Certificate Course in Machine Learning using Python [6 Weeks]

Solving Text Classification Problem: Spam Detection (Mini Project 3)

Text Classification Problem: Spam Detection

Problem Statement: Classify the incoming email as ham (good) or spam (bad)

#Reading sms.txt file

```
import pandas as pd
sms=pd.read_csv("sms.txt",header=None,names=['label','message'],sep='\t')
```

To view data

```
sms.head()
```

To view entries in label column

```
sms.label.value_counts()
```

Preparing X and y

```
X=sms.message
y=sms.label
```

Spilt X and y in training and testing data sets. When test_size is not given, its default value is 0.25

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y)
```

Making dictionary of message for processing

```
vect=CountVectorizer()
vect.fit(X_train) #making dictionary
X_train_matrix=vect.transform(X_train)
```

Note: 7526 unique words exist in X_train messages. These are called dictionary for these messages.

```
X_test_matrix=vect.transform(X_test)
X_test_matrix
```

Applying MultinomialNB Model on X_train_matrix and y_train

```
from sklearn.naive_bayes import MultinomialNB
nb=MultinomialNB()
nb.fit(X_train_matrix,y_train)
```

Predicting the output of X_test_matrix

```
#nb is the ML model which you have made
y_predict=nb.predict(X_test_matrix)
```

Printing the score of the model

```
print(nb.score(X_test_matrix,y_test))
```

Applying Logistic Regression

```
from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()
lr.fit(X_train_matrix,y_train)
print(lr.score(X_test_matrix,y_test))
```

Next

PREVIOUS ACTIVITY

◄ Text Classification

Jump to...

NEXT ACTIVITY

Python Demo: Spam Detection Problem ►

Stay in touch

Contact Us

- http://nielit.gov.in/gorakhpur/
- abhinav@nielit.gov.in or ajay.verma@nielit.gov.in