Agentic AI Lab

CSCR3215

B.Tech. (CSE)-VI Semester

Submitted to:
Mr Ayush Kumar

Submitted by:
Abhinav
(2023378721)

School of Engineering & Technology
Department of Computer Science & Engineering

Working of Chunking

1. Objective

The purpose of this code is to demonstrate different levels of text splitting used in Natural Language Processing (NLP). Text splitting is essential when working with large documents so that language models can process them efficiently.

2. Why Text Splitting is Needed

Large language models have token limits. Splitting text helps in chunking long documents into smaller, meaningful parts without losing context.

3. Level 1: Character-Based Splitting

In this level, text is split purely based on a fixed number of characters. It is simple but may break sentences or words abruptly.

4. Level 2: Word-Based Splitting

Here, text is split using word boundaries. This approach preserves words but may still break Sentences.

5. Level 3: Sentence-Based Splitting

This level splits text using sentence boundaries such as periods or punctuation. It maintains semantic meaning better than character or word splitting.

6. Level 4: Recursive Text Splitting

Recursive splitting tries multiple strategies in order. If a chunk is too large, it splits again using another rule. This approach balances size and meaning.

7. Level 5: Semantic or Context-Aware Splitting

The most advanced level splits text based on semantic similarity. Paragraphs or topics are kept together to preserve context.

8. Code Workflow

The code loads input text, applies different splitting strategies sequentially, and prints or stores the resulting chunks for analysis.

9. Applications

Text splitting is widely used in document search, chatbots, vector databases, retrieval-augmented generation (RAG), and summarization tasks.

10. Conclusion

This notebook demonstrates how increasing levels of text splitting improve contextual understanding while managing model limitations.
