```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import nltk
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from google.colab import files
import io
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
```

```
df=files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell

```
df=pd.read_csv("/content/spam.csv",encoding="latin")
df.head()
```

	v1	v2	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	ham	Go until jurong point, crazy.. Available only ...	NaN	NaN	NaN
1	ham	Ok lar... Joking wif u oni...	NaN	NaN	NaN
2	spam	Free entry in 2 a wkly	NaN	NaN	NaN

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   v1          5572 non-null   object
 1   v2          5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

```
df.isna().sum()
```

```
v1               0
v2               0
Unnamed: 2    5522
Unnamed: 3    5560
Unnamed: 4    5566
dtype: int64
```

```
df.rename({"v1":"label","v2":"Text"},inplace=True,axis=1)
df.tail()
```

	label	Text	Unnamed: 2	Unnamed: 3	Unnamed: 4
5567	spam	This is the 2nd time we have tried 2 contact u...	NaN	NaN	NaN
5568	ham	Will Ì_ b going to esplanade fr home?	NaN	NaN	NaN

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
```

```
---  ------      --------------  -----
 0   label       5572 non-null   object
 1   Text        5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: object(5)
memory usage: 217.8+ KB
```

```
df.shape
```

```
(5572, 5)
```

```
df.ndim
```

```
2
```

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['label'] = le.fit_transform(df['label'])
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5572 entries, 0 to 5571
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   label       5572 non-null   int64
 1   Text        5572 non-null   object
 2   Unnamed: 2  50 non-null     object
 3   Unnamed: 3  12 non-null     object
 4   Unnamed: 4  6 non-null      object
dtypes: int64(1), object(4)
memory usage: 217.8+ KB
```

```
nltk.download("stopwords")
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
import nltk
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
```

```
import re
corpus = []
length = len(df)
```

```
for i in range(0,length):
  text = re.sub("[^a-zA-Z0-9]"," ",df["Text"][i])
  text = text.lower()
  text = text.split()
  pe = PorterStemmer()
  stopword = stopwords.words("english")
  text = [pe.stem(word) for word in text if not word in set(stopword)]
  text = " ".join(text)
  corpus.append(text)
```

```
corpus
```

```
['go jurong point crazi avail bugi n great world la e buffet cine got amor wat',
 'ok lar joke wif u oni',
 'free entri 2 wkli comp win fa cup final tkt 21st may 2005 text fa 87121 receiv entri question std txt rate c appli
08452810075over18',
 'u dun say earli hor u c alreadi say',
 'nah think goe usf live around though',
 'freemsg hey darl 3 week word back like fun still tb ok xxx std chg send 1 50 rcv',
 'even brother like speak treat like aid patent',
 'per request mell mell oru minnaminungint nurungu vettam set callertun caller press 9 copi friend callertun',
 'winner valu network custom select receivea 900 prize reward claim call 09061701461 claim code kl341 valid 12 hour',
 'mobil 11 month u r entitl updat latest colour mobil camera free call mobil updat co free 08002986030',
 'gonna home soon want talk stuff anymor tonight k cri enough today',
 'six chanc win cash 100 20 000 pound txt csh11 send 87575 cost 150p day 6day 16 tsandc appli repli hl 4 info',
 'urgent 1 week free membership 100 000 prize jackpot txt word claim 81010 c www dbuk net lccltd pobox 4403ldnw1a7rw18',
 'search right word thank breather promis wont take help grant fulfil promis wonder bless time',
 'date sunday',
 'xxxmobilemovieclub use credit click wap link next txt messag click http wap xxxmobilemovieclub com n qjkgighjjgcbl',
 'oh k watch',
 'eh u rememb 2 spell name ye v naughti make v wet',
 'fine way u feel way gota b',
 'england v macedonia dont miss goal team news txt ur nation team 87077 eg england 87077 tri wale scotland 4txt 1 20
poboxox36504w45wq 16',
 'serious spell name',
 'go tri 2 month ha ha joke',
 'pay first lar da stock comin',
 'aft finish lunch go str lor ard 3 smth lor u finish ur lunch alreadi',
 'ffffffffff alright way meet sooner',
 'forc eat slice realli hungri tho suck mark get worri know sick turn pizza lol',
 'lol alway convinc',
 'catch bu fri egg make tea eat mom left dinner feel love',
 'back amp pack car let know room',
 'ahhh work vagu rememb feel like lol',
 'wait still clear sure sarcast x want live us',
 'yeah got 2 v apologet n fallen actin like spoilt child got caught till 2 go badli cheer',
 'k tell anyth',
 'fear faint housework quick cuppa',
 'thank subscript rington uk mobil charg 5 month pleas confirm repli ye repli charg',
 'yup ok go home look time msg xuhui go learn 2nd may lesson 8am',
 'oop let know roommat done',
 'see letter b car',
 'anyth lor u decid',
 'hello saturday go text see decid anyth tomo tri invit anyth',
 'pl go ahead watt want sure great weekend abiola',
 'forget tell want need crave love sweet arabian steed mmmmmm yummi',
 '07732584351 rodger burn msg tri call repli sm free nokia mobil free camcord pleas call 08000930705 deliveri tomorrow',
 'see',
 'great hope like man well endow lt gt inch',
 'call messag miss call',
 'get hep b immunis nigeria',
 'fair enough anyth go',
 'yeah hope tyler could mayb ask around bit',
 'u know stubborn even want go hospit kept tell mark weak sucker hospit weak sucker',
 'think first time saw class',
 'gram usual run like lt gt half eighth smarter though get almost whole second gram lt gt',
 'k fyi x ride earli tomorrow morn crash place tonight',
 'wow never realiz embarass accomod thought like sinc best could alway seem happi cave sorri give sorri offer sorri room
embarass',
 'sm ac sptv new jersey devil detroit red wing play ice hockey correct incorrect end repli end sptv',
```

