Experiment No - 01

Aim- To perform pre-processing of text (tokenization filtration, scopt validation,stop-words Removal, stemming etc.).

Theory:

Data preprocessing is a fundamental step while building a machine learning model.If the data is fairly pre-processed the result would be reliable.In NLP , the first step before building the machine learning model,is to preprocess.The data,lets see the various different steps that are followed while preprocessing the data also used for dimensionality reduction.

1. Tokenization.

2. Lower casing

3. Stop words removal

4. Stemming

5. Lemmatization

Each term is the axis in the vector. Space model. In multi-dimensional space, the text or document are constituted as vectors. The number of different words represents the number of dimensions.

The python library that is used to do the Par-processing tasks in NLP is nitk. You can install the nitk package using "pip install nltk

•Tokenization :- It is a method in which sentences are converted into words.

•Lowercasing:- The tokenized words converted into lower case format. (NLU- nlp), words having the same meaning like hip and NIP if they do not consist into lowercase then these both will Constitute as non-identical words in the vector space.

• stop words removal: These are the roost often used that do not have any significance while determining the two different documents like (a, an, the, etc) so they are to be remound, check the below image where from the sentence Introduction to Natural Language Processing" The "to" word is removed.

•Without removing stop words: get to see so tokens without removing stop-words, Now we shall remove stop words.

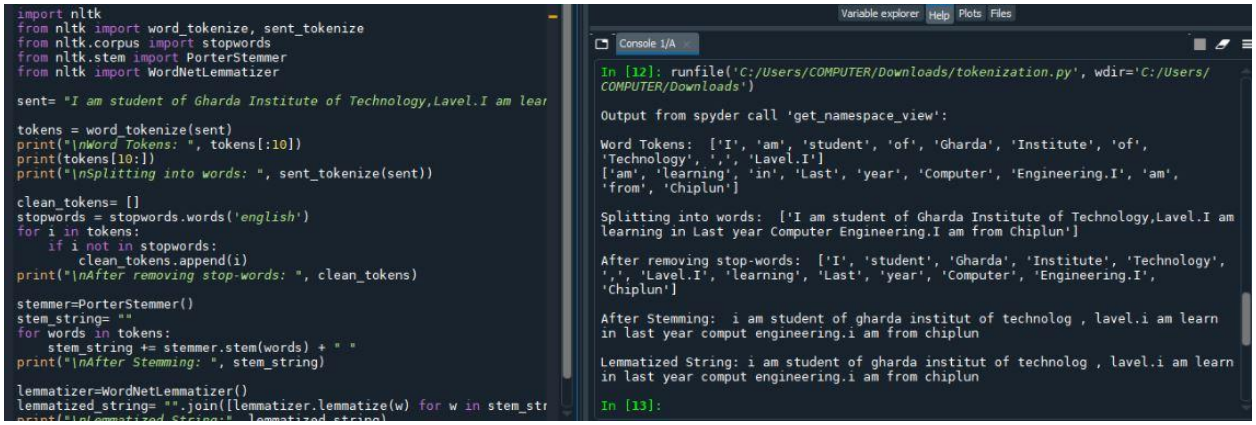
•stemmings:- It is the process in which the words. converted to its base from.

•Lemmatizations :-Different from stemming, lemmatization Lowers the words to word in the present language fox eq check the below image where word het and are changed to ha be respectively.

Code:-

```
import nltk
from nltk import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk import WordNetLemmatizer

sent= "I am student of Gharda Institute of Technology,Lavel.I am learning in Last
year Computer Engineering.I am from Chiplun"

tokens = word_tokenize(sent)
print("\nWord Tokens: ", tokens[:10])
print(tokens[10:])
print("\nSplitting into words: ", sent_tokenize(sent))

clean_tokens= []
stopwords = stopwords.words('english')
for i in tokens:
    if i not in stopwords:
        clean_tokens.append(i)
print("\nAfter removing stop-words: ", clean_tokens)

stemmer=PorterStemmer()
stem_string= ""
for words in tokens:
    stem_string += stemmer.stem(words) + " "
print("\nAfter Stemming: ", stem_string)

lemmatizer=WordNetLemmatizer()
lemmatized_string= "".join([lemmatizer.lemmatize(w) for w in stem_string])
print("\nLemmatized String:", lemmatized_string)
```

Output:-



```
import nltk
from nltk import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer
from nltk import WordNetLemmatizer

sent= "I am student of Gharda Institute of Technology,Lavel.I am lear

tokens = word_tokenize(sent)
print("\nWord Tokens: ", tokens[:10])
print(tokens[10:])
print("\nSplitting into words: ", sent_tokenize(sent))

clean_tokens= []
stopwords = stopwords.words('english')
for i in tokens:
    if i not in stopwords:
        clean_tokens.append(i)
print("\nAfter removing stop-words: ", clean_tokens)

stemmer=PorterStemmer()
stem_string= ""
for words in tokens:
    stem_string += stemmer.stem(words) + " "
print("\nAfter Stemming: ", stem_string)

lemmatizer=WordNetLemmatizer()
lemmatized_string= "".join([lemmatizer.lemmatize(w) for w in stem_str
```

```
Variable explorer  Help  Plots  Files

Console 1/A

In [12]: runfile('C:/Users/COMPUTER/Downloads/tokenization.py', wdir='C:/Users/
COMPUTER/Downloads')

Output from spyder call 'get_namespace_view':

Word Tokens:  ['I', 'am', 'student', 'of', 'Gharda', 'Institute', 'of',
'Technology', ',', 'Lavel.I']
['am', 'learning', 'in', 'Last', 'year', 'Computer', 'Engineering.I', 'am',
'from', 'Chiplun']

Splitting into words:  ['I am student of Gharda Institute of Technology,Lavel.I am
learning in Last year Computer Engineering.I am from Chiplun']

After removing stop-words:  ['I', 'student', 'Gharda', 'Institute', 'Technology',
',', 'Lavel.I', 'learning', 'Last', 'year', 'Computer', 'Engineering.I',
'Chiplun']

After Stemming:  i am student of gharda institut of technolog , lavel.i am learn
in last year comput engineering.i am from chiplun

Lemmatized String: i am student of gharda institut of technolog , lavel.i am learn
in last year comput engineering.i am from chiplun

In [13]:
```

Conclusion :- Thus I have performed pre-processing of text.