Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens. If the text is split into words, then its called as 'Word Tokenization' and if it's split into sentences then its called as 'Sentence Tokenization'.

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
corpus = """
JavaScript is the world's most popular programming language. JavaScript! is the programming language of the Web. JavaScript is easy to le
"""
```

```
# Paragraph -------------> Sentence
```

```
from nltk.tokenize import sent_tokenize
```

```
sent_tokenize(corpus)
```

```
["\nJavaScript is the world's most popular programming language.",
 'JavaScript!',
 'is the programming language of the Web.',
 'JavaScript is easy to learn.']
```

```
# Paragraph -------------> word
```

```
from nltk.tokenize import word_tokenize
```

```
word_tokenize(corpus)
```

```
['JavaScript',
 'is',
 'the',
 'world',
 "'s",
 'most',
 'popular',
 'programming',
 'language',
 '.',
 'JavaScript',
 '!',
 'is',
 'the',
 'programming',
 'language',
 'of',
 'the',
 'Web',
 '.',
 'JavaScript',
 'is',
 'easy',
 'to',
 'learn',
 '.']
```

```
len(word_tokenize(corpus))
```

```
26
```

```
from nltk.tokenize import wordpunct_tokenize
```

```
wordpunct_tokenize(corpus)
```

```
['JavaScript',
 'is',
     'the',
     'world',
     "'",
     's',
     'most',
     'popular',
     'programming',
     'language',
     '.',
     'JavaScript',
     '!',
     'is',
     'the',
     'programming',
     'language',
     'of',
     'the',
     'Web',
     '.',
     'JavaScript',
     'is',
     'easy',
     'to',
     'learn',
     '.']
```

```
len(wordpunct_tokenize(corpus))
```

```
    27
```

```
from nltk.tokenize import TreebankWordTokenizer
```

```
tokenizer=TreebankWordTokenizer()
```

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

```
    ['JavaScript',
     'is',
     'the',
     'world',
     "'s",
     'most',
     'popular',
     'programming',
     'language.',
     'JavaScript',
     '!',
     'is',
     'the',
     'programming',
     'language',
     'of',
     'the',
     'Web.',
     'JavaScript',
     'is',
     'easy',
     'to',
     'learn',
     '.']
```

```
len(tokenizer.tokenize(corpus))
```

```
    24
```