LAB 1-TOKENIZATION

2147126-Rajesh PV

```
import nltk

nltk.download();

from nltk.tokenize import sent_tokenize, word_tokenize,wordpunct_tokenize,TweetTokenizer,TreebankWordTokenizer,MWETokenizer


text_data="In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer"
```

1.**sentence tokenizer**

```
print(sent_tokenize(text_data));
```

output:

```
 ['In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths
globally.', "As of the end of 2020, there were 7.8 million women alive who were diagnosed
with breast cancer in the past 5 years, making it the world's most prevalent cancer"]
```

2.**word tokenizer**

```
print(word_tokenize(text_data));
```

output:

```
['In', '2020', ',', 'there', 'were', '2.3', 'million', 'women', 'diagnosed', 'with', 'breast', 'cancer',
'and', '685', '000', 'deaths', 'globally', '.', 'As', 'of', 'the', 'end', 'of', '2020', ',', 'there', 'were', '7.8',
'million', 'women', 'alive', 'who', 'were', 'diagnosed', 'with', 'breast', 'cancer', 'in', 'the', 'past', '5',
'years', ',', 'making', 'it', 'the', 'world', "'s", 'most', 'prevalent', 'cancer']
```

3.**White space tokenization**

```
print(text_data.split());
```

output:

```
 ['In', '2020,', 'there', 'were', '2.3', 'million', 'women', 'diagnosed', 'with', 'breast', 'cancer', 'and',
'685', '000', 'deaths', 'globally.', 'As', 'of', 'the', 'end', 'of', '2020,', 'there', 'were', '7.8', 'million',
'women', 'alive', 'who', 'were', 'diagnosed', 'with', 'breast', 'cancer', 'in', 'the', 'past', '5', 'years,', 'making', 'it', 'the', "world's", 'most', 'prevalent', 'cancer’]
```

4. **Punctuation-based tokenizer**

```
print(wordpunct_tokenize(text_data));
```

```
output: ['In', '2020', ',', 'there', 'were', '2', '.', '3', 'million', 'women', 'diagnosed', 'with', 'breast', 'cancer', 'and', '685', '000', 'deaths', 'globally', '.', 'As', 'of', 'the', 'end', 'of', '2020', ',', 'there', 'were', '7', '.', '8', 'million', 'women', 'alive', 'who', 'were', 'diagnosed', 'with', 'breast', 'cancer', 'in', 'the', 'past', '5', 'years', ',', 'making', 'it', 'the', 'world', "'", 's', 'most', 'prevalent', 'cancer']
```

5.**Treebankword Tokenizer**

```
s = "They'll save and invest more."

TreebankWordTokenizer().tokenize(s)
```

```
Output : Out[33]: ['hi', ',', 'my', 'name', 'ca', "n't", 'hello', ',']
```

6.**TweetTokenizer**

```
tknzr = TweetTokenizer()

s0 = "This is a cooool #dummysmiley: :-) :-P <3 and some arrows < > -> <--"

print(tknzr.tokenize(s0))
```

```
Output : ['This', 'is', 'a', 'cooool', '#dummysmiley', ':', ':-)', ':-P', '<3', 'and', 'some', 'arrows', '<', '>', '->', '<--']
```

7.**MWET Tokenizer**

```
tokenizer = MWETokenizer([('a', 'little'), ('a', 'little', 'bit'), ('a', 'lot')])

tokenizer.add_mwe(('in', 'spite', 'of'))

tokenizer.tokenize('Testing testing testing one two three'.split())
```

```
Output : Out[10]: ['Testing', 'testing', 'testing', 'one', 'two', 'three']
```

```
tokenizer.tokenize('In a little or a little bit or a lot in spite of'.split())
```

```
Output : Out[11]: ['In', 'a_little', 'or', 'a_little_bit', 'or', 'a_lot', 'in_spite_of’
```