Bio 

G

A project by Kerem Senler & Julia Nuss

1. Exploratory Data Analysis

- Loaded the PDFs
- Removed names and unnecessary information from the PDF-files

! Please note: Our whole code can run completely locally.
We created an online and an offline version.

Online Version

```
Required installations:
pip install langchain chromadb pypdf sentence-transformers ollama
ollama pull llama2
import os
from langchain.document_loaders import PyPDFLoader, DirectoryLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import Chroma
from langchain.llms import Ollama
from langchain.chains import RetrievalQA
```

Offline Version

```
no chroma db meaning everytime you run the code, it 'trains' the data
from the beginning.
import os
from langchain.document_loaders import PyPDFLoader, DirectoryLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import Chroma
from langchain.llms import Ollama
from langchain.chains import RetrievalQA
import time
```

2. Embedding and Storing Chunks, Connecting to VectorDB

- Embedded the documents and split into chunks
- Created vector store from document chunks

3. Connecting to LLM

- Connected to Ollama
- Set up a QA chain
- Created an interactive query loop

4. Evaluation

- How we evaluated our RAG system:
  - Asked multiple questions with a specific answer in our minds and evaluated the answers

5. Deployment

- Used Streamlit to deploy the model on an interactive website:
  - There is a BioRaG website that already contains our 20 PDF-documents about Machine Learning and AI used in biology, you can add other files as well and ask your questions

- Additionally we created a website that’s accessible for everybody to upload their own PDFs and ask questions on them: https://senlerk-pdf-rag-ui-h0kpio.streamlit.app/