

Міністерство освіти і науки України
Національний технічний університет України “Київський політехнічний
інститут імені Ігоря Сікорського”
Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Звіт

з лабораторної роботи №6

з дисципліни

“Програмування інтелектуальних інформаційних систем”

Виконала

ІІ-21 Скрипець О. О.

Перевірив

Баришич Л. М.

Київ 2024

Завдання

1. Поставити лангчейн і підтягнути в нього векторну БД
2. Продемонструвати роботу ВБД текстовим запитом
3. Підготувати ЛЛМ
4. Створити Hggiface Pipeline з БД + Модель
5. Продемонструвати роботу на промті

Не вдалося підключитися до сервісу reCAPTCHA. Перевірте інтернет-з'єднання й оновіть сторінку.

```
!pip install -q langchain
!pip install -q torch
!pip install -q transformers
!pip install -q sentence-transformers
!pip install -q datasets
!pip install -q faiss-cpu
```

```
50.4/50.4 kB 1.7 MB/s eta 0:00:00
1.0/1.0 MB 17.3 MB/s eta 0:00:00
407.2/407.2 kB 19.9 MB/s eta 0:00:00
295.8/295.8 kB 15.1 MB/s eta 0:00:00
76.4/76.4 kB 4.4 MB/s eta 0:00:00
78.0/78.0 kB 4.6 MB/s eta 0:00:00
141.9/141.9 kB 8.3 MB/s eta 0:00:00
54.5/54.5 kB 2.8 MB/s eta 0:00:00
58.3/58.3 kB 3.4 MB/s eta 0:00:00
255.2/255.2 kB 4.2 MB/s eta 0:00:00
471.6/471.6 kB 9.9 MB/s eta 0:00:00
116.3/116.3 kB 6.8 MB/s eta 0:00:00
134.8/134.8 kB 7.0 MB/s eta 0:00:00
194.1/194.1 kB 8.7 MB/s eta 0:00:00
27.5/27.5 MB 49.6 MB/s eta 0:00:00
```

```
!pip install -U langchain-community
```

```
Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Requirement already satisfied: langchain<0.4.0,>=0.3.3 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (0.3.3)
Requirement already satisfied: langchain-core<0.4.0,>=0.3.10 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (0.3.10)
Requirement already satisfied: langsmith<0.2.0,>=0.1.125 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (0.1.125)
Requirement already satisfied: numpy<2,>=1 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (1.26.4)
Collecting pydantic-settings<3.0.0,>=2.4.0 (from langchain-community)
Downloading pydantic_settings-2.6.0-py3-none-any.whl.metadata (3.5 kB)
Requirement already satisfied: requests<3,>=2 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (2.32.3)
Requirement already satisfied: tenacity!=8.4.0,<9.0.0,>=8.1.0 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (8.5.0)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (2.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (1.4.1)
Requirement already satisfied: multidict>=7.0,>=4.5 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (1.9.7)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp<4.0.0,>=3.8.3->langchain-community) (4.0.3)
Collecting marshmallow<4.0.0,>=3.18.0 (from dataclasses_json<0.7,>=0.5.7->langchain-community)
Downloading marshmallow-3.22.0-py3-none-any.whl.metadata (7.2 kB)
Collecting typing-inspect<1,>=0.4.0 (from dataclasses_json<0.7,>=0.5.7->langchain-community)
Downloading typing_inspect-0.9.0-py3-none-any.whl.metadata (1.5 kB)
Requirement already satisfied: langchain-text-splitters<0.4.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from langchain-community) (0.3.0)
Requirement already satisfied: pydantic<3.0.0,>=2.7.4 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.4.0,>=0.3.3->langchain-community) (2.7.4)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.4.0,>=0.3.3->langchain-community) (1.33)
Requirement already satisfied: packaging<25,>=23.2 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.4.0,>=0.3.3->langchain-community) (23.2)
Requirement already satisfied: typing-extensions>=4.7 in /usr/local/lib/python3.10/dist-packages (from langchain-core<0.4.0,>=0.3.3->langchain-community) (4.7.1)
Requirement already satisfied: httpx<1,>=0.23.0 in /usr/local/lib/python3.10/dist-packages (from langsmith<0.2.0,>=0.1.125->langchain-community) (0.23.0)