```
from sklearn.feature_extraction.text import CountVectorizer
CV = CountVectorizer(max_features=35000)
X = CV.fit_transform(corpus).toarray()
```

```
import pickle
pickle.dump(CV, open('cv1.pk1', 'wb'))
```

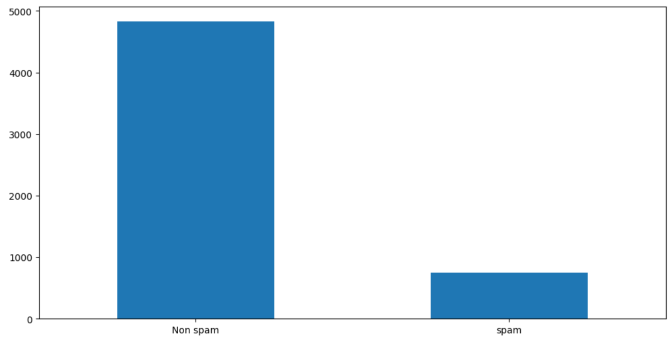
```
df.describe()
```

	label
count	5572.000000
mean	0.134063
std	0.340751
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000

```
df.shape
```

```
(5572, 5)
```

```
df["label"].value_counts().plot(kind="bar" ,figsize=(12,6))
plt.xticks(np.arange(2),('Non spam', 'spam'),rotation=0);
```



```
Y=pd.get_dummies(df['label'])
Y=Y.iloc[:,1].values
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.20,random_state =0)
```

```
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB()

model.fit(X_train,Y_train)
```

```
▾ MultinomialNB
MultinomialNB()
```

```
y_pred=model.predict(X_test)
y_pred
```

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

```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(Y_test, y_pred)
score = accuracy_score(Y_test,y_pred)
print(cm)
print('Accuracy score Is:-' ,score*100)
```

```
[[941   8]
 [  9 157]]
Accuracy score Is:- 98.47533632286995
```

```
import pickle
pickle.dump(model, open('spam.pk1', 'wb'))
```

```
loaded_model=pickle.load(open('spam.pk1', 'rb'))
loaded_model.predict(X_test)
loaded_model.score(X_test, Y_test)
```

```
0.9847533632286996
```

```
from hashlib import new
def new_review(new_review):
  new_review = new_review
  new_review = re.sub('[^a-zA-Z]', ' ',new_review)
  new_review = new_review.lower()
  new_review = new_review.split()
  ps  = PorterStemmer()
  all_stopwords = stopwords.words('english')
  all_stopwords.remove('not')
  new_review = [ps.stem(word) for word in new_review if not word in set(all_stopwords)]
  new_review = ' '.join(new_review)
  new_corpus = [new_review]
  new_X_test = CV.transform(new_corpus).toarray()
  new_Y_pred = loaded_model.predict(new_X_test)
  return new_Y_pred
  new_review = new_review(str(input("Enter new review...")))
  if new_review[0]==1:
    print("SPAM")
  else :
    print("NOT SPAM")
```

```
model.save('spam.h5')
```

```
from sklearn.svm import SVC
svm1=SVC(kernel='rbf')
svm1.fit(X_train,Y_train)
```

```
▾ SVC
SVC()
```

```
Y_pred4=svm1.predict(X_test)
from sklearn.metrics import accuracy_score
svm_rbf=accuracy_score(Y_test,Y_pred4)
svm_rbf
```

```
0.9730941704035875
```

```
svm2=SVC(kernel='sigmoid')
svm2.fit(X_train,Y_train)
```

```
▾         SVC
SVC(kernel='sigmoid')
```

```
Y_pred5=svm2.predict(X_test)
from sklearn.metrics import accuracy_score
svm_sig=accuracy_score(Y_test,Y_pred5)
svm_sig
```

```
0.9757847533632287
```

```
from sklearn.tree import DecisionTreeClassifier
dt=DecisionTreeClassifier()
dt.fit(X_train,Y_train)
```

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

```
Y_pred6=dt.predict(X_test)
from sklearn.metrics import accuracy_score
dec_tree=accuracy_score(Y_test,Y_pred6)
dec_tree
```

```
0.9748878923766816
```

```
models = pd.DataFrame({
    'Model': [ 'MultinomialNB','SVM-rbf','SVM-sigmoid','Decision Tree'],
    'Test Score' : [score,svm_rbf,svm_sig,dec_tree,]})
models.sort_values(by='Test Score',ascending=False)
```

	Model	Test Score
0	MultinomialNB	0.984753
2	SVM-sigmoid	0.975785
3	Decision Tree	0.974888
1	SVM-rbf	0.973094

New Section

New Section

● ×