Lab 6 Database Access and Dataset

In this lab, you will experiment with RAG-enhanced Retrieval without processing new data. Instead, you will work with preloaded book data from `IST345_Drucker_data` database to understand how RAG works.

For additional guidance, refer to Lab 6, which provides a detailed demonstration.

Dataset Overview

The provided dataset consists of three books authored by Peter F. Drucker:

- Book 1: *The Daily Drucker* (2004 version)
- Book 2: *The Effective Executive* (2002 version)
- Book 3: *The Essential Drucker* (2008 version)

Steps for Processing These Books into Database

Step 1: Extract Text from PDFs

- We use `PyMuPDFLoader` to parse PDFs and extract raw text.

Step 2: Perform Data Cleaning

- Remove unwanted pages (cover page, table of contents, empty pages, reference pages).
- Filter out short content (less than 60 characters).

Step 3: Modify Metadata

- Since this dataset contains book content, we **store** additional metadata, such as `title`, `author,` `year of publication`, `page number`, `and source`.

```
docs[0].metadata
```

```
[232] ✓ 0.0s
```

```
{'source': 'The Essential Drucker-2008.pdf',
 'title': 'The Essential Drucker',
 'author': 'Peter F. Drucker',
 'year': '2008',
 'page': 6}
```

Step 4: Chunk the Document for Retrieval

- For this dataset, we use `chunk_size = 512` and `chunk_overlap = 60`.

Step 5: Store Data in the Vector Database

- The cleaned and chunked text with metadata is loaded into the `IST345_Drucker_data` **database.**

`IST345_Drucker_data` Database Structure

1. **Table:** `langchain_pg_collection`

This table contains two distinct collections, each using a different embedding model.

Both collections contain identical data with the same chunk size and chunk overlap. The primary difference lies in how the text is encoded for retrieval.

Collection Name	Embedding Model Used
Book_data	`text-embedding-3-large`
Book_data_HF_MiniLM	`sentence-transformers/all-MiniLM-L6-v2`

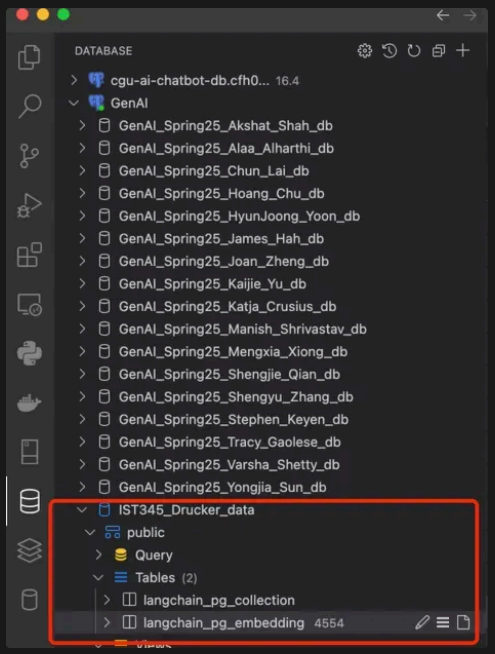
2. **Table:** `langchain_pg_embedding`

This table stores the actual book content with text embeddings generated by the respective models.

Checking Your Database Access

Before running the Lab, confirm your access to the `IST345_Drucker_data` database.

1. Open the Database extension in VS Code.
2. Navigate to `IST345_Drucker_data` > **public** > **Tables**.
3. Check for presence of the two tables (see below).



4. Verify access with a SQL Query

Run the following query to check if you can read data from the database:

```
SELECT * FROM langchain_pg_embedding LIMIT 5;
```

If the query returns results, you have accessed the database correctly and can proceed with Lab 6.