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in /usr/local/lib/python3.10/dist-packages (from langsmith<0.2.0,>=0.1.125->langchain-community) (3.9.14)
Requirement already satisfied: requests-toolbelt<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from langsmith<0.2.0,>=0.1.125->langchain-community) (1.0.0)
Collecting python-dotenv>=0.21.0 (from pydantic-settings<3.0.0,>=2.4.0->langchain-community)
Downloading python_dotenv-1.0.1-py3-none-any.whl.metadata (23 kB)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langchain-community) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langchain-community) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langchain-community) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3,>=2->langchain-community) (2024.7.4)
Requirement already satisfied: greenlet!=0.4.17 in /usr/local/lib/python3.10/dist-packages (from SQLAlchemy<3,>=1.4->langchain-community) (3.0.3)
Requirement already satisfied: anyio in /usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0->langsmith<0.2.0,>=0.1.125->langchain-community) (3.7.1)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0->langsmith<0.2.0,>=0.1.125->langchain-community) (1.0.5)
Requirement already satisfied: sniffio in /usr/local/lib/python3.10/dist-packages (from httpx<1,>=0.23.0->langsmith<0.2.0,>=0.1.125->langchain-community) (1.3.0)
Requirement already satisfied: h11<0.15,>=0.13 in /usr/local/lib/python3.10/dist-packages (from httpcore==1.*->httpx<1,>=0.23.0->langsmith<0.2.0,>=0.1.125->langchain-community) (0.14.0)
Requirement already satisfied: jsonpointer>=1.9 in /usr/local/lib/python3.10/dist-packages (from jsonpatch<2.0,>=1.33->langchain-community) (2.3)
Requirement already satisfied: annotated-types>=0.6.0 in /usr/local/lib/python3.10/dist-packages (from pydantic<3.0.0,>=2.7.4->langchain-community) (0.6.0)
Requirement already satisfied: pydantic-core==2.23.4 in /usr/local/lib/python3.10/dist-packages (from pydantic<3.0.0,>=2.7.4->langchain-community) (2.23.4)
Collecting mypy_extensions>=0.3.0 (from typing-inspect<1,>=0.4.0->dataclasses_json<0.7,>=0.5.7->langchain-community)
Downloading mypy_extensions-1.0.0-py3-none-any.whl.metadata (1.1 kB)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.10/dist-packages (from yarl<2.0,>=1.12.0->aiohttp<4.0.0,>=3.8.3->langchain-community) (0.2.0)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from anyio->httpx<1,>=0.23.0->langsmith<0.2.0,>=0.1.125->langchain-community) (1.2.0)
Downloading langchain_community-0.3.2-py3-none-any.whl (2.4 MB)
2.4/2.4 MB 31.3 MB/s eta 0:00:00
Downloading dataclasses_json-0.6.7-py3-none-any.whl (28 kB)
Downloading pydantic_settings-2.6.0-py3-none-any.whl (28 kB)
Downloading marshmallow-3.22.0-py3-none-any.whl (49 kB)
49.3/49.3 kB 2.9 MB/s eta 0:00:00
Downloading python_dotenv-1.0.1-py3-none-any.whl (19 kB)
Downloading typing_inspect-0.9.0-py3-none-any.whl (8.8 kB)
Downloading mypy_extensions-1.0.0-py3-none-any.whl (4.7 kB)
Installing collected packages: python-dotenv, mypy_extensions, marshmallow, typing-inspect, pydantic-settings, dataclasses-json,
Successfully installed dataclasses-json-0.6.7 langchain-community-0.3.2 marshmallow-3.22.0 mypy_extensions-1.0.0 pydantic-setting
```

```

from langchain.document_loaders import HuggingFaceDatasetLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from transformers import AutoTokenizer, AutoModelForQuestionAnswering
from transformers import AutoTokenizer, pipeline
from langchain import HuggingFacePipeline
from langchain.chains import RetrievalQA

```

Створюю векторну БД

```

from google.colab import drive
drive.mount('/content/drive')

```

```

with open('/content/drive/My Drive/huggingface_token.txt', 'w') as f:
    f.write('hf_TLZyHFZUtyhcTBupaUCCLxt0ZFeRslkVuJ') # Замініть 'your_token_here' на ваш реальний токен

