# EXPERIMENT NO. 15

| |
|---|
| **Student Name and Roll Number:** Piyush Gambhir – 21CSU349 |
| **Semester /Section:** Semester-V – AIML-V-B (AL-3) |
| **Link to Code:** NCU-Lab-Manual-And-End-Semester-Projects/NCU-CSL347-AAIES-Lab_Manual at main · Piyush-Gambhir/NCU-Lab-Manual-And-End-Semester-Projects (github.com) |
| **Date:** 25.11.2023 |
| **Faculty Signature:** |
| **Grade:** |

**Objective(s):**

- To study Large Language Models (LLM).
- To build an efficient model using Generative AI based LLMs.

**Outcome:**

Students will be able to understand how to perform a Natural Language processing task using Generative AI based LLMs

**Problem Statement:**

To implement simple PDF Document search using Open Source Generative AI model.

**Background Study:** Generative AI based Language Models (LLMs) utilize deep learning to generate human-like text. These models are highly versatile and find application in diverse areas such as creative writing, chatbots, language translation, and even document search. LLMs can effectively process and understand natural language, making them valuable tools for searching and summarizing vast collections of documents.

**Question Bank:**

1. What is a Large Language Model?
   A Large Language Model (LLM) is a type of artificial intelligence model designed to understand, generate, and manipulate human language. It uses deep learning techniques, often based on transformer architectures, to process and generate text, enabling it to perform tasks like language translation, text generation, question answering, and more.

2. What is a transformer pipeline?
   A transformer pipeline refers to a sequence of processing steps using a transformer-based model, like GPT-3.5, bard. to perform specific tasks. It involves input data being passed through the model to receive relevant outputs. For example, a text generation pipeline might involve input preprocessing, passing the input through the model, and then post-processing the model's generated text.
3. What are vector databases and embeddings?
   Vector databases and embeddings are techniques used in machine learning to represent data and enable efficient processing. A vector database stores data points as vectors in a multi-dimensional space, making it easier to search and compare them. Embeddings are lower-dimensional representations of data, often learned from raw data, which capture semantic relationships. For instance, word embeddings represent words in a way that preserves semantic similarities.

4. How does similarity search work?

   Vector databases and embeddings are techniques used in machine learning to represent data and enable efficient processing. A vector database stores data points as vectors in a multi-dimensional space, making it easier to search and compare them. Embeddings are lower-dimensional representations of data, often learned from raw data, which capture semantic relationships. For instance, word embeddings represent words in a way that preserves semantic similarities.

# Student Work Area

Algorithm/Flowchart/Code/Sample Outputs

# Experiment 15

### Problem Statement:

To implement simple PDF Document search using Open Source Generative AI model.

