SMS Classifier_Bernoulli NB

October 6, 2020

0.1 SMS Spam Classifier: Bernoulli Naive Bayes

The notebook is divided into the following sections: 1. Importing and preprocessing data 2. Building the model: Bernoulli Naive Bayes

```
In [2]: import pandas as pd
        docs = pd.read_table('SMSSpamCollection', header=None, names=['Class', 'sms'])
        docs.head()
```

```
Out[2]:   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 [3]: #df.column_name.value_counts() - gives no. of unique inputs in that columns
        docs.Class.value_counts()
```

```
Out[3]: ham     4825
        spam     747
        Name: Class, dtype: int64
```

```
In [4]: ham_spam=docs.Class.value_counts()
        ham_spam
```

```
Out[4]: ham     4825
        spam     747
        Name: Class, dtype: int64
```

```
In [5]: print("Spam % is ",(ham_spam[1]/float(ham_spam[0]+ham_spam[1]))*100)
```

```
Spam % is  13.406317300789663
```

```
In [6]: # mapping labels to 1 and 0
        docs['label'] = docs.Class.map({'ham':0, 'spam':1})
```

```
In [7]: docs.head()
```

```
Out[7]:   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 [8]: X=docs.sms
        y=docs.label

In [9]: X = docs.sms
        y = docs.label
        print(X.shape)
        print(y.shape)

(5572,)
(5572,)


In [10]: # 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 [11]: X_train.head()

Out[11]: 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 [12]: from sklearn.feature_extraction.text import CountVectorizer

         # vectorising the text
         vect = CountVectorizer(stop_words='english')

In [13]: vect.fit(X_train)

Out[13]: 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 [14]: vect.vocabulary_
```

```
Out[14]: {'shocking': 5675,
          'moan': 4243,
          'immed': 3381,
          'youwanna': 7177,
          'collected': 1783,
          'steed': 6012,
          '150p16': 294,
          'current': 1996,
          'female': 2629,
          'eightish': 2400,
          'das': 2043,
          '08719181503': 145,
          '8wp': 698,
          'soon': 5868,
          'mad1': 4022,
          'effects': 2391,
          'sometext': 5853,
          'woah': 7034,
          'manage': 4053,
          'obedient': 4546,
          'caught': 1587,
          'edu': 2383,
          'proper': 5078,
          'o2': 4544,
          'prescripiton': 5009,
          'lovly': 3965,
          'lakhs': 3736,
          'deluxe': 2111,
          'wamma': 6851,
          'suite342': 6133,
          'excellent': 2513,
          'ud': 7203,
          'kaiez': 3631,
          'tell': 6289,
          'puttin': 5125,
          'refund': 5267,
          'tirunelvali': 6429,
          '08701417012': 74,
          'ipaditan': 3486,
          '69988': 584,
          'm95': 4008,
          'blow': 1311,
          'takes': 6230,
          'company': 1812,
          'deltomorrow': 2110,
          'searching': 5553,
          'videosound': 6768,
          'cried': 1959,
```

```
'building': 1458,
'chaps': 1629,
'conveying': 1880,
'tmr': 6443,
'fault': 2605,
'jersey': 3566,
'449050000301': 485,
'cupboard': 1993,
'28days': 369,
'sponsors': 5949,
'78': 607,
'logo': 3907,
'tank': 6245,
'senor': 5594,
'aww': 1105,
'bras': 1390,
'truro': 6552,
'lux': 3990,
'magazine': 4028,
'ocean': 4556,
'euro': 2481,
'uterus': 6713,
'88888': 683,
'phne': 4816,
'weak': 6896,
'creepy': 1954,
'search': 5552,
'prasad': 4989,
'urawinner': 6685,
'wc1n3xx': 6895,
'favourite': 2611,
'apart': 949,
'im': 3373,
'eh': 2398,
'flag': 2688,
'transfred': 6520,
'stayin': 6003,
'afghanistan': 822,
'ajith': 861,
'slo': 5787,
'daaaaa': 2016,
'gimme': 2925,
'numbers': 4527,
'laptop': 3750,
'evrydy': 2506,
'hanging': 3108,
'reassurance': 5227,
'sacked': 5468,
```

```
'conference': 1840,
'million': 4184,
'trade': 6507,
'lacking': 3728,
'oso': 4645,
'financial': 2665,
'daywith': 2057,
'totes': 6499,
'title': 6433,
'08718727868': 137,
'http': 3298,
'borrow': 1348,
'sleepingwith': 5775,
'bits': 1281,
'anyplaces': 942,
'rearrange': 5223,
'fills': 2655,
'114': 261,
'quit': 5148,
'town': 6504,
'mentor': 4155,
'directly': 2198,
'prashanthettan': 4990,
'travel': 6524,
'crashing': 1940,
'lou': 3947,
'x29': 7113,
'transport': 6521,
'hip': 3212,
'xavier': 7117,
