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

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

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

```
text="Are you curious about tokenization?"
```

```
sent_tokenize_list=sent_tokenize(text)
print(sent_tokenize_list)
```

```
['Are you curious about tokenization?']
```

```
pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/d
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-package
```

```
from nltk.tokenize import word_tokenize

print("\nWord tokenizer:")
print(word_tokenize(text))
```

```
Word tokenizer:
['Are', 'you', 'curious', 'about', 'tokenization', '?']
```

```
from nltk.tokenize import WordPunctTokenizer
word_punct_tokenizer = WordPunctTokenizer()
print ("\nWord punct tokenizer:")
print (word_punct_tokenizer.tokenize(text))
```

```
Word punct tokenizer:
['Are', 'you', 'curious', 'about', 'tokenization', '?']
```

Stemming text **data**

```
from nltk.stem.porter import PorterStemmer
from nltk.stem.lancaster import LancasterStemmer
from nltk.stem.snowball import SnowballStemmer
```

```
words = ['AA', 'KURCHI', 'MADATHAPETTI', 'MINGITHE', 'MADAL',
'IRIGIPOVALI']


stemmers = ['MB', 'SL', 'GK']


stemmer_porter = PorterStemmer()
stemmer_lancaster = LancasterStemmer()
stemmer_snowball = SnowballStemmer('english')


formatted_row = '{:>16}' * (len(stemmers) + 1)
print ('\n', formatted_row.format('WORD', *stemmers), '\n')
```

```
                  WORD              MB              SL              GK
```

```
for word in words:
    stemmed_words = [
        stemmer_porter.stem(word),
        stemmer_lancaster.stem(word),
        stemmer_snowball.stem(word)
    ]

    print(formatted_row.format(word, *stemmed_words))
```

```
                  AA              aa              aa              aa
              KURCHI          kurchi           kurch          kurchi
        MADATHAPETTI    madathapetti     madathapett    madathapetti
            MINGITHE         mingith         mingith         mingith
               MADAL           madal             mad           madal
         IRIGIPOVALI     irigipovali      irigipoval     irigipovali
```

Double-click (or enter) to edit

Converting text to its base form using

lemmatization

```
from nltk.stem import WordNetLemmatizer


words = ['AA', 'KURCHI', 'MADATHAPETTI', 'MINGITHE', 'MADAL',
'IRIGIPOVALI']


lemmatizers = ['NOUN LEMMATIZER', 'VERB LEMMATIZER']


lemmatizer_wordnet = WordNetLemmatizer()


formatted_row = '{:>24}' * (len(lemmatizers) + 1)
print ('\n', formatted_row.format('WORD', *lemmatizers), '\n')
```

```
                          WORD         NOUN LEMMATIZER         VERB LEMMATIZER


for word in words:
  lemmatized_words = [lemmatizer_wordnet.lemmatize(word,
                                                   pos='n'),
                      lemmatizer_wordnet.lemmatize(word, pos='v')]
  print (formatted_row.format(word, *lemmatized_words))
```

```
---------------------------------------------------------------------------
LookupError                               Traceback (most recent call last)
/usr/local/lib/python3.10/dist-packages/nltk/corpus/util.py in __load(self)
     83                 try:
---> 84                     root = nltk.data.find(f"
{self.subdir}/{zip_name}")
     85                 except LookupError:
```

6 frames

```
LookupError:
**********************************************************************
  Resource wordnet not found.
  Please use the NLTK Downloader to obtain the resource:

  >>> import nltk
  >>> nltk.download('wordnet')

  For more information see: https://www.nltk.org/data.html

  Attempted to load corpora/wordnet.zip/wordnet/

  Searched in:
    - '/root/nltk_data'
    - '/usr/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/share/nltk_data'
    - '/usr/local/share/nltk_data'
    - '/usr/lib/nltk_data'
    - '/usr/local/lib/nltk_data'
```