NATURAL LANGUAGE PROCESSING

NLP

PHASE-0 - TEXT PREPROCESSING - 1

Natural Language Processing

What is NLP ?

→ NLP is a part of AI that helps computers understand, read and respond to human language. (like English, Hindi etc)

TOKENIZATION

Tokenization is the process of breaking text into small pieces, called tokens.

OR

Tokenization is the process to convert either a paragraph or a sentence into tokens.

Topics :-

1. Corpus means → Paragraph
2. Documents means → Sentences
3. Vocabulary means → Unique Words
4. Words

Word Tokenization

Break text into words

Sentence Tokenization

Break text into Sentences.

Character Tokenization

Break text into characters.

Example :-

Corpus (paragraph) { "My name is Ritik. I am a python developer"

⇓

Tokens { Sentences }

⇓

Documents (Sentences) { "My name is Ritik"
"I am a Python Developer"

⇓

Tokens { Words }

Words { 'My', 'Name', 'is'

Example :-

" I like to drink Apple Juice. My friend likes Mango Juice. "

=> For unique words, here we have total 11 words

But if we ask for Unique words

=> 10 words

because "juice" came twices.

Vocabulary → Unique Words

TOKENIZATION CODE EXAMPLE :-

```
''' bash
    pip install nltk
'''
```

installing module nltk (Natural language Toolkit)

```
corpus = " Iam Ritik . I am a python developer!
          I am using Jupyter Notebook."
```

creatin corpus (paragraph)

sent_tokenize ->

use to split paragraph into sentences

```
from nltk.tokenize import sent_tokenize

document = sent_tokenize (corpus)

print (document)
```

['I am Ritik.', 'I am Python Developer', 'I am using Jupyter Notebook.']

```
print ( type(document))   # list
```

Note :- Output will be a list type.
:- It uses "." and "!" to split the paragraph and create sentences as elements of list.

```
for i in document :-
    print (i)
                    # I am Ritik.
# printing elements of    I am Python developer!
  list.                 I am using Jupyter Notebook.
```

```
# word_tokenize
    use to split sentence into words

from nltk.tokenize import word_tokenize

word = word.tokenize (corpus)

Print (word)

# ['I', 'am', 'Python' ..... 'Jupter', 'Notebook', '.']
```

Note :- Output will be list type.

```
for i in word :
    print(word)
```

```
# I
am
Ritik
.
.
.
Notebook
.
```

—x—x—x—x—x—x—x—x—x—x—

Extras :- TreebankWordTokenizer

```
from nltk.tokenize impor TreebankWordTokenizer

tokenizer = TokenizerI()

print( tokenizer.tokenize (corpus))
```

Output : ['I', 'am', 'Ritik.', 'I', 'am'.....
.....'Jupyter', 'Notebook', '.']

Note :- It only splits the last (.) as element and the rest attached along with words.
eg:- 'Ritik.'