GMAIL SPAM DETECTION WITH NATURAL LANGUAGE PROCESSING

OBJECTIVE

The objective of this project is to build a prediction model to predict whether a mail is spam or not by NLP

NATURAL LANGUAGE PROCESSING

- **Natural Language Processing (NLP) is the study of making computers understand how humans naturally speak, write and communicate.**

- We will be using Python library NLTK (Natural Language Toolkit) for doing natural language processing in English Language. The **Natural language toolkit (NLTK)** is a collection of Python libraries designed especially for identifying and tag parts of speech found in the text of natural language like English.

REQUIREMENTS :

- **Install NLTK**

```
>>>pip install nltk
```

- **Import NTLK**

```
>>>import nltk
```

- **Download NTLK**

```
>>>nltk.download()
```

SAMPLE DATA (SPAM.CSV)

	A	B
1	type	text
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	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
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	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, Â£1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	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
10	spam	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	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
12	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.
13	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
14	spam	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	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 and a blessing at all time
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if that's the way u feel. That's the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ãº1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	Iâ€˜m going to try for 2 months ha ha only joking
24	ham	So Ã¼ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?

STEPS INVLOVED

1. Importing Dataset
2. Preprocessing Dataset
3. Vectorization
4. Training and Classification
5. Model Evaluation

01 IMPORTING DATASET

REQUIREMENTS

- PANDAS LIBRARY (python)
- DATASET of restaurant review (spam.csv)

DATASET DETAILS:

- **Columns** : total 2

1)type

2)text

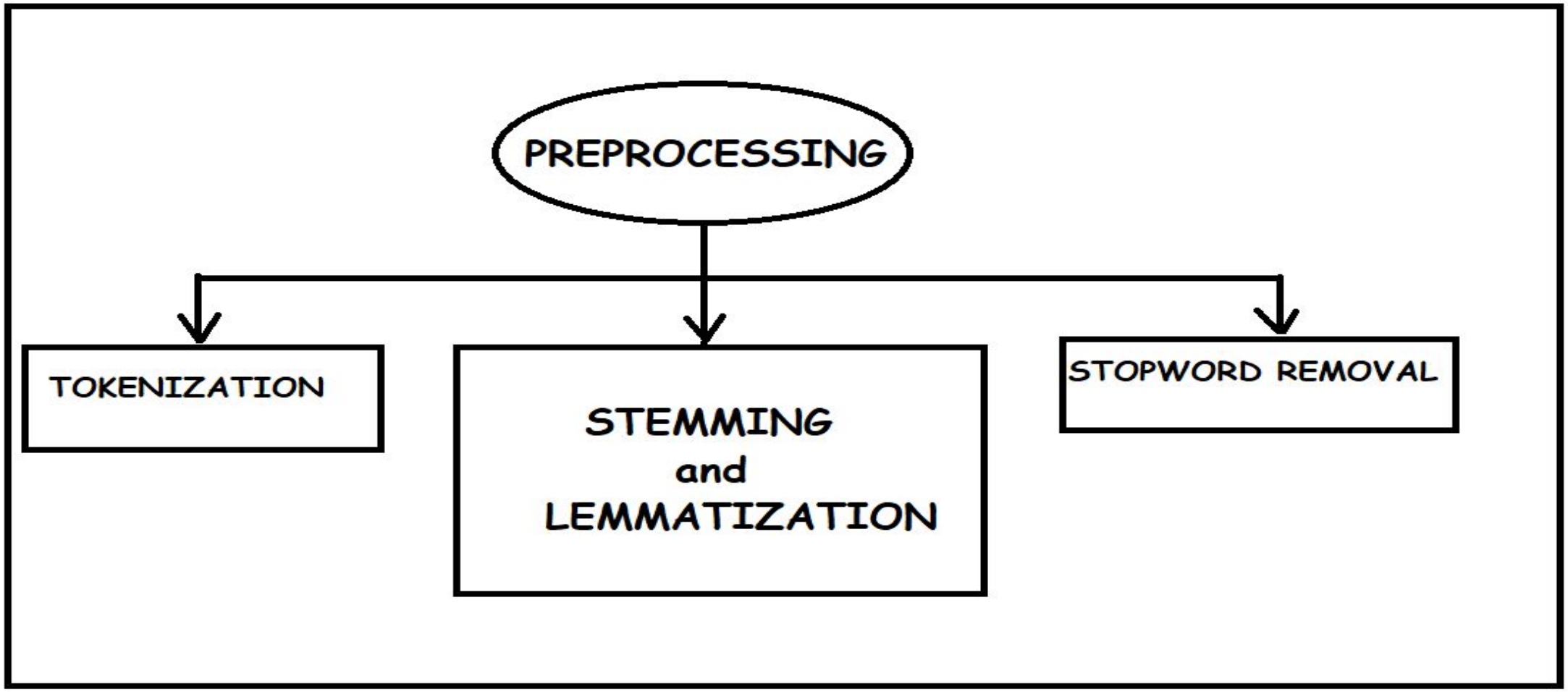
PANDAS :

Initially we need to import pandas library to import the reviews data set .