```

Mounted at /content/drive

```

# Specify the dataset name and the column containing the content
dataset_name = "databricks/databricks-dolly-15k"
page_content_column = "context" # or any other column you're interested in

```

```

# Create a loader instance
loader = HuggingFaceDatasetLoader(dataset_name, page_content_column)

```

```

# Load the data
data = loader.load()

```

```

# Display the first 15 entries
data[:2]

```

README.md: 100% 8.20k/8.20k [00:00<00:00, 93.1kB/s]

databricks-dolly-15k.jsonl: 100% 13.1M/13.1M [00:00<00:00, 25.6MB/s]

Generating train split: 100% 15011/15011 [00:00<00:00, 35974.15 examples/s]

```

[Document(metadata={'instruction': 'When did Virgin Australia start operating?', 'response': 'Virgin Australia commenced services on 31 August 2000 as Virgin Blue, with two aircraft on a single route.', 'category': 'closed_qa'}, page_content="Virgin Australia, the trading name of Virgin Australia Airlines Pty Ltd, is an Australian-based airline. It is the largest airline by fleet size to use the Virgin brand. It commenced services on 31 August 2000 as Virgin Blue, with two aircraft on a single route. It suddenly found itself as a major airline in Australia\'s domestic market after the collapse of Ansett Australia in September 2001. The airline has since grown to directly serve 32 cities in Australia, from hubs in Brisbane, Melbourne and Sydney."), Document(metadata={'instruction': 'Which is a species of fish? Tope or Rope', 'response': 'Tope', 'category': 'classification'}, page_content=""))]

```

```

# Create an instance of the RecursiveCharacterTextSplitter class with specific parameters.
# It splits text into chunks of 1000 characters each with a 150-character overlap.
text_splitter = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=150)

# 'data' holds the text you want to split, split the text into documents using the text splitter.
docs = text_splitter.split_documents(data)

```

```

# Define the path to the pre-trained model you want to use
modelPath = "sentence-transformers/all-MiniLM-l6-v2"

```

```

# Create a dictionary with model configuration options, specifying to use the CPU for computations
model_kwargs = {'device':'cpu'}

```

```

# Create a dictionary with encoding options, specifically setting 'normalize_embeddings' to False
encode_kwargs = {'normalize_embeddings': False}

```

```

# Initialize an instance of HuggingFaceEmbeddings with the specified parameters
embeddings = HuggingFaceEmbeddings(
    model_name=modelPath, # Provide the pre-trained model's path
    model_kwargs=model_kwargs, # Pass the model configuration options
    encode_kwargs=encode_kwargs # Pass the encoding options
)

```

```
<ipython-input-16-a34a09b47e7d>:11: LangChainDeprecationWarning: The class `HuggingFaceEmbeddings` was deprecated in LangChain 0.2.14.
  embeddings = HuggingFaceEmbeddings(
modules.json: 100% 349/349 [00:00<00:00, 15.7kB/s]
config_sentence_transformers.json: 100% 116/116 [00:00<00:00, 5.93kB/s]
README.md: 100% 10.7k/10.7k [00:00<00:00, 207kB/s]
sentence_bert_config.json: 100% 53.0/53.0 [00:00<00:00, 2.19kB/s]
config.json: 100% 612/612 [00:00<00:00, 36.7kB/s]
model.safetensors: 100% 90.9M/90.9M [00:00<00:00, 135MB/s]
tokenizer_config.json: 100% 350/350 [00:00<00:00, 7.96kB/s]
vocab.txt: 100% 232k/232k [00:00<00:00, 3.85MB/s]
text = "This is a test document."
query_result = embeddings.embed_query(text)
query_result[:3]

[-0.038338541984558105, 0.12346471846103668, -0.02864297851920128]

1 Embeddings loaded: 100% 100/100 [00:00<00:00, 1.07kB/s]
db = FAISS.from_documents(docs, embeddings)

question = "What is cheesemaking?"
searchDocs = db.similarity_search(question)
print(searchDocs[0].page_content)

"The goal of cheese making is to control the spoiling of milk into cheese. The milk is traditionally from a cow, goat, sheep or buffalo"
```