

B. TECH 3rd YEAR STUDENT

AGENTIC AI



BACHELOR OF TECHNOLOGY

In

Computer Science and Engineering

Created By

Ajay Gahlawat

2023437648

LAB 2- DATA_CHUNKING

WORKING CODE-

```
from langchain.text_splitter import (
    CharacterTextSplitter,
    RecursiveCharacterTextSplitter,
    TokenTextSplitter,
    SentenceTransformersTokenTextSplitter
)

text = """
Artificial Intelligence is transforming the world.
It enables machines to learn from data and make decisions.
Text splitting is an important preprocessing step in NLP.
Large documents must be divided into smaller chunks.
"""

character_splitter = CharacterTextSplitter(
    separator="\n",
    chunk_size=100,
    chunk_overlap=20
)

character_chunks = character_splitter.split_text(text)

print("Character Text Splitter Output:")
```

```
print(character_chunks)
```

```
recursive_splitter = RecursiveCharacterTextSplitter(  
    chunk_size=100,  
    chunk_overlap=20  
)
```

```
recursive_chunks = recursive_splitter.split_text(text)  
print("\nRecursive Character Text Splitter Output:")  
print(recursive_chunks)
```

```
token_splitter = TokenTextSplitter(  
    chunk_size=30,  
    chunk_overlap=5  
)
```

```
token_chunks = token_splitter.split_text(text)  
print("\nToken Text Splitter Output:")  
print(token_chunks)
```

```
sentence_transformer_splitter = SentenceTransformersTokenTextSplitter(  
    chunk_size=30,  
    chunk_overlap=5  
)
```

```
sentence_chunks = sentence_transformer_splitter.split_text(text)
print("\nSentence Transformer Token Splitter Output:")
print(sentence_chunks)
```