For this we use:

```
>>>import pandas as pd
>>>data =pd.read_csv(spam.csv)
```

02 PREPROCESSING DATASET



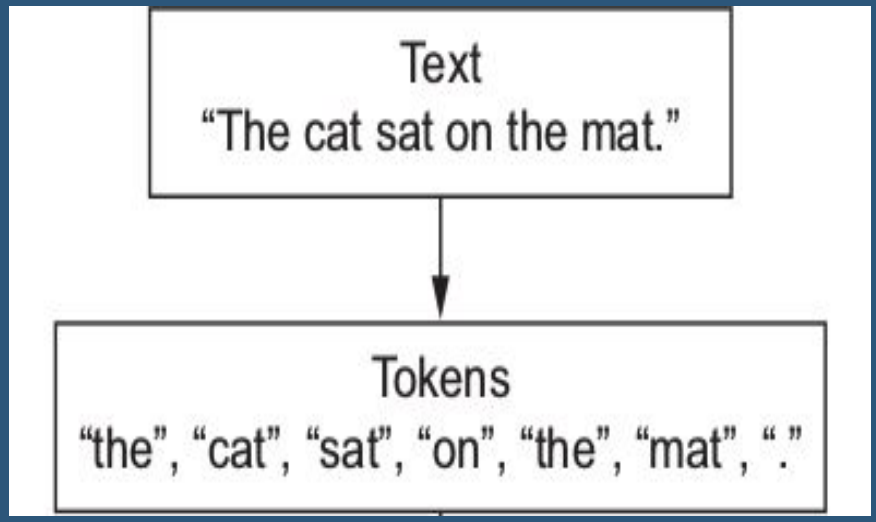
TOKENIZATION

It is splitting up of data into tokens(comma separated values)

Natural Language toolkit has very important module **tokenize** which further comprises of sub-modules that is sentence tokenize and word tokenize we use word tokenize.

REQUIREMENTS :

- word_tokenize method from NLTK



STEMMING

This is the idea of removing the suffix of a word and reducing different forms of a word to a core root.

REQUIREMENTS :

- **stemmer from nltk.stem package**

There are different stemmers in this package

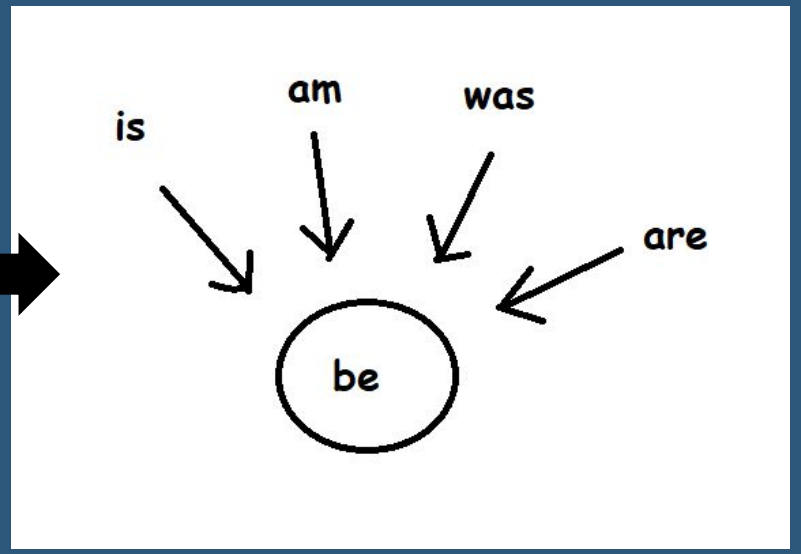
1) Snowball
2) Porter
3) Lancaster

LEMMITIZATION

Lemmatization is the algorithmic process of finding the lemma of a word depending on their meaning. Lemmatization aims to remove inflectional endings. It helps in returning the base or dictionary form of a word, which is known as the lemma. The major difference between stemming and lemmatizing is stemming can often create non-existent words, whereas lemmas are actual words.

REQUIREMENTS :

- **Lemmitizer from nltk.stem**



STOPWORDS REMOVAL

A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore.

