In [7]:
```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

In [8]:
```
df = pd.read_csv('mail_data.csv')
```

In [9]:
```
print(df)
```

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

[5572 rows x 2 columns]
```

In [10]:
```
data = df.where((pd.notnull(df)), '')
```

In [11]:
```
data.head(10)
```

Out[11]:

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

In [12]:
```
data.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 [13]:
```
data.shape
```

Out[13]:
```
(5572, 2)
```

In [14]:
```
data.loc[data['Category'] == 'spam', 'Category',]=0
data.loc[data['Category'] == 'ham', 'Category',]=1
```

In [15]:
```
X=data['Message']
Y=data['Category']
```

In [16]:
```
print(X)
```

```
0       Go until jurong point, crazy.. Available only ...
1                           Ok lar... Joking wif u oni...
2       Free entry in 2 a wkly comp to win FA Cup fina...
3       U dun say so early hor... U c already then say...
4       Nah I don't think he goes to usf, he lives aro...
                              ...                        
5567    This is the 2nd time we have tried 2 contact u...
5568                Will ü b going to esplanade fr home?
5569    Pity, * was in mood for that. So...any other s...
5570    The guy did some bitching but I acted like i'd...
5571                         Rofl. Its true to its name
Name: Message, Length: 5572, dtype: object
```

In [17]:
```
print(Y)
```

```
0       1
1       1
2       0
3       1
4       1
       ..
5567    0
5568    1
5569    1
5570    1
5571    1
Name: Category, Length: 5572, dtype: object
```

In [18]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, random_state = 3)
```

In [19]:
```
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

In [20]:
```
print(Y.shape)
print(Y_train.shape)
print(Y_test.shape)
```

```
(5572,)
(4457,)
(1115,)
```

In [21]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
feature_extraction = TfidfVectorizer(min_df = 1, stop_words = 'english', lowercase=True)

X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)

Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

In [22]:
```
print(X_train)
```

```
3075                  Don know. I did't msg him recently.
1787    Do you know why god created gap between your f...
1614                          Thnx dude. u guys out 2nite?
4304                                 Yup i'm free...
3266    44 7732584351, Do you want a New Nokia 3510i c...
                              ...                        
789     5 Free Top Polyphonic Tones call 08701728737,...
968     What do u want when i come back?.a beautiful n...
1667    Guess who spent all last night phasing in and ...
3321    Eh sorry leh... I din c ur msg. Not sad alread...
1688    Free Top ringtone -sub to weekly ringtone-get ...
Name: Message, Length: 4457, dtype: object
```

In [23]:
```
print(X_train_features)
```

```
  (0, 5413)     0.6198254967574347
  (0, 4456)     0.4168658090846482
  (0, 2224)     0.413103377943378
  (0, 3811)     0.34780165336091333
  (0, 2329)     0.38783870336935383
  (1, 4080)     0.18880584110891163
  (1, 3185)     0.29694482957694585
  (1, 3325)     0.31610586766078863
  (1, 2957)     0.3398297002864083
  (1, 2746)     0.3398297002864083
  (1, 918)      0.22871581159877646
  (1, 1839)     0.27849035900561455
  (1, 2758)     0.3226407885943799
  (1, 2956)     0.3303699595553024
  (1, 1091)     0.3303699595553024
  (1, 3046)     0.25037127926213518
  (1, 3811)     0.17419952275504033
  (2, 407)      0.5092725360051008
  (2, 3156)     0.41072393183112698
  (2, 2404)     0.4528771107006745
  (2, 6601)     0.6056811524587518
  (3, 2870)     0.5864269879324768
  (3, 7414)     0.8100020912469564
  (4, 50)       0.23633754072626942
  (4, 5497)     0.15743785051118356
  :     :
  (4454, 4602)  0.26697657324445391
  (4454, 3142)  0.32014451677763156
  (4455, 2247)  0.37052851863170466
  (4455, 2469)  0.35441545511837946
  (4455, 5646)  0.33545678464631296
  (4455, 6810)  0.29731757715898277
  (4455, 6091)  0.23103841516927642
  (4455, 7113)  0.30536590342067704
  (4455, 3872)  0.31089114917886658
  (4455, 4715)  0.30714144758811196
  (4455, 6916)  0.19636985317119715
  (4455, 3922)  0.31287563163368587
  (4455, 4456)  0.24920025316229423
  (4456, 141)   0.292943737785358
  (4456, 647)   0.30133182431707617
  (4456, 6311)  0.30133182431707617
  (4456, 5569)  0.4619395404299172
  (4456, 6028)  0.21034888000987115
  (4456, 7154)  0.24083218452280053
  (4456, 7150)  0.3677554681447669
  (4456, 6249)  0.17573831794959716
  (4456, 6307)  0.2752760476857975
  (4456, 334)   0.22280077711654938
  (4456, 5778)  0.16243864490190795
  (4456, 2870)  0.31523196273113385
```

In [26]:
```
print(X_test_features)
```

```
  (0, 7271)     0.19403270081790069
  (0, 6920)     0.20571591693537986
  (0, 5373)     0.23656908724638063
  (0, 5213)     0.19885473575902182
  (0, 4386)     0.18353333634030898098
  (0, 1549)     0.26464988483971888
  (0, 1405)     0.31768639389314351
  (0, 1361)     0.25132445289897426
  (0, 1082)     0.24510684362545027
  (0, 1041)     0.28016206931555726
  (0, 405)      0.23813163030013606
  (0, 306)      0.23975986557206702
  (0, 20)       0.30660832384591537
  (0, 14)       0.26797874473132896
  (0, 9)        0.2852706805264544
  (0, 1)        0.23813163030013606
  (1, 7368)     0.29957800964520975
  (1, 6732)     0.42473488670029325
  (1, 6588)     0.32989377075962767
  (1, 6507)     0.26731535902873493
  (1, 6214)     0.3621564482127515
  (1, 4729)     0.22965776503163893
  (1, 4418)     0.34576968913168
  (1, 3491)     0.4960939561018028
  (2, 7205)     0.22344717215670331
  :     :
  (1110, 3167)  0.5718357066163949
  (1111, 7353)  0.4991205841293424
  (1111, 6787)  0.4005017571427885
  (1111, 6033)  0.4714849709283488
  (1111, 3227)  0.4438493577273525
  (1111, 2440)  0.4137350055985486
  (1112, 7071)  0.33558524648843113
  (1112, 6777)  0.32853717524096393
  (1112, 6297)  0.30568996872268727
  (1112, 5778)  0.2280742809854942
  (1112, 5695)  0.3381604952481646
  (1112, 5056)  0.25591830435595413
  (1112, 4170)  0.33078356237138663
  (1112, 2329)  0.24185689037749
  (1112, 1683)  0.40170874362720344
  (1112, 1109)  0.35334496762883244
  (1113, 4080)  0.3045947361955407
  (1113, 4038)  0.37023505294137066
  (1113, 3811)  0.2810398058655096
  (1113, 3281)  0.33232508601719535
  (1113, 3113)  0.33840833425155675
  (1113, 2852)  0.5956422293158835
  (1113, 2224)  0.3337959267453531
  (1114, 4557)  0.5196253874825217
  (1114, 4033)  0.8543942045002639
```

In [25]:
```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train_features, Y_train)
```

Out[25]:
```
LogisticRegression()
```

In [27]:
```
prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

In [28]:
```
print('Acc on training data: ', accuracy_on_training_data)
```

```
Acc on training data:  0.9670181736594121
```

In [29]:
```
prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
```

In [30]:
```
print('acc on test data: ', accuracy_on_test_data)
```

```
acc on test data:  0.9659192825112107
```

In [39]:
```
input_your_mail = ["Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's"]

input_data_features = feature_extraction.transform(input_your_mail)

prediction = model.predict(input_data_features)

print(prediction)

if(prediction[0]==1):
    print('Ham mail')
else:
    print('Spam mail')
```

```
[0]
Spam mail
```

In [ ]:
```

```