



THE
WORLD IS
HERE
@



SHARDA
UNIVERSITY
Beyond Boundaries

AGENTIC AI LAB

NAME-TAUSHIR ALAM

SYSTEM ID-2023509207

ROLL NO-2301010904

WORKING CODE OF CHUNKING MEHTOD

1.install lib

```
pip install -U langchain-text-splitters
```

2.sample text

```
text = """This is the text I would like to chunk up.
```

It is the example text for this exercise.

Chunking helps large documents become searchable and useful for AI systems.

Recursive splitting preserves sentence and word boundaries."""

3.charcater level

```
def character_split(text, chunk_size=30):  
    chunks = []  
    for i in range(0, len(text), chunk_size):  
        chunks.append(text[i:i+chunk_size])  
    return chunks  
  
chunks = character_split(text, 30)  
  
for i, c in enumerate(chunks):  
    print(f"Chunk {i+1}: {c}")
```

4.word level splitting

```
def word_split(text, chunk_size=6):  
    words = text.split()  
    chunks = []  
    for i in range(0, len(words), chunk_size):  
        chunk = " ".join(words[i:i+chunk_size])  
        chunks.append(chunk)  
    return chunks
```

```
chunks = word_split(text, 6)
for i, c in enumerate(chunks):
    print(f"Chunk {i+1}: {c}")

5.sentence level

import re

def sentence_split(text):
    sentences = re.split(r'(?<=[.!?])\s+', text)
    return sentences

chunks = sentence_split(text)

for i, c in enumerate(chunks):
    print(f"Chunk {i+1}: {c}")

6.recurisvie character

from langchain_text_splitters import RecursiveCharacterTextSplitter
splitter = RecursiveCharacterTextSplitter(
    chunk_size=80,
    chunk_overlap=20
)
chunks = splitter.split_text(text)

for i, c in enumerate(chunks):
    print(f"\n--- Chunk {i+1} ---")
    print(c)

token level

from transformers import AutoTokenizer
from langchain_text_splitters import TokenTextSplitter
tokenizer = AutoTokenizer.from_pretrained("gpt2")
```

```
splitter = TokenTextSplitter(  
    chunk_size=50,  
    chunk_overlap=10,  
    tokenizer=tokenizer  
)
```

```
chunks = splitter.split_text(text)
```

```
for i, c in enumerate(chunks):  
    print(f"\n--- Chunk {i+1} ---")  
    print(c)
```