'floor': 2708,
'speeding': 5920,
'boundaries': 1355,
'fixd': 2683,
'qet': 5131,
'parking': 4723,
'lambda': 3737,
'noise': 4480,
'11pm': 265,
'donno': 2268,
'bathroom': 1172,
'height': 3173,
'portal': 4940,
'art': 1019,
'08006344447': 56,
'morphine': 4290,
'spark': 5905,
'habit': 3077,
```

```
'hearing': 3161,
'planet': 4869,
'life': 3830,
'plane': 4868,
'3aj': 442,
'coco': 1769,
'idc': 3350,
'heads': 3153,
'facts': 2564,
'nething': 4432,
'career': 1554,
'movies': 4305,
'33': 430,
'09050000928': 156,
'luv': 3988,
'positions': 4945,
'rcvd': 5198,
'alto18': 889,
'giv': 2933,
'rebooting': 5231,
'toshiba': 6494,
'hungry': 3314,
't91': 6213,
'fab': 2557,
'nike': 4464,
'remain': 5293,
'blog': 1301,
'nearly': 4406,
'yetunde': 7163,
'sh': 5625,
'ias': 3334,
'july': 3615,
'66': 569,
'arts': 1020,
'pattern': 4747,
'efreefone': 2393,
'online': 4599,
'hospitals': 3272,
'yeovil': 7157,
'gold': 2961,
'08702840625': 80,
'noun': 4515,
'british': 1419,
'followin': 2722,
'rights': 5386,
'nursery': 4528,
'august': 1075,
'09050000460': 153,
```

```
'chat': 1643,
'stifled': 6025,
'kano': 3640,
'evone': 2502,
'gm': 2943,
'miles': 4181,
'jstfrnd': 3606,
'09061221066': 183,
'dip': 2195,
'maaaan': 4010,
'comuk': 1832,
'smells': 5804,
'able': 734,
'contention': 1869,
'ride': 5382,
'dramastorm': 2298,
'causing': 1589,
'hides': 3205,
'mis': 4207,
'cocksuckers': 1768,
'grr': 3039,
'opinions': 4614,
'1843': 309,
'3650': 434,
'08714342399': 112,
'cheese': 1664,
'londn': 3913,
'alot': 886,
'shoot': 5678,
'hanger': 3106,
'po': 4897,
'spoons': 5953,
'throwing': 6394,
'86688': 664,
'syria': 6209,
'archive': 989,
'kickboxing': 3668,
'milk': 4182,
'stitch': 6028,
'whispers': 6965,
'adults': 807,
'keeping': 3651,
'countinlots': 1913,
'kiefer': 3675,
'kids': 3673,
'530': 539,
'motive': 4297,
'quarter': 5137,
```

```
'units': 6648,
'themob': 6353,
'lst': 3973,
'drinks': 2312,
'yalrigu': 7137,
'rpl': 5434,
'topped': 6487,
'academic': 745,
'09065171142': 210,
'entered': 2441,
'uh': 6612,
'130': 279,
'330': 431,
'recd': 5234,
'scream': 5542,
'gopalettan': 2984,
'seat': 5555,
'ko': 3705,
'grandmas': 3009,
'everyday': 2493,
'tai': 6224,
'happens': 3115,
'careers': 1555,
'm221bp': 4000,
'summon': 6140,
'driver': 2314,
'dnt': 2236,
'gdeve': 2885,
'stapati': 5987,
'quoting': 5155,
'audrie': 1074,
'moving': 4307,
'____': 718,
'typical': 6605,
'treats': 6531,
'8000930705': 618,
'09061104283': 179,
'freak': 2776,
'weaknesses': 6898,
'burger': 1466,
'angry': 915,
'surprised': 6172,
'regarding': 5273,
'bay': 1174,
'exhibition': 2524,
'anymore': 940,
'glass': 2939,
'waited': 6835,
```

```
'instituitions': 3454,
'andrews': 914,
'charts': 1639,
'teenager': 6283,
'del': 2099,
'moms': 4267,
'weekends': 6921,
'guilty': 3061,
'spending': 5925,
'nuclear': 4524,
'admin': 793,
'ben': 1231,
'troubleshooting': 6544,
'shaking': 5634,
'wasn': 6871,
'received': 5239,
'radio': 5160,
'fran': 2771,
'jacket': 3528,
'previous': 5025,
'vijay': 6771,
'remembered': 5296,
'clock': 1738,
'bennys': 1235,
'matrix3': 4101,
'malarky': 4048,
'pookie': 4928,
'40mph': 470,
'conducts': 1838,
'disconnect': 2210,
'playing': 4878,
'postponed': 4958,
'dane': 2029,
'wan': 6852,
'cuddled': 1985,
'gigolo': 2924,
'bilo': 1265,
'barely': 1151,
'yesterday': 7162,
'chosen': 1705,
'reslove': 5340,
'09066382422': 225,
'lucy': 3982,
'biz': 1282,
'june': 3618,
'write': 7089,
'ese': 2472,
'sugar': 6128,
```

```
'chinatown': 1688,
'gbp5': 2883,
'jeri': 3563,
'slots': 5789,
'boost': 1341,
'device': 2151,
'womdarfull': 7039,
'foot': 2733,
'definite': 2093,
'shexy': 5654,
'ago': 840,
'elama': 2404,
'av': 1084,
'95': 707,
'listn': 3875,
'ure': 6686,
'bthere': 1444,
'secrets': 5563,
'fassyole': 2596,
'plyr': 4894,
'murdered': 4346,
'cola': 1776,
'wed': 6910,
'wish': 7006,
'kavalan': 3646,
'danger': 2031,
'ready': 5209,
'bob': 1325,
'or2stoptxt': 4624,
'ipad': 3485,
'69866': 579,
'thesis': 6360,
'sexychat': 5621,
'issues': 3509,
'topic': 6485,
'08002888812': 52,
'minnaminunginte': 4198,
'developer': 2150,
'spinout': 5933,
'apps': 978,
'qatar': 5129,
'thousands': 6386,
'hallaq': 3093,
'ammae': 900,
'removed': 5306,
'med': 4125,
'care': 1551,
'zealand': 7192,
```

```
'dun': 2343,
'shah': 5629,
'doinat': 2258,
'chick': 1675,
'dartboard': 2042,
