```
# Tokenisation:
# Tokenisation is the process of breaking down a piece of text. like a senetence or a paragraph, into individual words or "Tokens"
# It is the process of converting a sequence of text into smaller parts,known as tokens.
# These tokens can be as small as characters or as long as words.


import nltk # Natural language tulkit
nltk.download("punkt")
```

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

```
# word tokenisation using the split()function
my_text= """A paragraph is a self-contained unit of discourse in writing dealing with a particular point or idea.
 Though not required by the orthographic conventions of any language with a writing system,
  paragraphs are a conventional means of organizing extended segments of prose."""

print(my_text.split())
```

```
['A', 'paragraph', 'is', 'a', 'self-contained', 'unit', 'of', 'discourse', 'in', 'writing', 'dealing', 'with', 'a', 'particular', 'p
```

```
txt="How such documents are actually stored depends on the file format!."


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


# Word tokenizer
print(word_tokenize(txt))
```

```
['How', 'such', 'documents', 'are', 'actually', 'stored', 'depends', 'on', 'the', 'file', 'format', '!', '.']
```

```
# Sentence Tokenizer
print(sent_tokenize(txt))
```

```
['How such documents are actually stored depends on the file format!.']
```