REQUIREMENTS :

- Import stopwords from nltk.corpus

>>>from nltk.corpus import stopwords
>>stopwords.words(‘english’)

```
from nltk.corpus import stopwords
nltk.download('stopwords')

language = "english"
stop_words = set(stopwords.words(language))
print(stop_words)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/shubham/nltk_data...
```

```
{"hasn't", 'hers', 'do', 'is', 'in', 'on', 'been', 'does', 'mightn', "you'd", "couldn't", 'co
uldn', "she's", 'it', 'i', 'own', "mightn't", 'which', 'have', 'or', 'themselves', "needn't",
'has', 'when', 'his', 'further', 'off', 'our', 'of', 'how', "hadn't", 'any', 'are', 'very',
'them', 'into', 'same', 'isn', 'because', 'd', 'wasn', 're', 'each', 'an', 'after', 'agains
t', 'until', 'don', 'll', 'they', 'while', 'under', 'had', 'with', 'here', 'just', "didn't",
'only', 'not', 'now', "you'll", 'having', 'ourselves', 'did', 'hasn', "haven't", 'the', "yo
u're", 't', 'so', 'he', 'too', 'we', 'once', 'hadn', 'that', 'these', "shan't", 'doesn', 'wo
n', 'aren', 'between', "it's", 'few', 'haven', 'she', 'ma', 'by', 've', "mustn't", 'above',
'nor', 'my', 'ain', 'as', 'ours', 'her', 'but', 'most', 'some', 'for', 'a', 'yours', 'shan',
'where', "weren't", "don't", 'why', 'up', 'about', 'and', 'no', 'mustn', 'you', 'herself', 't
heirs', 'again', 'can', 's', 'should', 'through', 'their', 'him', 'wouldn', 'such', 'both',
'if', 'whom', 'myself', 'yourselves', 'what', "you've", 'itself', 'over', 'y', 'from', 'you
r', 'am', 'will', 'those', "doesn't", 'weren', 'be', 'then', 'to', 'yourself', "that'll", 'it
s', 'himself', 'other', "wasn't", 'didn', 'this', "aren't", 'below', 'shouldn', 'was', 'o',
'who', "shouldn't", 'at', 'during', "won't", 'all', 'needn', 'than', "isn't", "wouldn't", 'do
ing', 'were', 'being', 'me', 'down', "should've", 'more', 'm', 'there', 'before', 'out'}
```

```
[nltk_data]   Unzipping corpora/stopwords.zip.
```

04 VECTORIZATION

The vectorization is a technique used to convert textual data to numerical format. Using vectorization, a matrix is created where each column represents a feature and each row represents an individual review.

We have two ways for vectorization :

TF (Term Frequency)

Term Frequency is defined as how frequently the word appear

in the document .

$$TF = \frac{No\ of\ time\ word\ appear\ in\ the\ document}{Total\ no\ of\ word\ in\ the\ document}$$

Term Frequency-Inverse Document Frequency(TF-IDF)

TD-IDF basically tells importance of the word in the corpus or dataset

- It is the combination of Term frequency and **Inverse Document Frequency** .
- Inverse Document frequency is another concept which is used for finding out importance of the word. It is based on the fact that less frequent words are more informative and important.

Requirements:

- **Import TfidfVectorizer from sklearn**

IDF(t) = log_e(Total number of documents / Number of documents with term t in it)

WI,J = TF I,J X IDF

- WI,J=weight which signifies how important a word is for individual text message
- TF I,J =(no. of times I occurred)/(total no. of terms in J)

05 TRAINING AND CLASSIFICATION

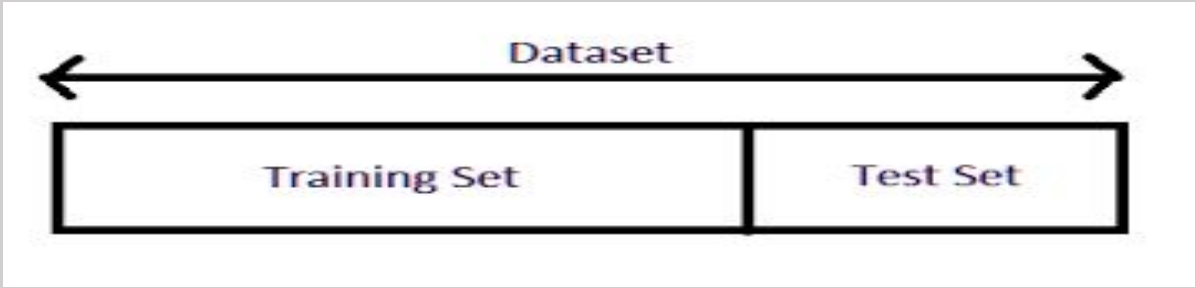
TRANING DATA:

Further the data is splitted into training and testing set based on size of the dataset .

REQUIREMENTS:

- Import train_test_split from sklearn.model_selection

```
>>> from sklearn.model_selection import train_test_split
>>>X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state =0)
```



CLASSIFICATION:

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

REQUIREMENTS:

- Import Logistic regression from sklearn.linear_model
- from sklearn.svm import LinearSVC

05 MODEL EVALUTAION

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

Classification Accuracy:

Accuracy is a common evaluation metric for classification problems.
It's the number of correct predictions made as a ratio of all predictions made.

REQUIREMENTS:

- import accuracy_score from sklearn.metrics

```
>>>result = model.score(X_test, y_test)
```