In [280…
```
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, classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
```

In [12]:
```
data = pd.read_csv('spam.csv')
```

In [14]:
```
data.keys()
```

Out[14]:
```
Index(['v1', 'v2', 'Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'], dtype='object')
```

In [16]:
```
data.head()
```

Out[16]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	NaN	NaN	NaN
4	ham	Nah I don't think he goes to usf, he lives aro...	NaN	NaN	NaN

In [18]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

In [20]:
```
data["v1"].value_counts()
```

Out[20]:
```
v1
ham     4825
spam     747
Name: count, dtype: int64
```

In [22]:
```
data.describe()
```

Out[22]:

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
count	5572	5572	50	12	6
unique	2	5169	43	10	5
top	ham	Sorry, I'll call later	bt not his girlfrnd... G o o d n i g h t . . .@"	MK17 92H. 450Ppw 16"	GNT:-)"
freq	4825	30	3	2	2

In [24]:
```
data = data.drop(columns=['Unnamed: 2', 'Unnamed: 3', 'Unnamed: 4'])
```

In [26]:
```
data
```

Out[26]:

	v1	v2
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 Ì_ 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 [28]:
```
data.describe()
```

Out[28]:

	v1	v2
count	5572	5572
unique	2	5169
top	ham	Sorry, I'll call later
freq	4825	30

In [30]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   v1      5572 non-null   object
 1   v2      5572 non-null   object
dtypes: object(2)
memory usage: 87.2+ KB
```

In [32]:
```
data = data.where((pd.notnull(data)), '')
```

In [36]:
```
data.shape
```

Out[36]:
```
(5572, 2)
```

In [38]:
```
data.loc[data['v1'] == 'spam', 'v1',] = 0
```

In [40]:
```
data.loc[data['v1'] == 'ham', 'v1',] = 1
```

In [44]:
```
X = data['v2']
y = data['v1']
```

In [46]:
```
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 Ì_ 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: v2, Length: 5572, dtype: object
```

In [48]:
```
print(y)
```

```
0       1
1       1
2       0
3       1
4       1
       ..
5567    0
5568    1
5569    1
5570    1
5571    1
Name: v1, Length: 5572, dtype: object
```

In [66]:
```
X_train, X_test, y_train,y_test = train_test_split(X,y, test_size=0.2, random_state
```

In [68]:
```
print(X.shape)
print(X_train.shape)
print(X_test.shape)
```

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

In [70]:
```
print(y.shape)
print(y_train.shape)
print(y_test.shape)
```

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

In [234…
```
feature_extraction = TfidfVectorizer(min_df = 1, stop_words = 'english', lowercase
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 [238…
```
print(X_train_features)
```

```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 34759 stored elements and shape (4457, 7511)>
  Coords        Values
  (0, 4513)     0.2909649098524696
  (0, 3380)     0.21807195185332803
  (0, 3262)     0.25877035357606315
  (0, 3136)     0.440116181574609
  (0, 2122)     0.38613577623520473
  (0, 3386)     0.3219352588930141
  (0, 6599)     0.20296878731699391
  (0, 4296)     0.3891385935794867
  (0, 3979)     0.2410582143632299
  (0, 741)      0.3219352588930141
  (1, 7443)     0.35056971070320353
  (1, 6442)     0.5652509076654626
  (1, 6417)     0.4769136859540388
  (1, 6872)     0.4306015894277422
  (1, 4061)     0.380431198316959
  (2, 5825)     0.4917598465723273
  (2, 2226)     0.413484525934624
  (2, 3917)     0.40088501350982736
  (2, 2109)     0.42972812260098503
  (2, 933)      0.4917598465723273
  (3, 7453)     0.5202633571003087
  (3, 1842)     0.3708680641487708
  (3, 1599)     0.5927091854194291
  (3, 6140)     0.4903863168693604
  (4, 1842)     0.36051481797205776
  :     :
  (4452, 4636)  0.4030918768627523
  (4453, 1762)  0.45610005640082985
  (4453, 7273)  0.5787739591782677
  (4453, 999)   0.6760129013031282
  (4454, 5370)  0.42618909997886
  (4454, 7346)  0.31166263834107377
  (4454, 1049)  0.31932060116006045
  (4454, 2001)  0.4166919007849217
  (4454, 3088)  0.34475593009514444
  (4454, 2086)  0.3809693742808703
  (4454, 3029)  0.42618909997886
  (4455, 4773)  0.35860460546223444
  (4455, 3763)  0.16807158405536876
  (4455, 4251)  0.30616657078392584
  (4455, 2108)  0.3136468384526087
  (4455, 7407)  0.3028481995557642
  (4455, 7358)  0.2915949626395065
  (4455, 2764)  0.3226323745940581
  (4455, 6361)  0.25697343671652706
  (4455, 6433)  0.38998123077430413
  (4455, 1148)  0.38998123077430413
  (4456, 4557)  0.48821933148688146
  (4456, 1386)  0.4460036316446079
  (4456, 6133)  0.5304350313291551
  (4456, 6117)  0.5304350313291551
```

