NBA 2 Spam detection

December 18, 2022

[1]:
```
import pandas as pd
```

[2]:
```
df = pd.read_csv("spam.csv")
df.head()
```

[2]:
```
  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…
```

[3]:
```
df.groupby('Category').describe()
```

[3]:
```
          Message                                                                \
            count unique                                                    top
Category
ham          4825   4516                                Sorry, I'll call later
spam          747    641  Please call our customer service representativ…


          freq
Category
ham         30
spam         4
```

[4]:
```
df['spam']=df['Category'].apply(lambda x: 1 if x=='spam' else 0)
df.head()
```

[4]:
```
  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
```

[5]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(df.Message,df.spam)
```

[6]:
```
from sklearn.feature_extraction.text import CountVectorizer
v = CountVectorizer()
X_train_count = v.fit_transform(X_train.values)
X_train_count.toarray()[:2]
```

[6]:
```
array([[0, 0, 0, …, 0, 0, 0],
       [0, 0, 0, …, 0, 0, 0]], dtype=int64)
```

[7]:
```
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB()
model.fit(X_train_count,y_train)
```

[7]:
```
MultinomialNB()
```

[8]:
```
emails = [
    'Hey mohan, can we get together to watch footbal game tomorrow?',
    'Upto 20% discount on parking, exclusive offer just for you. Dont miss this␣
 ↪reward!'
]
emails_count = v.transform(emails)
model.predict(emails_count)
```

[8]:
```
array([0, 1], dtype=int64)
```

[9]:
```
X_test_count = v.transform(X_test)
model.score(X_test_count, y_test)
```

[9]:
```
0.9856424982053122
```

[10]:
```
from sklearn.pipeline import Pipeline
clf = Pipeline([
    ('vectorizer', CountVectorizer()),
    ('nb', MultinomialNB())
])
```

[11]:
```
clf.fit(X_train, y_train)
```

[11]:
```
Pipeline(steps=[('vectorizer', CountVectorizer()), ('nb', MultinomialNB())])
```

[12]:
```
clf.score(X_test,y_test)
```

[12]:
```
0.9856424982053122
```

[13]:
```
clf.predict(emails)
```

[13]:
```
array([0, 1], dtype=int64)
```

[ ]: