

Name: Prathamesh Arvind Jadhav
Roll No: 4059

Experiment No:1

Experiment Title:

Implement Tokenization by Word and Sentence

Aim:

To implement tokenization techniques in Natural Language Processing (NLP) using word and sentence tokenization.

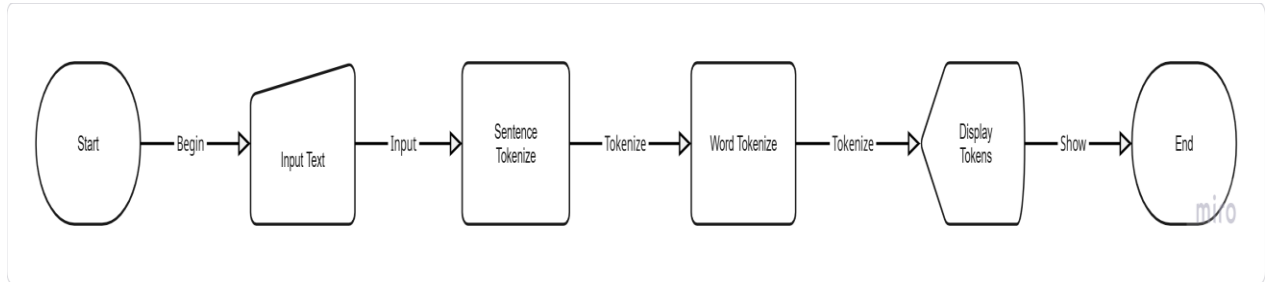
Objective:

- Understand the concept of tokenization in NLP.
 - Perform word-level and sentence-level tokenization on a sample text.
 - Implement tokenization using Python with NLTK/spaCy.
-

Procedure & Flowchart:

- **Procedure:**
 1. Install required NLP libraries like NLTK or spaCy.
 2. Import necessary modules for tokenization.
 3. Define a sample paragraph or input string.
 4. Apply sentence tokenization to split text into sentences.
 5. Apply word tokenization to split sentences into words.
 6. Print the results of both tokenizations.

- **Flowchart:**



Code / Implementation:

✓ To implement tokenization by word

```
[1] import nltk
    nltk.download('punkt')

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

[5] def tokenize_by_word(text):
    return nltk.word_tokenize(text)

[4] nltk.download('punkt_tab')

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

sample_text = "Hello! My name is Prathamesh Jadhav."
tokens = tokenize_by_word(sample_text)
print(tokens)

['Hello', '!', 'My', 'name', 'is', 'Prathamesh', 'Jadhav', '.']
```

✓ Sentence Tokenization using NLTK

```
✓ 1s ▶ # Sentence Tokenization using NLTK

import nltk
from nltk.tokenize import sent_tokenize

# Download the required data
nltk.download('punkt')

# Sample text
text = "Hello! My name is Prathamesh Jadhav. I am learning Natural Language Processing using Python."

# Sentence Tokenization
sentences = sent_tokenize(text)

print("Sentence Tokenization:")
for i, sentence in enumerate(sentences, 1):
    print(f"{i}. {sentence}")
```

↗ Sentence Tokenization:

1. Hello!
2. My name is Prathamesh Jadhav.
3. I am learning Natural Language Processing using Python.

[nltk_data] Downloading package punkt to /root/nltk_data...

[nltk_data] Package punkt is already up-to-date!