In [240…

```
print(X_train)
```

```
3075    Mum, hope you are having a great day. Hoping t...
1787                          Yes:)sura in sun tv.:)lol.
1614    Me sef dey laugh you. Meanwhile how's my darli...
4304                Yo come over carlos will be here soon
3266                    Ok then i come n pick u at engin?
                              ...
789                         Gud mrng dear hav a nice day
968             Are you willing to go for aptitude class.
1667    So now my dad is gonna call after he gets out ...
3321    Ok darlin i supose it was ok i just worry too ...
1688                     Nan sonathaya soladha. Why boss?
Name: v2, Length: 4457, dtype: object
```

In [242…

```
print(X_test_features)
```

```
<Compressed Sparse Row sparse matrix of dtype 'float64'
        with 7766 stored elements and shape (1115, 7511)>
  Coords        Values
  (0, 1537)     0.667337188824809
  (0, 4294)     0.5159375448718375
  (0, 6007)     0.537093591660729
  (1, 1)        0.21260233518669944
  (1, 43)       0.24547458936715755
  (1, 321)      0.28671640581392144
  (1, 520)      0.1934450786526249
  (1, 602)      0.28671640581392144
  (1, 2899)     0.1385795841356552
  (1, 3300)     0.37297727661877506
  (1, 3365)     0.28671640581392144
  (1, 4045)     0.250549335510249
  (1, 5250)     0.28671640581392144
  (1, 5347)     0.2733682162643466
  (1, 5501)     0.28671640581392144
  (1, 6579)     0.2733682162643466
  (1, 6599)     0.14954692788663673
  (1, 7222)     0.23059492898537964
  (2, 2939)     0.47195476517479323
  (2, 2941)     0.6068486133983123
  (2, 4070)     0.44361668503137164
  (2, 6648)     0.3410121739015846
  (2, 6701)     0.30969080396105314
  (3, 1606)     0.28517759021090444
  (3, 2649)     0.303870736800912
  :     :
  (1111, 2458)  0.42325261089251354
  (1111, 3259)  0.44776220819286267
  (1111, 6093)  0.467191431141905
  (1111, 6848)  0.3968546202564372
  (1111, 7415)  0.49457538286455366
  (1112, 2114)  0.32870972643480745
  (1112, 2704)  0.3704547809702327
  (1112, 2780)  0.3745139316876871
  (1112, 3259)  0.3631408033721114
  (1112, 3432)  0.3631408033721114
  (1112, 4282)  0.35091845697551116
  (1112, 4903)  0.47703903024985594
  (1113, 1657)  0.44289971323548966
  (1113, 3239)  0.488439471695463
  (1113, 3963)  0.3910346709289789
  (1113, 5806)  0.488439471695463
  (1113, 6846)  0.4168758749641195
  (1114, 2352)  0.270495916357943
  (1114, 2862)  0.38140394975458775
  (1114, 2899)  0.2421646568502054
  (1114, 3564)  0.40844238751288037
  (1114, 5073)  0.3194139844000448
  (1114, 5565)  0.5010303679312903
  (1114, 6902)  0.3063326681877805
  (1114, 7295)  0.33014792863496223
```

In [244…

```
print(X_test)
```

```
2632                        I WILL CAL YOU SIR. In meeting
454     Loan for any purpose å£500 - å£75,000. Homeown...
983     LOOK AT THE FUCKIN TIME. WHAT THE FUCK YOU THI...
1282    Ever green quote ever told by Jerry in cartoon...
4610                                  Wat time Ì_ finish?
                              ...
4827    Lol no. Just trying to make your day a little ...
5291      Xy trying smth now. U eat already? We havent...
3325    Huh so fast... Dat means u havent finished pai...
3561    Still chance there. If you search hard you wil...
1136    Dont forget you can place as many FREE Request...
Name: v2, Length: 1115, dtype: object
```

In [246…
```
print(y_train)
```

```
3075    1
1787    1
1614    1
4304    1
3266    1
       ..
789     1
968     1
1667    1
3321    1
1688    1
Name: v1, Length: 4457, dtype: int32
```

In [248…
```
print(y_test)
```

```
2632    1
454     0
983     1
1282    1
4610    1
       ..
4827    1
5291    1
3325    1
3561    1
1136    0
Name: v1, Length: 1115, dtype: int32
```

In [250…
```
model = LogisticRegression()
```

In [252…
```
model.fit(X_train_features, y_train_features)
```

Out[252…

▾ LogisticRegression

LogisticRegression()

In [254…
```
prediction = model.predict(X_train_features)
accuracy = accuracy_score(y_train_features, prediction)
```

In [258…
```
print("Accuracy on Training data :", accuracy)
```

```
Accuracy on Training data : 0.9661207089970832
```

In [260…
```
prediction_test = model.predict(X_test_features)
accuracy_test = accuracy_score(y_test, prediction_test)
```