'bevies': 1246,
'sarcasm': 5497,
'occurs': 4555,
'ofstuff': 4570,
'original': 4639,
'kalstiya': 3635,
'eldest': 2406,
'lunsford': 3985,
'week': 6918,
'present': 5012,
'checkup': 1658,
'erupt': 2468,
'specialise': 5913,
'84025': 650,
'nitz': 4473,
'rcv': 5197,
'married': 4082,
'anythingtomorrow': 945,
'08': 44,
'chess': 1670,
'dream': 2304,
'ned': 4411,
'naked': 4371,
'6months': 588,
'revealed': 5369,
'69911': 582,
'tuesday': 6569,
'terrific': 6310,
'moved': 4302,
'zed': 7194,
'09064012160': 202,
'sane': 5490,
'vegetables': 6746,
'ground': 3033,
'31p': 426,
'waheed': 6830,
'payed': 4755,
'battery': 1173,
'expiry': 2539,
'peril': 4784,
'george': 2904,
'box420': 1371,
'moments': 4266,
```

```
'hmm': 3225,
'tht': 6398,
'wot': 7076,
'figures': 2650,
'appropriate': 974,
'scotsman': 5537,
'tbs': 6265,
'ahhhh': 846,
'bro': 1421,
'bowl': 1358,
'116': 262,
'juicy': 3610,
'elaya': 2405,
'darlin': 2038,
'gettin': 2913,
'rinu': 5392,
'02073162414': 10,
'unintentionally': 6645,
'dinner': 2193,
'phone750': 4819,
'score': 5532,
'lionm': 3862,
'09066361921': 218,
'wuldnt': 7108,
'polyphonic': 4922,
'waht': 6833,
'bk': 1283,
'eastenders': 2366,
'ors': 4641,
'decimal': 2078,
'dang': 2030,
'revision': 5372,
'pink': 4851,
'yellow': 7155,
'slowly': 5793,
'eurodisinc': 2483,
'7oz': 613,
'861': 662,
'sofa': 5837,
'bluetooth': 1316,
'forms': 2757,
'sometme': 5855,
'81303': 633,
'cuddle': 1984,
'sugababes': 6126,
'sir': 5742,
'vivek': 6794,
'corrct': 1896,
```

```
'exciting': 2515,
'alwys': 891,
'selected': 5579,
'gailxx': 2850,
'applebees': 966,
'netvision': 4434,
'word': 7056,
'hasbro': 3129,
'stop2stop': 6040,
'finishing': 2673,
'82468': 639,
'aftr': 829,
'april': 981,
'tgxxrz': 6332,
'cancel': 1532,
'event': 2490,
'companion': 1811,
'silent': 5726,
'base': 1160,
'garden': 2866,
'deal': 2062,
'favour': 2610,
'helpline': 3185,
'manual': 4066,
'lv': 3992,
'reason': 5224,
'8lb': 694,
'types': 6604,
'fones': 2728,
'springs': 5963,
'900': 699,
'bluray': 1320,
'programs': 5062,
'snap': 5823,
'mmmmm': 4236,
'merely': 4158,
'pete': 4803,
'2wks': 407,
'3mins': 451,
'expecting': 2531,
'safe': 5473,
'addie': 788,
'muchxxlove': 4332,
'vasai': 6739,
'genuine': 2901,
'social': 5836,
'driving': 2316,
'fishhead': 2680,
```

```
'bye': 1495,
'language': 3747,
'box403': 1370,
'wherevr': 6962,
'degrees': 2097,
'erutupalam': 2469,
'rentl': 5312,
'68866': 573,
'notified': 4511,
'alert': 870,
'fake': 2572,
'pack': 4682,
'becausethey': 1200,
'lennon': 3808,
'1win150ppmx3': 329,
'course': 1920,
'network': 4435,
'appointments': 972,
'agency': 834,
'nigpun': 4463,
'hooch': 3255,
'thats': 6346,
'section': 5565,
'reward': 5373,
'txt250': 6591,
'bulbs': 1460,
'pushes': 5122,
'thinked': 6369,
'nevering': 4439,
'verified': 6751,
'england': 2433,
'singing': 5737,
'flyng': 2716,
'deliver': 2106,
'changed': 1624,
'flower': 2711,
'6031': 558,
'scarcasim': 5521,
'dual': 2332,
'pobox12n146tf150p': 4901,
'various': 6735,
'08000839402': 48,
'750': 600,
'mobno': 4251,
'tb': 6264,
'instead': 3453,
'settings': 5613,
'major': 4041,
```

```
'450p': 487,
'0a': 243,
'amma': 899,
'11mths': 264,
'gayle': 2878,
'designation': 2134,
'brum': 1437,
'water': 6882,
'followed': 2721,
'blake': 1288,
'n9dx': 4362,
'consensus': 1854,
'fb': 2612,
'madam': 4024,
'bbq': 1179,
'cos': 1901,
'westonzoyland': 6951,
'falling': 2578,
'monthlysubscription': 4281,
'haughaighgtujhyguj': 3135,
'upload': 6678,
'wright': 7088,
'blood': 1308,
'sp': 5898,
'couldn': 1911,
'tie': 6408,
'maniac': 4064,
'shy': 5712,
'ls278bb': 3972,
'mila': 4180,
'album': 867,
'41782': 472,
'paying': 4757,
'private': 5040,
'urgoin': 6692,
'1st': 324,
'environment': 2453,
'invest': 3474,
'team': 6274,
'dealing': 2064,
'enter': 2440,
'yay': 7147,
'purpose': 5118,
'180': 308,
'childish': 1681,
'takin': 6231,
'spk': 5938,
'lasagna': 3757,
```

```
'4403ldnw1a7rw18': 480,
'heart': 3162,
'wrench': 7086,
'area': 991,
'yavnt': 7144,
'lodge': 3901,
'river': 5398,
'carly': 1564,
'09090204448': 231,
'spanish': 5902,
'chatter': 1646,
'uwana': 6717,
'weakness': 6897,
'account': 760,
'cock': 1767,
'ru': 5441,
'wins': 6999,
'chip': 1693,
'requests': 5327,
'08712405020': 106,
'lonely': 3916,
'dvd': 2350,
'weiyi': 6932,
'filth': 2658,
'common': 1806,
'wah': 6828,
