```
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
```

```
df = pd.read_csv('/content/sample_data/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 comp to win FA Cup fina...	NaN	NaN	NaN
3	ham	U dun say so early hor... U c already then say...	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

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

```
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 =
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-9-06dc39eeeaae> in <cell line: 2>()
      1 from sklearn.model_selection import train_test_split
----> 2 X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.20,
random_state = 0)

NameError: name 'X' is not defined
```

SEARCH STACK OVERFLOW

```
print("Before OverSampling, counts of label '1': {}".format(sum(y_train == 1)))
print("Before OverSampling, counts of label '0': {} \n".format(sum(y_train == 0)))
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state = 2)
X_train_res, y_train_res = sm.fit_resample(X_train, y_train.ravel())
print('After OverSampling, the shape of train_X: {}'.format(X_train_res.shape))
print('After OverSampling, the shape of train_y: {} \n'.format(y_train_res.shape))
print("After OverSampling, counts of label '1': {}".format(sum(y_train_res == 1)))
print("After OverSampling, counts of label '': {}".format(sum(y_train_res == 0)))
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-10-eac4e9b0f9ac> in <cell line: 1>()
----> 1 print("Before OverSampling, counts of label '1': {}".format(sum(y_train ==
1)))
      2 print("Before OverSampling, counts of label '0': {} \n".format(sum(y_train
== 0)))
      3 from imblearn.over_sampling import SMOTE
      4 sm = SMOTE(random_state = 2)
      5 X_train_res, y_train_res = sm.fit_resample(X_train, y_train.ravel())

NameError: name 'y_train' is not defined
```

```
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)
```

```
  File "<ipython-input-14-588c02c481f6>", line 2
    text = re.sub("[^a-zA-Z0-9]"," ",df["text"][i])
    ^
IndentationError: expected an indented block
```

SEARCH STACK OVERFLOW

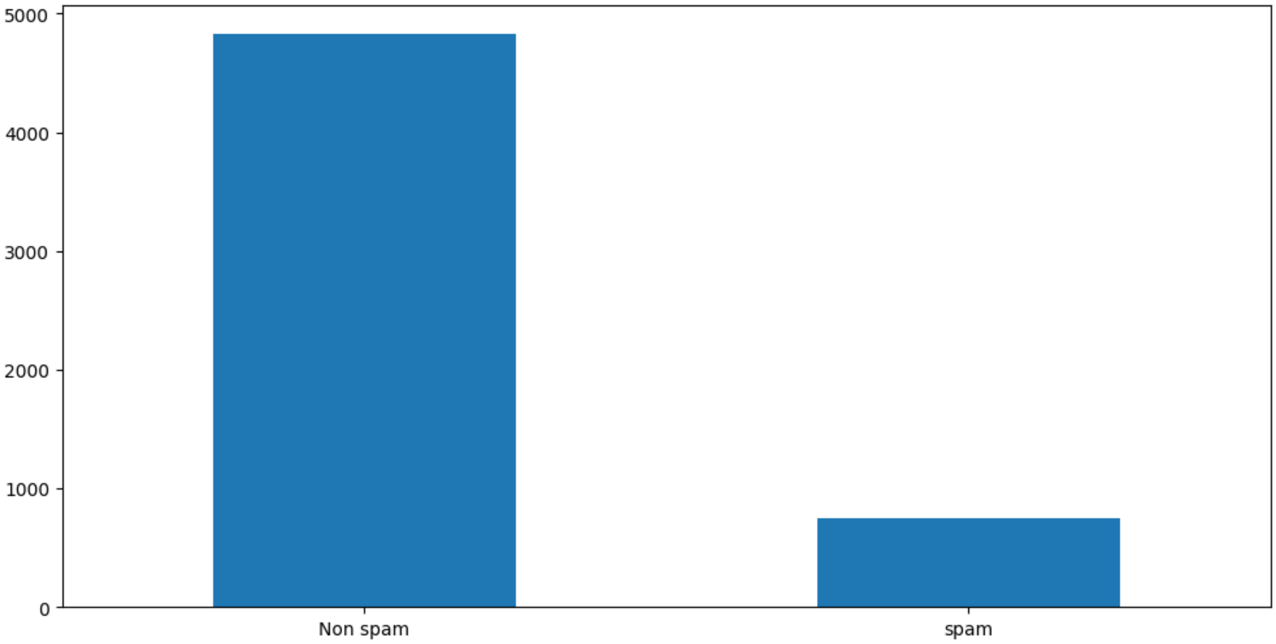
```
from sklearn.feature_extraction.text import CountVectorizer
cv=CountVectorizer(max_features=35000)
x=cv.fit_transform(corpus).toarray()
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-16-e89d879d2a0a> in <cell line: 3>()
      1 from sklearn.feature_extraction.text import CountVectorizer
      2 cv=CountVectorizer(max_features=35000)
----> 3 x=cv.fit_transform(corpus).toarray()

                              1 frames

/usr/local/lib/python3.9/dist-packages/sklearn/feature_extraction/text.py in
_count_vocab(self, raw_documents, fixed_vocab)
   1292             vocabulary = dict(vocabulary)
   1293             if not vocabulary:
-> 1294                 raise ValueError(
   1295                     "empty vocabulary; perhaps the documents only contain
stop words"
   1296                 )

ValueError: empty vocabulary; perhaps the documents only contain stop words
```

SEARCH STACK OVERFLOW

```
import pickle
pickle.dump(cv, open('cv1.pkl', 'wb'))


df["label"].value_counts().plot(kind="bar", figsize=(12,6))
```

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



Colab paid products - Cancel contracts here

✓ 0s completed at 12:03 PM