPR       TPR     Threshold
          0    0.000000  0.000000  2.000000e+00
          1    0.000000  0.308108  1.000000e+00
          2    0.000000  0.313514  1.000000e+00
```

3	0.000000	0.335135	1.000000e+00
4	0.000000	0.340541	1.000000e+00
5	0.000000	0.351351	1.000000e+00
6	0.000000	0.367568	1.000000e+00
7	0.000000	0.400000	1.000000e+00
8	0.000000	0.410811	1.000000e+00
9	0.000000	0.594595	1.000000e+00
10	0.000000	0.605405	1.000000e+00
11	0.000000	0.616216	1.000000e+00
12	0.000000	0.627027	1.000000e+00
13	0.000000	0.675676	9.999999e-01
14	0.000000	0.686486	9.999995e-01
15	0.000000	0.718919	9.999985e-01
16	0.000000	0.729730	9.999979e-01
17	0.000000	0.945946	8.232595e-01
18	0.001656	0.945946	6.035141e-01
19	0.003311	0.945946	5.574840e-01
20	0.011589	0.945946	2.930076e-01
21	0.011589	0.951351	2.832433e-01
22	0.013245	0.951351	2.567566e-01
23	0.013245	0.967568	2.534251e-01
24	0.022351	0.967568	1.345685e-01
25	0.024007	0.967568	1.345397e-01
26	0.028146	0.967568	1.345108e-01
27	0.038907	0.967568	1.344819e-01
28	0.044702	0.967568	1.015514e-01
29	0.046358	0.967568	9.392548e-02
..	...	...	...
80	0.322020	1.000000	1.812409e-03
81	0.370861	1.000000	7.269237e-04
82	0.372517	1.000000	7.259448e-04
83	0.384934	1.000000	5.958465e-04
84	0.386589	1.000000	5.943177e-04
85	0.437086	1.000000	3.038344e-04
86	0.439570	1.000000	2.982017e-04
87	0.468543	1.000000	1.801308e-04
88	0.470199	1.000000	1.791812e-04
89	0.592715	1.000000	2.597581e-05
90	0.594371	1.000000	2.531364e-05
91	0.644040	1.000000	8.393958e-06
92	0.652318	1.000000	8.206213e-06
93	0.698675	1.000000	3.224466e-06
94	0.700331	1.000000	3.209868e-06
95	0.780629	1.000000	1.884950e-07
96	0.782285	1.000000	1.873539e-07
97	0.802980	1.000000	6.411598e-08
98	0.804636	1.000000	6.357147e-08
99	0.853477	1.000000	6.439797e-09

```
100   0.855132   1.000000   6.354992e-09
101   0.865894   1.000000   4.003403e-09
102   0.867550   1.000000   3.870068e-09
103   0.940397   1.000000   2.323773e-11
104   0.942053   1.000000   2.233382e-11
105   0.959437   1.000000   6.325596e-13
106   0.961093   1.000000   6.086604e-13
107   0.970199   1.000000   4.174215e-14
108   0.972682   1.000000   3.968331e-14
109   1.000000   1.000000   6.227131e-41

[110 rows x 3 columns]
```

In [141]:
```
# plotting the ROC curve
%matplotlib inline
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.title('ROC')
plt.plot(false_positive_rate, true_positive_rate)
```

Out[141]: [<matplotlib.lines.Line2D at 0x1a1d8a69b0>]