'asking': 1036,
'daytime': 2056,
'says': 5518,
'karaoke': 3641,
'provided': 5089,
'newquay': 4446,
'scary': 5523,
'steam': 6010,
'accommodation': 755,
'80878': 630,
'weirdy': 6931,
'oredi': 4632,
'marry': 4083,
'yarasu': 7142,
'xam': 7116,
'bedrm': 1205,
'post': 4952,
'know': 3700,
'stereo': 6016,
'tampa': 6244,
'penis': 4774,
'breakin': 1399,
```

```
'toclaim': 6452,
'cops': 1892,
'lab': 3726,
'dwn': 2352,
'code': 1770,
'desert': 2132,
'floating': 2706,
'admission': 796,
'txtx': 6601,
'08700469649': 71,
'pre': 4995,
'jo': 3578,
'confirmd': 1844,
'087123002209am': 96,
'sake': 5476,
'drinkin': 2310,
'scenery': 5524,
'easy': 2368,
'2day': 376,
'outfor': 4656,
'soz': 5896,
'cer': 1609,
'm8': 4006,
'booty': 1342,
'friendship': 2803,
'girls': 2931,
'chillin': 1686,
'box334sk38ch': 1367,
'celebrate': 1602,
'ahmad': 847,
'karnan': 3642,
'successful': 6115,
'urself': 6697,
'parchi': 4714,
'cup': 1992,
'buyer': 1484,
'07': 20,
'mt': 4321,
'checking': 1656,
'rstm': 5438,
'pg': 4807,
'amk': 897,
'todays': 6454,
'ho': 3232,
'paragon': 4711,
'eshxxxxxxxxxx': 2473,
'fucked': 2822,
'drvgsto': 2329,
```

```
'shrub': 5705,
'geeee': 2888,
'msg': 4315,
'history': 3215,
'go2sri': 2950,
'sneham': 5827,
'excuses': 2518,
'dancin': 2027,
'hook': 3257,
'wake': 6838,
'centre': 1608,
'probthat': 5052,
'ones': 4596,
'question': 5142,
'members': 4144,
'miracle': 4205,
'rs': 5437,
'grow': 3035,
'jobyet': 3581,
'09050000555': 154,
'jen': 3559,
'shhhhh': 5655,
'virtual': 6786,
'tons': 6478,
'premier': 5001,
'hat': 3132,
'investigate': 3475,
'expires': 2538,
'subscrition': 6111,
'inshah': 3446,
'shirt': 5666,
'jackson': 3530,
'wlcome': 7027,
'talking': 6238,
'realising': 5214,
'stealing': 6009,
'waste': 6873,
'shipped': 5664,
'minmobsmorelkpobox177hp51fl': 4196,
'mths': 4324,
'cld': 1727,
'lifeis': 3831,
'ummifying': 6620,
'badass': 1129,
'picking': 4835,
'explosive': 2543,
'gving': 3070,
'watchin': 6879,
```

```
'swing': 6199,
'weird': 6928,
'cps': 1929,
'pushbutton': 5121,
'screaming': 5544,
'muchand': 4331,
'uni': 6641,
'lar': 3751,
'wiskey': 7012,
'jos': 3598,
'dedicated': 2084,
'useful': 6703,
'cme': 1756,
'urmom': 6695,
'stylist': 6094,
'site': 5748,
'begun': 1217,
'openings': 4610,
'xxx': 7127,
'plan': 4867,
'jeevithathile': 3558,
'kisi': 3688,
'brownie': 1432,
'goodnite': 2979,
'game': 2858,
'chores': 1704,
'shinco': 5659,
'4info': 507,
'imp': 3384,
'kanagu': 3637,
'lifting': 3835,
'formatting': 2756,
'12mths': 277,
'willpower': 6986,
'hp20': 3292,
'paul': 4749,
'ending': 2424,
'somerset': 5852,
'ends': 2426,
'messenger': 4168,
'plum': 4889,
'tablets': 6217,
'physics': 4830,
'taylor': 6262,
'gn': 2945,
'profiles': 5059,
'urgnt': 6691,
'desires': 2135,
```

```
'spend': 5924,
'pix': 4858,
'xin': 7120,
'thriller': 6391,
'lyf': 3994,
'small': 5796,
'swalpa': 6181,
'collapsed': 1780,
'friday': 2796,
'drpd': 2322,
'gentleman': 2899,
'07734396839': 26,
'bian': 1254,
'savamob': 5510,
'youphone': 7174,
'decided': 2076,
'echo': 2375,
'abeg': 727,
'88066': 676,
'b4280703': 1114,
'flavour': 2695,
'perspective': 4799,
'buddy': 1450,
'deliveredtomorrow': 2108,
'videochat': 6765,
'6th': 591,
'hack': 3078,
'arguments': 1001,
'living': 3885,
'happening': 3114,
'nvm': 4535,
'pases': 4730,
'home': 3244,
'surf': 6167,
'sk3': 5756,
'knocking': 3699,
'mth': 4323,
'sender': 5591,
'rules': 5449,
'iyo': 3522,
'en': 2421,
'yeh': 7153,
'09058094599': 175,
'07946746291': 40,
'4xx26': 518,
'toilet': 6457,
'zhong': 7196,
'lined': 3854,
```

```
'apples': 967,
'lips': 3867,
'bout': 1356,
'changes': 1625,
'testing': 6317,
'0844': 62,
'bx526': 1494,
'watchng': 6881,
'loud': 3948,
'hesitant': 3194,
'continue': 1871,
'croydon': 1965,
'shampain': 5637,
'rolled': 5413,
'reasonable': 5225,
'roger': 5410,
'rtf': 5439,
'shining': 5661,
'up4': 6669,
'gastroenteritis': 2871,
'possible': 4951,
'escape': 2471,
'usps': 6710,
'bcmsfwc1n3xx': 1185,
'pressies': 5018,
'4txt': 513,
'16': 302,
'california': 1507,
'yards': 7143,
'ouch': 4650,
'inperialmusic': 3441,
'clash': 1723,
'morning': 4289,
'mushy': 4349,
'voda': 6796,
'l8r': 3720,
'checked': 1655,
'duvet': 2349,
'wc1n': 6894,
'jenxxx': 3561,
'randomly': 5181,
'ultimately': 6618,
'stressful': 6058,
'30': 412,
'finding': 2666,
'vid': 6763,
'ke': 3650,
'reception': 5243,
```

```
'82242': 636,
'yah': 7135,
'stomach': 6033,
'names': 4376,
'sweetheart': 6193,
'stock': 6029,
'badly': 1130,
'interfued': 3462,
'taste': 6251,
'ppm150': 4975,
'adsense': 805,
'09058099801': 177,
'fromwrk': 2816,
'tattoos': 6256,
'sed': 5569,
'exorcist': 2528,
'disappeared': 2205,
'yogasana': 7169,
'cleaning': 1729,
'queen': 5139,
'ertini': 2466,
'24th': 357,
'otside': 4648,
'progress': 5063,
'terrorist': 6312,
'pocketbabe': 4909,
'okors': 4584,
'custcare': 2000,
'cool': 1888,
'incredible': 3410,
'acid': 766,
'reaching': 5204,
'sister': 5745,
'greetings': 3025,
'blackberry': 1285,
'philosophy': 4815,
'69698': 577,
'gets': 2910,
'fortune': 2758,
'save': 5511,
'tones2u': 6472,
'skip': 5764,
'reservations': 5334,
'rgds': 5375,
'simple': 5731,
'september': 5604,
'5free': 543,
'ask': 1032,
```

```
           'grocers': 3029,
           'ovr': 4671,
           'hm': 3223,
           'divert': 2226,
           'worlds': 7066,
           'fumbling': 2830,
           'blokes': 1305,
           'employer': 2420,
           'unable': 6623,
           'santacalling': 5492,
           'member': 4143,
           'foned': 2727,
           'manageable': 4054,
           'early': 2359,
           'pics': 4837,
           'tenerife': 6302,
           'lkpobox177hp51fl': 3886,
           '02072069400': 9,
           'exorcism': 2527,
           'outdoors': 4654,
           'fix': 2682,
           'refilled': 5264,
           'diet': 2175,
           'bit': 1277,
           '63miles': 566,
           'suppliers': 6155,
           'fink': 2674,
           'paid': 4688,
           'outta': 4665,
           'sale': 5480,
           '7ws': 616,
           'matthew': 4103,
           'yr': 7180,
           'ibuprofens': 3341,
           'jacuzzi': 3531,
           'premarica': 5000,
           'ph': 4808,
           'patent': 4741,
           'packs': 4684,
           'nature': 4394,
           ...}

In [15]: vect.get_feature_names()

Out[15]: ['00',
          '000',
          '008704050406',
          '0121',
```

```
'01223585236',
'01223585334',
'0125698789',
'02',
'0207',
'02072069400',
'02073162414',
'02085076972',
'021',
'03',
'04',
'0430',
'05',
'050703',
'0578',
'06',
'07',
'07008009200',
'07090201529',
'07090298926',
'07123456789',
'07732584351',
'07734396839',
'07742676969',
'0776xxxxxxx',
'07781482378',
'07786200117',
'078',
'07801543489',
'07808',
'07808247860',
'07808726822',
'07815296484',
'07821230901',
'07880867867',
'0789xxxxxxx',
'07946746291',
'0796xxxxxx',
'07973788240',
'07xxxxxxxxx',
'08',
'0800',
'08000407165',
'08000776320',
'08000839402',
'08000930705',
'08000938767',
'08001950382',
```

```
'08002888812',
'08002986030',
'08002986906',
'08002988890',
'08006344447',
'0808',
'08081263000',
'08081560665',
'0825',
'083',
'0844',
'08448714184',
'0845',
'08450542832',
'08452810071',
'08452810073',
'08452810075over18',
'0870',
'08700435505150p',
'08700469649',
'08700621170150p',
'08701213186',
'08701417012',
'08701417012150p',
'0870141701216',
'087016248',
'08701752560',
'0870241182716',
'08702840625',
'08704050406',
'08704439680',
'08706091795',
'0870737910216yrs',
'08707509020',
'08707808226',
'08708034412',
'08709222922',
'08709501522',
'0871',
'087104711148',
'08712101358',
'08712103738',
'0871212025016',
'08712300220',
'087123002209am',
'08712317606',
'08712400602450p',
'08712400603',
```

```
'08712402050',
'08712402578',
'08712402779',
'08712402902',
'08712402972',
'08712404000',
'08712405020',
'08712405022',
'08712460324',
'0871277810710p',
'0871277810810',
'0871277810910p',
'08714342399',
'08714712379',
'08714712388',
'08714712394',
'08714712412',
'08715203028',
'08715203649',
'08715203652',
'08715203685',
'08715203694',
'08715500022',
'08715705022',
'08717168528',
'08717205546',
'0871750',
'08717898035',
'08718711108',
'08718720201',
'08718723815',
'08718725756',
'08718726270',
'087187262701',
'08718726970',
'08718726971',
'08718726978',
'08718727868',
'08718727870',
'087187278701501ppm',
'08718730555',
'08718730666',
'08718738001',
'08718738002',
'08719180248',
'08719181503',
'08719181513',
'08719899217',
```

```
'08719899229',
'08719899230',
'09',
'09050000301',
'09050000332',
'09050000460',
'09050000555',
'09050000878',