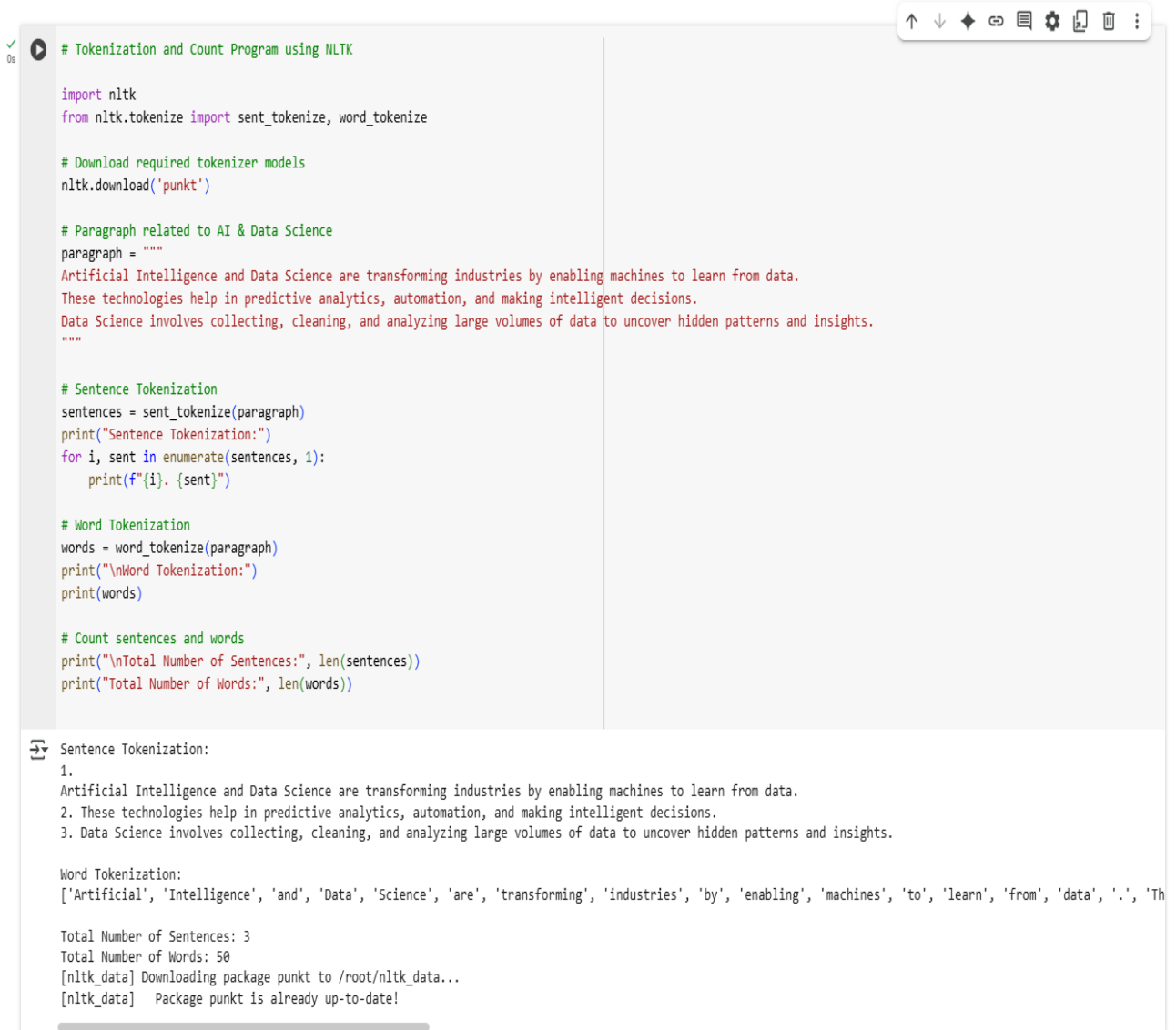
Student Activity - Code & Output

Student Task:

1. Prepare a tokenization program using a different paragraph related to your subject.
2. Identify and count the number of words and sentences.

Sample Code for Student Activity:

Tokenization + Count



```
# Tokenization and Count Program using NLTK

import nltk
from nltk.tokenize import sent_tokenize, word_tokenize

# Download required tokenizer models
nltk.download('punkt')

# Paragraph related to AI & Data Science
paragraph = """
Artificial Intelligence and Data Science are transforming industries by enabling machines to learn from data.
These technologies help in predictive analytics, automation, and making intelligent decisions.
Data Science involves collecting, cleaning, and analyzing large volumes of data to uncover hidden patterns and insights.
"""

# Sentence Tokenization
sentences = sent_tokenize(paragraph)
print("Sentence Tokenization:")
for i, sent in enumerate(sentences, 1):
    print(f"{i}. {sent}")

# Word Tokenization
words = word_tokenize(paragraph)
print("\nWord Tokenization:")
print(words)

# Count sentences and words
print("\nTotal Number of Sentences:", len(sentences))
print("Total Number of Words:", len(words))
```

Sentence Tokenization:

1. Artificial Intelligence and Data Science are transforming industries by enabling machines to learn from data.
2. These technologies help in predictive analytics, automation, and making intelligent decisions.
3. Data Science involves collecting, cleaning, and analyzing large volumes of data to uncover hidden patterns and insights.

Word Tokenization:

```
['Artificial', 'Intelligence', 'and', 'Data', 'Science', 'are', 'transforming', 'industries', 'by', 'enabling', 'machines', 'to', 'learn', 'from', 'data', '.', 'Th
```

Total Number of Sentences: 3
Total Number of Words: 50
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!

Questions & Answers:

1. What is tokenization in NLP? (CO1)

Answer:

Tokenization in NLP is the process of splitting a large text into smaller units such as words or sentences. These units (called tokens) are essential for analyzing and processing natural language using machine learning models or rule-based systems.

2. Differentiate between word and sentence tokenization. (CO1)

Answer:

- **Word Tokenization** breaks a sentence into individual words or tokens (e.g., "NLP is fun" → ['NLP', 'is', 'fun']).
 - **Sentence Tokenization** breaks a paragraph or text into sentences (e.g., "NLP is fun. It's a growing field." → ['NLP is fun.', "It's a growing field."]).
-

3. Write a Python program to implement sentence and word tokenization. (CO2)

Answer:

```
from nltk.tokenize import sent_tokenize, word_tokenize
import nltk
nltk.download('punkt')

text = "NLP is transforming technology. Tokenization is the first step."
print("Sentence Tokenization:", sent_tokenize(text))
print("Word Tokenization:", word_tokenize(text))
```

4. Why is tokenization important in the preprocessing phase of NLP tasks? (CO3)

Tokenization is important because it breaks down raw text into manageable pieces (tokens), enabling further analysis like parsing, tagging, or machine learning. It standardizes text input and helps models focus on meaningful units of language.

Conclusion:

In this experiment, we successfully implemented word and sentence tokenization using Python's NLTK library. Tokenization is a foundational step in any NLP pipeline, and understanding it is crucial for further processing tasks such as POS tagging, parsing, or named entity recognition.