In [6]:

```
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 [7]:

```
raw_mail_data = pd.read_csv("C:/Users/shami/Downloads/mail_dataset.csv")
```

In [8]:

```
raw_mail_data
```

Out[8]:

	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 × 2 columns

In [9]:

```
mail_data = raw_mail_data.where((pd.notnull(raw_mail_data)),'')
```

In [10]:

```
mail_data.head()
```

Out[10]:

	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 [11]:

```
mail_data.shape
```

Out[11]:

```
(5572, 2)
```

In [13]:

```
mail_data[mail_data['Category'] == 'ham']
```

Out[13]:

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
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...
6	ham	Even my brother is not like to speak with me. ...
...	...	...
5565	ham	Huh y lei...
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

4825 rows × 2 columns

In [14]:

```
mail_data[mail_data['Category'] == 'spam']
```

Out[14]:

	Category	Message
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...
...	...	...
5537	spam	Want explicit SEX in 30 secs? Ring 02073162414...
5540	spam	ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...
5547	spam	Had your contract mobile 11 Mnths? Latest Moto...
5566	spam	REMINDER FROM O2: To get 2.50 pounds free call...
5567	spam	This is the 2nd time we have tried 2 contact u...

747 rows × 2 columns

In [15]:

```
mail_data.loc[mail_data['Category'] == 'spam', 'Category'] = 0
mail_data.loc[mail_data['Category'] == 'ham', 'Category'] = 1
```

In [16]:

```
X = mail_data['Message']
Y = mail_data['Category']
```

In [17]:
```
X
```

Out[17]:
```
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 [18]:
```
Y
```

Out[18]:
```
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 [19]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.2,random_state=3
```

In [21]:
```
print(X_train.shape)
print(X_test.shape)
```

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

In [23]:
```
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 [24]:
```
X_train
```

Out[24]:
```
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 087018728737,...
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 [26]:
```
print(X_train_features)
```

```
  (0, 2329)	0.38783870336935383
  (0, 3811)	0.34780165336891333
  (0, 2224)	0.413103377943378
  (0, 4456)	0.4168658090846482
  (0, 5413)	0.6198254967574347
  (1, 3811)	0.17419952275504033
  (1, 3046)	0.2503712792613518
  (1, 1991)	0.33036995955537024
  (1, 2956)	0.33036995955537024
  (1, 2758)	0.3226407885943799
  (1, 1839)	0.2784903590561455
  (1, 918)	0.22871581159877646
  (1, 2746)	0.3398297002864083
  (1, 2957)	0.3398297002864083
  (1, 3325)	0.31610586766078863
  (1, 3185)	0.29694482957694585
  (1, 4080)	0.18880584110891163
  (2, 6601)	0.6056811524587518
  (2, 2404)	0.45287711070606745
  (2, 3156)	0.4107239318312698
  (2, 407)	0.509272536051008
  (3, 7414)	0.8100020912469564
  (3, 2870)	0.5864269879324768
  (4, 2870)	0.41872147309323743
  (4, 487)	0.2899118421746198
  :	:
  (4454, 2855)	0.47210665083641806
  (4454, 2246)	0.47210665083641806
  (4455, 4456)	0.24920025316220423
  (4455, 3922)	0.31287563163368587
  (4455, 6916)	0.19636985317119715
  (4455, 4715)	0.30714144758811196
  (4455, 3872)	0.3108911491788658
  (4455, 7113)	0.30536590342067704
  (4455, 6091)	0.23103841516927642
  (4455, 6810)	0.29731757715898277
  (4455, 5646)	0.33545678464631296
  (4455, 2469)	0.35441545511837946
  (4455, 2247)	0.37052851863170466
  (4456, 2870)	0.31523196273113385
  (4456, 5778)	0.16243064490100795
  (4456, 334)	0.2220077711654938
  (4456, 6307)	0.2752760476857975
  (4456, 6249)	0.17573831794959716
  (4456, 7150)	0.3677554681447669
  (4456, 7154)	0.24083218452280053
  (4456, 6028)	0.21034888000987115
  (4456, 5569)	0.4619395404299172
  (4456, 6311)	0.30133182431707617
  (4456, 647)	0.30133182431707617
  (4456, 141)	0.292943737785358
```

In [27]:
```
model = LogisticRegression()
```

In [28]:
```
model.fit(X_train_features,Y_train)
```

Out[28]:
```
▾ LogisticRegression ⓘ ? (https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegressio
LogisticRegression()
```

In [29]:

```
prediction_on_training_data = model.predict(X_train_features)
accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)
```

In [30]:

```
accuracy_on_training_data
```

Out[30]:

```
0.9676912721561588
```

In [32]:

```
prediction_on_test_data = model.predict(X_test_features)
accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)
accuracy_on_test_data
```

Out[32]:

```
0.9668161434977578
```

In [42]:

```
input_mail = ["Sunshine Quiz Wkly Q! Win a top Sony DVD player if u know which count
input_data_features = feature_extraction.transform(input_mail)
prediction = model.predict(input_data_features)

if(prediction[0] == 1):
    print("It is a Ham mail...")
else:
    print("It is a spam mail!!!")
```

```
It is a spam mail!!!
```

In [ ]: