spam-detection

September 21, 2024

1 SPAM DETECTION

1.0.1 Importing necessary libraries

[1]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score
```

1.0.2 Reading the spam-ham file

[2]:
```
df=pd.read_csv(r'C:\Users\Windows\Desktop\projects\spam.csv')
```

1.0.3 Quick View of the dataset

[3]:
```
df.shape
```

[3]:
```
(5572, 2)
```

[4]:
```
df.head()
```

[4]:

	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]:
```
df.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
```

[6]:
```
df.dtypes
```

[6]:
```
Category    object
Message     object
dtype: object
```

2 EDA

2.0.1 Checking for duplicates and null values

[7]:
```
df.duplicated().sum()
```

[7]:
```
np.int64(415)
```

[8]:
```
df=df.drop_duplicates()
```

[9]:
```
df.duplicated().sum()
```

[9]:
```
np.int64(0)
```

[10]:
```
df.isnull().sum().sort_values(ascending = False)
```

[10]:
```
Category    0
Message     0
dtype: int64
```

2.0.2 Categorical to Numeric

[11]:
```
mapping = {'spam':1,'ham':0}
df['Category']=df['Category'].map(mapping)
```

[12]:
```
df['Category']
```

[12]:
```
0       0
1       0
2       1
3       0
4       0
       ..
5567    1
5568    0
```

```
5569    0
5570    0
5571    0
Name: Category, Length: 5157, dtype: int64
```

2.0.3 Features(x) and Target value(y)

[13]:
```
x = df['Message']
y=df['Category']
```

2.0.4 Spilitting the data

[14]:
```
x_train , x_test , y_train , y_test = train_test_split(x,y,train_size = 0.25,␣
↪random_state=42)
```

2.0.5 Model Preparation

[15]:
```
clf=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('nb',MultinomialNB())
])
```

2.0.6 Model Training

[16]:
```
clf.fit(x_train,y_train)
```

[16]:
```
Pipeline(steps=[('vectorizer', CountVectorizer()), ('nb', MultinomialNB())])
```

2.0.7 Model Prediction

[17]:
```
y_pred = clf.predict(x_test)
```

2.1 Some Unseen Input here

Here I given Two email Two detect 1st One is ham and the other one looking spam

[18]:
```
emails=[
    'Sounds great! Are you home now?',
    'Will u meet ur dream partner soon? Is ur career off 2 a flyng start? 2␣
↪find out free, txt HORO followed by ur star sign, e. g. HORO ARIES'
]
```

3 Predict Email ←→

[19]:
```
clf.predict(emails)
```

[19]:
```
array([0, 1])
```

It got it correct

3.1 ACCURACY OF MODEL

[20]:
```
accuracy = accuracy_score (y_test , y_pred)
accuracy
```

[20]:
```
0.9788004136504653
```

98% ACCURATE !

4 THANK YOU :)

[ ]:
```

```