'09050000928',
'09050001295',
'09050001808',
'09050002311',
'09050003091',
'09050090044',
'09050280520',
'09053750005',
'09056242159',
'09057039994',
'09058091854',
'09058091870',
'09058094454',
'09058094455',
'09058094507',
'09058094565',
'09058094583',
'09058094594',
'09058094597',
'09058094599',
'09058098002',
'09058099801',
'09061104276',
'09061104283',
'09061209465',
'09061213237',
'09061221061',
'09061221066',
'09061701444',
'09061701461',
'09061701939',
'09061702893',
'09061743386',
'09061743806',
'09061743810',
'09061743811',
'09061744553',
'09061749602',
'09061790121',
'09061790125',
```

```
'09061790126',
'09063440451',
'09063458130',
'0906346330',
'09064011000',
'09064012103',
'09064012160',
'09064015307',
'09064017295',
'09064018838',
'09064019014',
'09064019788',
'09065069120',
'09065069154',
'09065171142',
'09065174042',
'09065394514',
'09065989180',
'09065989182',
'09066350750',
'09066358152',
'09066358361',
'09066361921',
'09066362231',
'09066364311',
'09066364349',
'09066364589',
'09066368470',
'09066380611',
'09066382422',
'09066612661',
'09066649731from',
'09066660100',
'09071512433',
'09077818151',
'09090204448',
'09094100151',
'09094646631',
'09095350301',
'09096102316',
'09099725823',
'09099726395',
'09099726429',
'09099726481',
'09099726553',
'09111032124',
'09701213186',
'0a',
```

```
'10',
'100',
'1000',
'1000s',
'100p',
'100percent',
'1013',
'1030',
'10am',
'10k',
'10p',
'10ppm',
'10th',
'11',
'1120',
'113',
'1131',
'114',
'116',
'118p',
'11mths',
'11pm',
'12',
'1205',
'120p',
'121',
'1225',
'123',
'125',
'1250',
'125gift',
'128',
'12hrs',
'12mths',
'13',
'130',
'1327',
'139',
'14',
'140',
'1405',
'140ppm',
'145',
'1450',
'146tf150p',
'14thmarch',
'15',
'150',
```

```
'1500',
'150p',
'150p16',
'150pm',
'150ppermesssubscription',
'150ppm',
'150pw',
'151',
'153',
'15pm',
'16',
'165',
'1680',
'169',
'177',
'18',
'180',
'1843',
'18p',
'18yrs',
'195',
'1956669',
'1apple',
'1b6a5ecef91ff9',
'1cup',
'1er',
'1hr',
'1im',
'1lemon',
'1mega',
'1million',
'1pm',
'1st',
'1st4terms',
'1stchoice',
'1thing',
'1tulsi',
'1win150ppmx3',
'1winaweek',
'1winawk',
'1x150p',
'1yf',
'20',
'200',
'2000',
'2003',
'2004',
'2005',
```

```
'2006',
'2007',
'2025050',
'20m12aq',
'20p',
'21',
'21870000',
'21st',
'22',
'220',
'220cm2',
'2309',
'23f',
'23g',
'24',
'24hrs',
'24m',
'24th',
'25',
'250',
'250k',
'255',
'25p',
'26',
'2667',
'26th',
'27',
'28',
'2814032',
'28days',
'28th',
'28thfeb',
'29',
'2b',
'2c',
'2channel',
'2day',
'2docd',
'2end',
'2ez',
'2find',
'2getha',
'2geva',
'2go',
'2gthr',
'2hrs',
'2kbsubject',
'2lands',
```

```
'2marrow',
'2moro',
'2morow',
'2morro',
'2morrow',
'2morrowxxxx',
'2mro',
'2mrw',
'2mwen',
'2nd',
'2nite',
'2optout',
'2px',
'2rcv',
'2stop',
'2stoptxt',
'2u',
'2watershd',
'2waxsto',
'2wks',
'2wt',
'2wu',
'2yr',
'2yrs',
'30',
'300',
'3000',
'300603',
'300p',
'3030',
'30apr',
'30ish',
'30pm',
'30pp',
'30s',
'31',
'3100',
'310303',
'31p',
'32',
'32000',
'326',
'33',
'330',
'350',
'3510i',
'3650',
'36504',
```

```
'3680',
'3750',
'37819',
'38',
'382',
'391784',
'3aj',
'3d',
'3days',
'3g',
'3gbp',
'3hrs',
'3lions',
'3lp',
'3miles',
'3mins',
'3mobile',
'3optical',
'3pound',
'3qxj9',
'3rd',
'3ss',
'3uz',
'3wks',
'3x',
'3xx',
'40',
'400',
'400mins',
'400thousad',
'402',
'40411',
'40533',
'40gb',
'40mph',
'41685',
'41782',
'420',
'4217',
'42810',
'430',
'434',
'44',
'440',
'4403ldnw1a7rw18',
'44345',
'447797706009',
'447801259231',
```

```
'448712404000',
'449050000301',
'45',
'450p',
'450ppw',
'45239',
'45pm',
'47',
'4719',
'4742',
'47per',
'48',
'4882',
'48922',
'49',
'49557',
'4a',
'4d',
'4eva',
'4few',
'4fil',
'4get',
'4got',
'4info',
'4msgs',
'4mths',
'4qf2',
'4t',
'4th',
'4txt',
'4u',
'4utxt',
'4w',
'4wrd',
'4xx26',
'4years',
'50',
'500',
'5000',
'50award',
'50ea',
'50gbp',
'50p',
'50perweeksub',
'50perwksub',
'50pm',
'50ppm',
'50rcvd',
```

```
'50s',
'515',
'5226',
'523',
'5249',
'526',
'528',
'530',
'54',
'542',
'5digital',
'5free',
'5ish',
'5k',
'5min',
'5mls',
'5p',
'5pm',
'5th',
'5wb',
'5we',
'5wkg',
'5wq',
'5years',
'60',
'600',
'6031',
'6089',
'60p',
'61',
'61200',
'61610',
'62468',
'630',
'63miles',
'645',
'65',
'66',
'6669',
'674',
'67441233',
'68866',
'69101',
'69669',
'69696',
'69698',
