Library required for Preprocessing

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
```

```
import nltk
```

```
nltk.download()
```

```
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader> d

Download which package (l=list; x=cancel)?
  Identifier> punkt
    Downloading package punkt to /root/nltk_data...
      Unzipping tokenizers/punkt.zip.

---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
---------------------------------------------------------------------------
Downloader> q
True
```

Sentence Tokenization

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

```
text = '''Myself Chirag From VCET , From CSE(DS) department . Studying in Final Year'''
```

```
text
```

```
'Myself Chirag From VCET , From CSE(DS) department . Studying in Final Year'
```

```
sentences = sent_tokenize (text)
```

```
sentences
```

```
['Myself Chirag From VCET , From CSE(DS) department .',
 'Studying in Final Year']
```

Word Tokenization

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

```
words = word_tokenize (text)
```

```
words
```

```
['Myself',
 'Chirag',
 'From',
 'VCET',
 ',',
 'From',
 'CSE',
 '(',
 'DS',
 ')',
 'department',
 '.',
 'Studying',
 'in',
 'Final',
 'Year']
```

```
for w in words:
    print (w)
```

```
Myself
Chirag
From
VCET
,
From
CSE
(
DS
)
department
.
Studying
in
Final
Year
```

Levels of Sentences Tokenization using Comprehension

```
sent_tokenize (text)
```

```
['Myself Chirag From VCET , From CSE(DS) department .',
 'Studying in Final Year']
```

```
[word_tokenize (text) for t in sent_tokenize(text)]
```

```
[['Myself',
  'Chirag',
  'From',
  'VCET',
  ',',
  'From',
  'CSE',
  '(',
  'DS',
  ')',
  'department',
  '.',
  'Studying',
  'in',
  'Final',
  'Year'],
 ['Myself',
  'Chirag',
  'From',
  'VCET',
  ',',
  'From',
  'CSE',
  '(',
  'DS',
  ')',
  'department',
  '.',
  'Studying',
  'in',
  'Final',
  'Year']]
```

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

```
wordpunct_tokenize (text)
```

```
['Myself',
 'Chirag',
 'From',
 'VCET',
 ',',
 'From',
 'CSE',
 '(',
 'DS',
 ')',
 'department',
 '.',
 'Studying',
 'in',
 'Final',
 'Year']
```

Filteration of Text by converting into lower case

```
text.lower()
```

```
'myself chirag from vcet , from cse(ds) department . studying in final year'
```

```
text.upper()
```

```
'MYSELF CHIRAG FROM VCET , FROM CSE(DS) DEPARTMENT . STUDYING IN FINAL YEAR'
```