Prajwal Gupta

RA1911003010660

AI LAB-11

Aim- Implementation of NLP programs

Problem Formulation- Solving a dataset using NLP.

Problem Statement- Building a spam classifier to predict if the given SMS are spam or not using NLP.

Algorithm used (Problem Solving)- Naïve Bayes

Human Language is modified into fragments or tokens which can be understood by the machine. Like in this problem, SMS collected from different sources will be tokenized and then analyzed and classified as whether they are spam or not.

After using NLP to modify the data, Naïve Bayes is used for the classification task. Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Dataset-

1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though
6	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
7	ham	Even my brother is not like to speak with me. They treat me like aids patent.
8	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
9	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
10	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
11	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
12	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
13	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
14	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful
15	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
16	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
17	ham	Oh k...i'm watching here:)
18	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
19	ham	Fine if that□s the way u feel. That□s the way its gota b
20	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/ú1.20 POBOXox36504W45WQ 16+
21	ham	Is that seriously how you spell his name?

The dataset has 5574 rows, i.e. 5574 data entries. The SMS are classified as 'ham' and 'spam'. 'Spam' means that the SMS is spam and 'ham' means that the SMS is spam.

Code-

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

In [ ]:
```
messages = pd.read_csv('SMSSpamCollection', sep='\t',
                           names=["label", "message"])
```

In [ ]:
```
messages
```

Out[ ]:

	label	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 × 2 columns

In [ ]:
```
import re
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```
Out[ ]:
```
True
```

In [ ]:
```
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
corpus = []
```

In [ ]:
```
for i in range(0, len(messages)):
    review = re.sub('[^a-zA-Z]', ' ', messages['message'][i])
    review = review.lower()
    review = review.split()

    review = [ps.stem(word) for word in review if not word in stopwords.words('english')]
    review = ' '.join(review)
    corpus.append(review)
```

In [ ]:
```
corpus
```

In [ ]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000)
X = cv.fit_transform(corpus).toarray()
```

In [ ]:
```
len(X)
```

Out[ ]:
```
5572
```

In [ ]:
```
y = pd.get_dummies(messages['label'])
y =y.iloc[:,1].values
y
```

Out[ ]:
```
array([0, 0, 1, ..., 0, 0, 0], dtype=uint8)
```

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

```
5572
```

In [ ]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
```

In [ ]:
```
from sklearn.naive_bayes import MultinomialNB
spam_detect_model = MultinomialNB().fit(X_train, y_train)
```

In [ ]:
```
y_pred = spam_detect_model.predict(X_test)
```

In [ ]:
```
from sklearn.metrics import confusion_matrix
confusion_m = confusion_matrix(y_test, y_pred)
confusion_m
```

Out[ ]:
```
array([[946,   9],
       [  8, 152]])
```

In [ ]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

Out[ ]:
```
0.9847533632286996
```

Output-

In [ ]:
```
from sklearn.metrics import confusion_matrix
confusion_m = confusion_matrix(y_test, y_pred)
confusion_m
```

Out[ ]:
```
array([[946,   9],
       [  8, 152]])
```

In [ ]:
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test, y_pred)
```

Out[ ]:
```
0.9847533632286996
```

Result- Hence NLP is implemented to solve a SMS spam classifier.