In [262…
```
print("Accuracy on Testing data :", accuracy_test)
```

```
Accuracy on Testing data : 0.9623318385650225
```

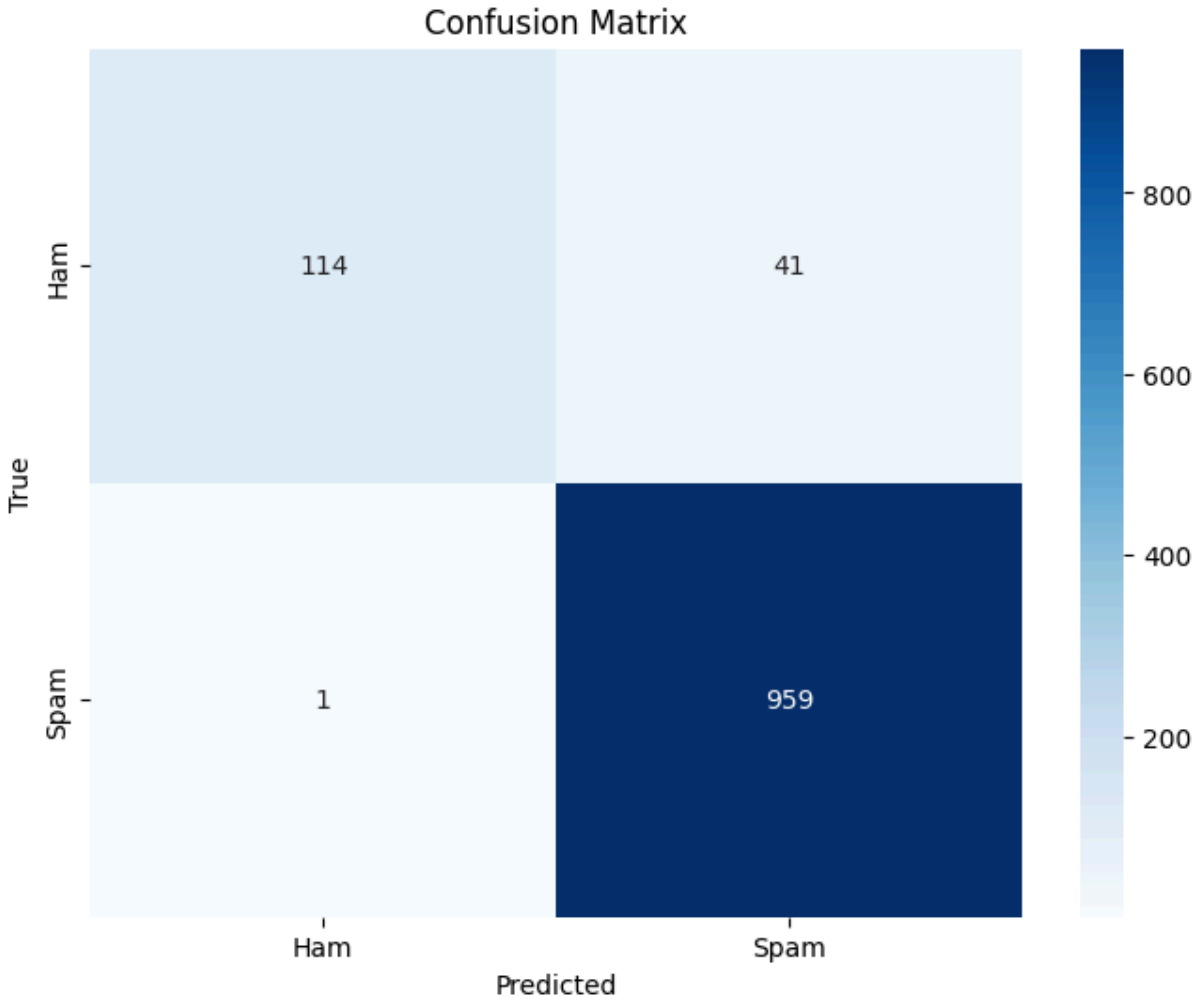
In [278…
```
predictions = model.predict(X_test_features)
report = classification_report(y_test, predictions)
print(report)
```

```
              precision    recall  f1-score   support

           0       0.99      0.74      0.84       155
           1       0.96      1.00      0.98       960

    accuracy                           0.96      1115
   macro avg       0.98      0.87      0.91      1115
weighted avg       0.96      0.96      0.96      1115
```

In [282…
```
predictions = model.predict(X_test_features)
cm = confusion_matrix(y_test, predictions)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Ham', 'Spam'], yti
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()
```



In [284…

```
input_your_mail = ['Go until jurong point, crazy.. Available only in bugis n great
input_data_features = feature_extraction.transform(input_your_mail)
prediction = model.predict(input_data_features)
if (prediction[0]==1):
    print("Ham mail")
else:
    print("Spam mail")
```

```
Ham mail
```