Laboratory file

on

AGENTIC AI

School of Engineering and Technology

Department of Computer Science and Engineering

Subject code _ **CSCR 3215**

SUBMITTED BY:

Name : Mayank
System ID: 2023436982

SUBMITTED TO:

Mr. Ayush Singh

Sharda University

Greater Noida, Uttar Pradesh

Lab 02: Chunking Method – Multi-Level Text Splitting

Objective:

The objective of this project is to implement and compare multiple text chunking strategies to efficiently divide large unstructured documents into smaller, meaningful segments for use in Retrieval-Augmented Generation (RAG), semantic search, and Large Language Model (LLM) applications.

Methodology:

1. Dataset Collection:
   Textual data from various sources such as plain text files, markdown documents, code files, and PDFs containing mixed content (text, tables, and images) are used.
2. Data Preprocessing:
   Documents are loaded and cleaned. Depending on the format, specialized loaders and parsers extract textual and structural elements.
3. Chunking Techniques Implemented:
   - Character-based chunking
   - Recursive character chunking
   - Document-structure-based chunking
   - Semantic chunking using embeddings
   - Agentic chunking using LLM reasoning
4. Tools & Libraries:
   LangChain text splitters, embedding models (e.g., OpenAI embeddings), vector similarity techniques, and Unstructured library for multimodal PDF parsing.

Working:

1. The input document is first loaded into the system.
2. Based on the selected level, the text is split using different strategies:
   - Fixed-size character slicing
   - Recursive splitting using paragraphs, lines, and words
   - Structure-aware splitting (headings, code blocks, functions)
3. In semantic chunking, sentences are converted into embeddings, and similarity between them is computed to group related content.
4. In agentic chunking, an LLM analyzes the document context and decides logical chunk boundaries.
5. Each chunk is stored with metadata and can be used for retrieval or downstream LLM tasks.

Outcomes:

1. Efficient handling of long documents within LLM token limits.
2. Improved contextual coherence in chunks compared to naive splitting.
3. Better retrieval accuracy in RAG systems due to semantic grouping.
4. Capability to process multimodal PDFs containing text, tables, and images.