

```
#EXCLLER DAY-8(198)
# Install NLTK (if not installed)
!pip install nltk

import nltk

# Download necessary NLTK resources
nltk.download('punkt')

def tokenize_text(paragraph):
    """Tokenizes text into sentences and words."""
    sentences = nltk.sent_tokenize(paragraph) # Sentence tokenization
    words = [nltk.word_tokenize(sentence) for sentence in sentences] # Word tokenization
    return sentences, words

# Sample paragraph for tokenization
paragraph = """
Tokenization is the process of breaking text into smaller units called tokens.
These tokens can be words, sentences, or even smaller units. Tokenization is an
important step in text preprocessing.
"""

# Tokenize the text
sentences, words = tokenize_text(paragraph)

# Print tokenized sentences
print("Sentences:")
for sentence in sentences:
    print(sentence)

# Print tokenized words
print("\nWords:")
for word_list in words:
    print(word_list)
```

```

Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.1.8)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!

```

```

-----
LookupError                                Traceback (most recent call last)
<ipython-input-15-b62f946928e4> in <cell line: 0>()

```

```

22
23 # Tokenize the text
--> 24 sentences, words = tokenize_text(paragraph)
25
26 # Print tokenized sentences

```

⬆ 5 frames

```

/usr/local/lib/python3.11/dist-packages/nltk/data.py in find(resource_name, paths)
577     sep = "*" * 70
578     resource_not_found = f"\n{sep}\n{msg}\n{sep}\n"
--> 579     raise LookupError(resource_not_found)
580
581

```

LookupError:

\*\*\*\*\*

Resource `punkt_tab` not found.  
Please use the NLTK Downloader to obtain the resource:

```

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

```

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

Attempted to load `tokenizers/punkt_tab/english/`