Lab 5: Embeddings

CSI403 - Full Stack Development

Faculty of Information Technology

Sripatum University

Weight: 3.75%

Objectives

- ✓ Run embedding.py
- ✓ Understand chunking
- ✓ Add new documents

Repository: `https://github.com/amornpan/Generic-RAG`

Task 1: Study embedding.py

Review the pipeline:

1. **Load** documents from `md_corpus/`
2. **Chunk** documents into smaller pieces
3. **Create embeddings** using bge-m3 model
4. **Index** to OpenSearch

bge-m3 produces 1024-dimensional vectors supporting multilingual text

Task 2: Run Indexing

```
cd Generic-RAG
conda activate rag_env
python embedding.py
```

This will:

- Read all `.md` files from `md_corpus/`
- Split into chunks
- Generate embeddings
- Store in OpenSearch

Task 3: Verify

```
curl http://localhost:9200/documents/_count
```

Expected response:

```
{
  "count": 42,
  "_shards": { "total": 1, "successful": 1 }
}
```

Task 4-5: Add New Documents

Step 1: Add 3 new `.md` files to `md_corpus/` folder

Step 2: Re-index

```
python embedding.py
```

Step 3: Verify count increased

```
curl http://localhost:9200/documents/_count
```

Task 6: Test Search

Search for content from your new documents:

- Use the API endpoint `/search`
- Or test via Swagger UI
- Verify your new content appears in results

Deliverables

Item	Check
Indexing completed	☐
Documents verified	☐
New documents added	☐
Search tested	☐

Deadline: Sunday 23:59