Enhanced RAG Techniques

Advanced Chunking And Preprocessing

pdf, txt, word, json
⇓
Document

Document Splitter
Recursive Character

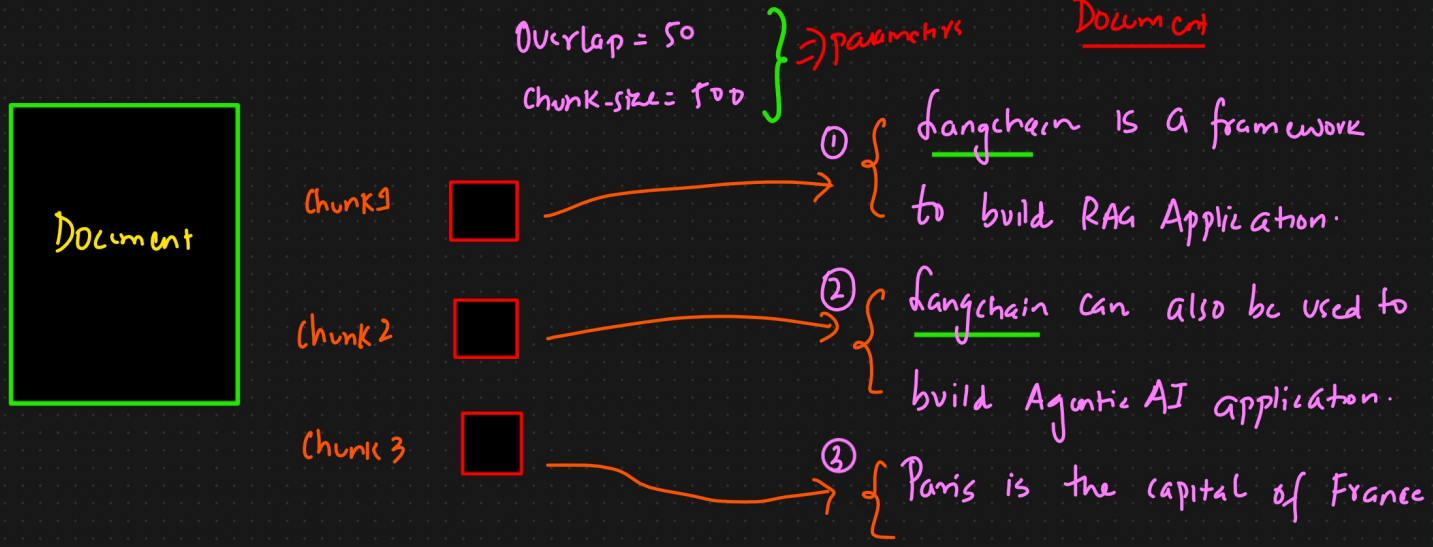
1) Semantic Chunking

Semantic chunking is the process of splitting a document into meaningful units (chunks) based on semantic similarity — not just by number of tokens or lines.

This is important in RAG systems because:

Better chunks → better retrieval → better grounding → better answers.

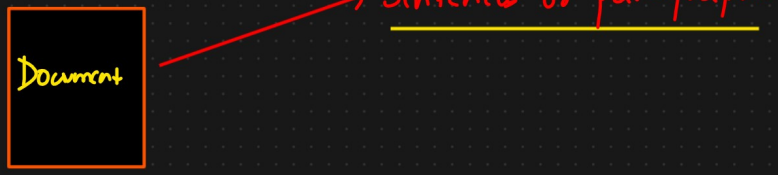
Chunks should be self-contained, contextually rich, and logically separated.



Contextual rich logically separated chunks.

How does It Work ?

1) Document Segmentation



2) Sentence Embedding : Each sentence is converted into a vector representation

OpenAI, Hf => Sentence Transformers

3) Semantic Similarity Check : Cosine similarity between adjacent embedding

S1 ↔ S2 => Similar => 0.95

S2 ↔ S3

S3 ↔ S4

{ Threshold => 0.8
=> 0.75
=> 0.95 }

4) Merging Of Sentences : Merge adjacent sentences if they are semantically similar

5) Form chunks : Merge S1 and S2 => [S1, S2] => chunk 1

=> [S3] => chunk 2.

Chunks

1. LangChain is a framework for building LLM-powered apps.
2. It integrates with tools like OpenAI and Pinecone.
3. The Eiffel Tower is located in Paris.
4. France is a popular tourist destination.

⇓ Semantic chunking

Olp

["LangChain is a framework...", "It integrates with tools..."]
["The Eiffel Tower is located in Paris."]
["France is a popular tourist destination."]