'69855',
'69866',
```

```
'69888',
'69888nyt',
'69911',
'69969',
'69988',
'6hrs',
'6ish',
'6missed',
'6months',
'6ph',
'6pm',
'6th',
'6wu',
'6zf',
'700',
'7250',
'7250i',
'730',
'731',
'75',
'750',
'75max',
'762',
'7634',
'7684',
'77',
'7732584351',
'78',
'786',
'7876150ppm',
'7am',
'7cfca1a',
'7ish',
'7oz',
'7pm',
'7th',
'7ws',
'800',
'8000930705',
'80062',
'8007',
'80082',
'80086',
'80122300p',
'80155',
'80182',
'8027',
'80488',
```

```
'80608',
'8077',
'80878',
'81010',
'81151',
'81303',
'81618',
'820554ad0a1705572711',
'82242',
'82277',
'82324',
'82468',
'83049',
'83110',
'83118',
'83222',
'83332',
'83338',
'83355',
'83600',
'83738',
'84',
'84025',
'84122',
'84128',
'84199',
'84484',
'85',
'85023',
'85069',
'85222',
'85233',
'8552',
'86021',
'861',
'864233',
'86688',
'86888',
'87021',
'87066',
'87070',
'87077',
'87121',
'87131',
'872',
'87239',
'87575',
'88039',
```

```
'88066',
'88088',
'88222',
'88600',
'88800',
'8883',
'88877',
'88888',
'89034',
'89070',
'89080',
'89105',
'89545',
'89555',
'89693',
'89938',
'8am',
'8ball',
'8lb',
'8p',
'8pm',
'8th',
'8wp',
'900',
'9061100010',
'910',
'9153',
'92h',
'930',
'9307622',
'945',
'95',
'9755',
'9758',
'99',
'9996',
'9ae',
'9am',
'9pm',
'9t',
'9th',
'9yt',
'____',
'a21',
'a30',
'aah',
'aaniye',
'aaoooright',
```

```
'aathi',
'abbey',
'abdomen',
'abeg',
'abel',
'aberdeen',
'abi',
'ability',
'abiola',
'abj',
'able',
'abnormally',
'aboutas',
'abroad',
'absence',
'absolutely',
'absolutly',
'abstract',
'abt',
'abta',
'ac',
'academic',
'acc',
'accent',
'accenture',
'accept',
'access',
'accessible',
'accidant',
'accident',
'accidentally',
'accommodation',
'accomodate',
'accomodations',
'accordin',
'accordingly',
'account',
'accounting',
'accounts',
'achan',
'ache',
'achieve',
'acid',
'acl03530150pm',
'acnt',
'aco',
'act',
'acted',
```

```
'actin',
'acting',
'action',
'activ8',
'activate',
'active',
'activities',
'actor',
'actual',
'actually',
'ad',
'adam',
'add',
'addamsfa',
'added',
'addicted',
'addie',
'adding',
'address',
'adewale',
'adjustable',
'admin',
'administrator',
'admirer',
'admission',
'admit',
'adore',
'adoring',
'adp',
'adress',
'adrian',
'adrink',
'ads',
'adsense',
'adult',
'adults',
'advance',
'adventure',
'advice',
'advise',
'advising',
'advisors',
'aeronautics',
'aeroplane',
'affair',
'affairs',
'affection',
'affections',
```

```
'affidavit',
'afford',
'afghanistan',
'afraid',
'africa',
'african',
'aft',
'afternon',
'afternoon',
'aftr',
'ag',
'age',
'age16',
'age23',
'agency',
'agent',
'agents',
'ages',
'agidhane',
'aging',
'ago',
'agree',
'ah',
'aha',
'ahead',
'ahhh',
'ahhhh',
'ahmad',
'aids',
'aig',
'aight',
'ain',
'aint',
'air',
'air1',
'airport',
'airtel',
'aiya',
'aiyah',
'aiyar',
'aiyo',
'ajith',
'ak',
'aka',
'akon',
'al',
'alaipayuthe',
'album',
```

```
'alcohol',
'aldrine',
'alert',
'alertfrom',
'alerts',
'alex',
'alfie',
'algarve',
'algebra',
'ali',
'alian',
'alibi',
'alive',
'allah',
'allday',
'allow',
'allowed',
'allows',
'alot',
'alright',
'alrite',
'alto18',
'alwa',
'alwys',
'amanda',
'amazing',
'ambitious',
'american',
'ami',
'amk',
'amla',
'amma',
'ammae',
'ammo',
'amore',
'amp',
'amplikater',
'amrita',
'ams',
'amt',
'amused',
'analysis',
'anand',
'anderson',
'andre',
'andres',
'andrews',
'angry',
```

```
'animal',
'animation',
'anna',
'annie',
'anniversary',
'announced',
'announcement',
'annoyin',
'annoying',
'anot',
'ans',
'ansr',
'answer',
'answered',
'answerin',
'answering',
'answers',
'answr',
'antelope',
'antha',
'anthony',
'anti',
'antibiotic',
'anybody',
'anymore',
'anyones',
'anyplaces',
'anythiing',
'anythin',
'anythingtomorrow',
'anytime',
'anyways',
'aom',
'apart',
'apartment',
'apes',
'apeshit',
'aphex',
'apnt',
'apo',
'apologetic',
'apologise',
'apologize',
'apology',
'app',
'apparently',
'appeal',
'appear',
```

```
 'appendix',
 'applausestore',
 'applebees',
 'apples',
 'application',
 'apply',
 'applying',
 'appointment',
 'appointments',
 'appreciate',
 'appropriate',
 'approve',
 'approved',
 'approx',
 'apps',
 'appt',
 'appy',
 'april',
 'aproach',
 'aptitude',
 'aquarius',
 'ar',
 'arab',
 'arabian',
 'arcade',
 'archive',
 'ard',
 'area',
 'aren',
 'arent',
 'arestaurant',
 'aretaking',
 'argentina',
 'argh',
 'argue',
 'arguing',
 ...]
