mail-spam-detection-98-accuracy-1

April 3, 2024

[14]:
```
# This Python 3 environment comes with many helpful analytics libraries␣
 ↪installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/
 ↪docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list␣
 ↪all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that␣
 ↪gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved␣
 ↪outside of the current session

data=pd.read_csv('/kaggle/input/spam-email/spam.csv')
data
data.columns
data.info()
data.isna().sum()
data['Spam']=data['Category'].apply(lambda x:1 if x=='spam' else 0)
data.head(5)
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(data.Message,data.
 ↪Spam,test_size=0.25)
#CounterVectorizer Convert the text into matrics
from sklearn.feature_extraction.text import CountVectorizer

from sklearn.naive_bayes import MultinomialNB
```

```
from sklearn.pipeline import Pipeline
clf=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('nb',MultinomialNB())
])

clf.fit(X_train,y_train)
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'
]

clf.predict(emails); clf.predict(emails)

clf.score(X_test,y_test)
```

```
/kaggle/input/spam-email/spam.csv
<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
```

[14]:
```
0.9856424982053122
```

[ ]:
```
data=pd.read_csv('/kaggle/input/spam-email/spam.csv')
data
```

[ ]:

	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]

[ ]:
```
data.columns
```

[ ]:
```
Index(['Category', 'Message'], dtype='object')
```

[4]:
```
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
```

Dropped The Column Unnamed: 0

[5]:
```
data.isna().sum()
```

[5]:
```
Category    0
Message     0
dtype: int64
```

[6]:
```
data['Spam']=data['Category'].apply(lambda x:1 if x=='spam' else 0)
data.head(5)
```

[6]:

	Category	Message	Spam
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

[10]:
```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(data.Message,data.
 ↪Spam,test_size=0.25)
```

[11]:
```
#CounterVectorizer Convert the text into matrics
from sklearn.feature_extraction.text import CountVectorizer
```

Naive Bayes Have three Classifier(Bernouli,Multinominal,Gaussian) Here I use Multinominal Bayes Because here data in a discrete form discrete data(e.g movie ratings ranging 1 to 5 as each rating will have certain frequency to represent)

[ ]:
```
from sklearn.naive_bayes import MultinomialNB
```

[ ]:
```
from sklearn.pipeline import Pipeline
clf=Pipeline([
    ('vectorizer',CountVectorizer()),
    ('nb',MultinomialNB())
])
```

1 Tarining The Model

[ ]:
```
clf.fit(X_train,y_train)
```

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

[ ]:
```
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'
]
```

Predict Email

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

2 Prediction Of Model

[ ]:
```
clf.score(X_test,y_test)
```