Week 6: Embeddings + Document Indexing

Full Stack RAG with Local LLM

Semester 2/2568

Agenda

1. Embeddings
2. Local Embedding
3. Document Chunking
4. Quiz 2 + Lab 5

What are Embeddings?

Text to Vector Conversion

- Convert text to numbers
- Capture semantic meaning
- Similar texts = Similar vectors

Example:

- "Hello" → [0.12, -0.34, 0.56, ...]
- 1024 dimensions (bge-m3)

Embedding Models

Model	**Dim**	**Type**
OpenAI ada-002	1536	Cloud
BAAI/bge-m3	**1024**	**Local!**

We use bge-m3 - No API Key!

HuggingFace Embedding

```python
from llama_index.embeddings.huggingface import HuggingFaceEmbedding

embed_model = HuggingFaceEmbedding(
    model_name="BAAI/bge-m3",
    trust_remote_code=True
)

# Create embedding - No API Key!
vector = embed_model.get_text_embedding("Hello")
print(len(vector))  # 1024
```

Chunking

```
from llama_index.core.node_parser import SentenceSplitter

splitter = SentenceSplitter(
    chunk_size=1024,
    chunk_overlap=200
)

chunks = splitter.split_text(document_text)
```

Quiz 2 (5%)

Topics: FastAPI, OpenSearch, Pydantic, REST API

Lab 5: Embeddings (3.75%)

Tasks:

1. Study embedding.py
2. Run document indexing
3. Add new documents
4. Test search

Questions? Good luck on Quiz 2!