```

```
In [16]: # transform
         X_train_transformed = vect.transform(X_train)
         X_test_tranformed =vect.transform(X_test)
```

```
In [17]: from sklearn.naive_bayes import BernoulliNB

         # instantiate bernoulli NB object
         bnb = BernoulliNB()

         # fit
```

```
bnb.fit(X_train_transformed,y_train)

# predict class
y_pred_class = bnb.predict(X_test_tranformed)

# predict probability
y_pred_proba =bnb.predict_proba(X_test_tranformed)

# accuracy
from sklearn import metrics
metrics.accuracy_score(y_test, y_pred_class)
```

Out[17]:
```
0.9770279971284996
```

In [18]:
```
bnb
```

Out[18]:
```
BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)
```

In [19]:
```
metrics.confusion_matrix(y_test, y_pred_class)
```

Out[19]:
```
array([[1207,    1],
       [  31,  154]])
```

In [20]:
```
confusion = metrics.confusion_matrix(y_test, y_pred_class)
print(confusion)
#[row, column]
TN = confusion[0, 0]
FP = confusion[0, 1]
FN = confusion[1, 0]
TP = confusion[1, 1]
```

```
[[1207    1]
 [  31  154]]
```

In [21]:
```
sensitivity = TP / float(FN + TP)
print("sensitivity",sensitivity)
```

```
sensitivity 0.8324324324324325
```

In [22]:
```
specificity = TN / float(TN + FP)

print("specificity",specificity)
```

```
specificity 0.9991721854304636
```

In [23]:
```
precision = TP / float(TP + FP)

print("precision",precision)
print(metrics.precision_score(y_test, y_pred_class))
```

```
precision 0.9935483870967742
0.9935483870967742
```

```
In [24]: 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 0.9935483870967742
PRECISION SCORE : 0.9935483870967742
RECALL SCORE : 0.8324324324324325
F1 SCORE : 0.9058823529411765
```

```
In [25]: y_pred_proba
```

```
Out[25]: array([[1.00000000e+00, 1.54647985e-10],
                [1.00000000e+00, 1.69177313e-10],
                [9.99999997e-01, 2.52488636e-09],
                ...,
                [9.99994219e-01, 5.78141360e-06],
                [3.64570741e-06, 9.99996354e-01],
                [1.00000000e+00, 2.90374743e-12]])
```

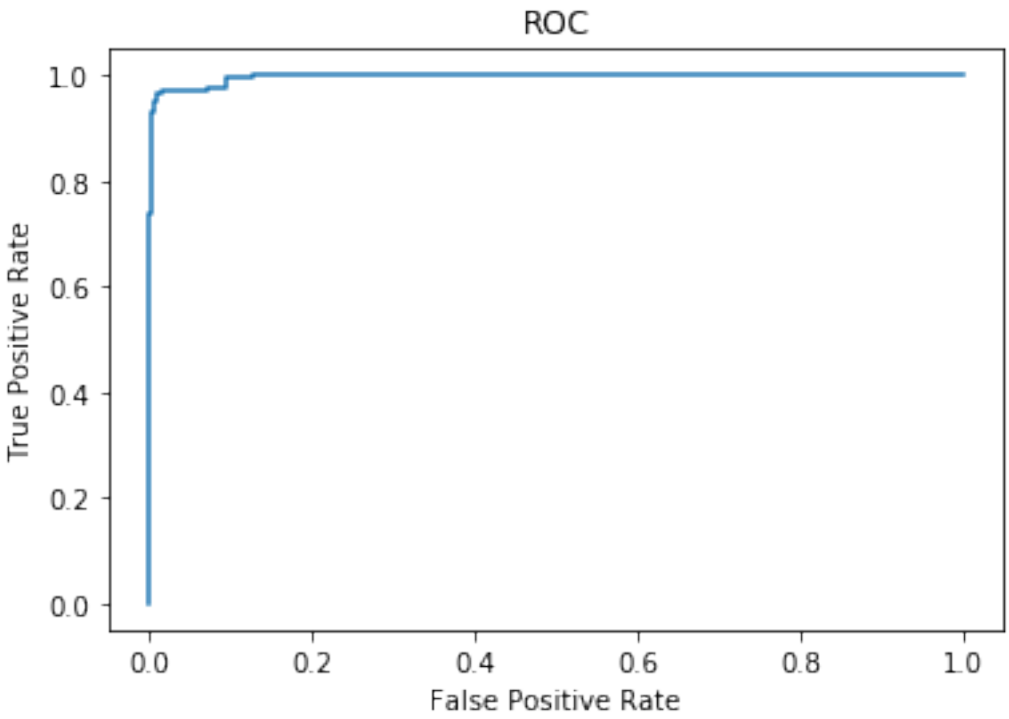
```
In [26]: from sklearn.metrics import confusion_matrix as sk_confusion_matrix
         from sklearn.metrics import roc_curve, auc
         import matplotlib.pyplot as plt
         false_positive_rate, true_positive_rate, thresholds = roc_curve(y_test, y_pred_proba[
         roc_auc = auc(false_positive_rate, true_positive_rate)
```

```
In [27]: print (roc_auc)
```

```
0.9967692858421334
```

```
In [31]: %matplotlib inline
         plt.ylabel('True Positive Rate')
         plt.xlabel('False Positive Rate')
         plt.title('ROC')
         plt.plot(false_positive_rate, true_positive_rate)
```

```
Out[31]: [<matplotlib.lines.Line2D at 0x1a1d720978>]
```


ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0