### Install Dependencies:

In [ ]:

```
! pip install langchain pypdf faiss-cpu sentence-transformers
```

```
Requirement already satisfied: langchain in c:\users\mainp\appdata\local\programs\python\pytho
n311\lib\site-packages (0.0.344)
Requirement already satisfied: pypdf in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (3.17.1)
Requirement already satisfied: faiss-cpu in c:\users\mainp\appdata\local\programs\python\pytho
n311\lib\site-packages (1.7.4)
Requirement already satisfied: sentence-transformers in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (2.2.2)
Requirement already satisfied: PyYAML>=5.3 in c:\users\mainp\appdata\local\programs\python\pyt
hon311\lib\site-packages (from langchain) (6.0.1)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in c:\users\mainp\appdata\local\programs\pyt
hon\python311\lib\site-packages (from langchain) (2.0.23)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from langchain) (3.9.1)
Requirement already satisfied: anyio<4.0 in c:\users\mainp\appdata\local\programs\python\pytho
n311\lib\site-packages (from langchain) (3.7.1)
Requirement already satisfied: dataclasses-json<0.7,>=0.5.7 in c:\users\mainp\appdata\local\pr
ograms\python\python311\lib\site-packages (from langchain) (0.6.3)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in c:\users\mainp\appdata\local\programs\p
ython\python311\lib\site-packages (from langchain) (1.33)
Requirement already satisfied: langchain-core<0.1,>=0.0.8 in c:\users\mainp\appdata\local\prog
rams\python\python311\lib\site-packages (from langchain) (0.0.8)
Requirement already satisfied: langsmith<0.1.0,>=0.0.63 in c:\users\mainp\appdata\local\progra
ms\python\python311\lib\site-packages (from langchain) (0.0.67)
Requirement already satisfied: numpy<2,>=1 in c:\users\mainp\appdata\local\programs\python\pyt
hon311\lib\site-packages (from langchain) (1.25.0)
Requirement already satisfied: pydantic<3,>=1 in c:\users\mainp\appdata\local\programs\python
\python311\lib\site-packages (from langchain) (2.5.2)
Requirement already satisfied: requests<3,>=2 in c:\users\mainp\appdata\local\programs\python
\python311\lib\site-packages (from langchain) (2.31.0)
Requirement already satisfied: tenacity<9.0.0,>=8.1.0 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from langchain) (8.2.3)
Requirement already satisfied: transformers<5.0.0,>=4.6.0 in c:\users\mainp\appdata\local\prog
rams\python\python311\lib\site-packages (from sentence-transformers) (4.35.2)
Requirement already satisfied: tqdm in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (from sentence-transformers) (4.65.0)
Requirement already satisfied: torch>=1.6.0 in c:\users\mainp\appdata\local\programs\python\py
thon311\lib\site-packages (from sentence-transformers) (2.1.1+cu118)
Requirement already satisfied: torchvision in c:\users\mainp\appdata\local\programs\python\pyt
hon311\lib\site-packages (from sentence-transformers) (0.16.1+cu118)
Requirement already satisfied: scikit-learn in c:\users\mainp\appdata\local\programs\python\py
thon311\lib\site-packages (from sentence-transformers) (1.2.2)
Requirement already satisfied: scipy in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (from sentence-transformers) (1.10.1)
Requirement already satisfied: nltk in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (from sentence-transformers) (3.8.1)
Requirement already satisfied: sentencepiece in c:\users\mainp\appdata\local\programs\python\p
ython311\lib\site-packages (from sentence-transformers) (0.1.99)
Requirement already satisfied: huggingface-hub>=0.4.0 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from sentence-transformers) (0.19.4)
Requirement already satisfied: attrs>=17.3.0 in c:\users\mainp\appdata\local\programs\python\p
ython311\lib\site-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (23.1.0)
Requirement already satisfied: multidict<7.0,>=4.5 in c:\users\mainp\appdata\local\programs\py
thon\python311\lib\site-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in c:\users\mainp\appdata\local\programs\python
\python311\lib\site-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.9.3)
Requirement already satisfied: frozenlist>=1.1.1 in c:\users\mainp\appdata\local\programs\pyth
on\python311\lib\site-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in c:\users\mainp\appdata\local\programs\pytho
n\python311\lib\site-packages (from aiohttp<4.0.0,>=3.8.3->langchain) (1.3.1)
Requirement already satisfied: idna>=2.8 in c:\users\mainp\appdata\local\programs\python\pytho
n311\lib\site-packages (from anyio<4.0->langchain) (3.4)
Requirement already satisfied: sniffio>=1.1 in c:\users\mainp\appdata\local\programs\python\py
thon311\lib\site-packages (from anyio<4.0->langchain) (1.3.0)
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in c:\users\mainp\appdata\local\prog
rams\python\python311\lib\site-packages (from dataclasses-json<0.7,>=0.5.7->langchain) (3.20.
1)
Requirement already satisfied: typing-inspect<1,>=0.4.0 in c:\users\mainp\appdata\local\progra
ms\python\python311\lib\site-packages (from dataclasses-json<0.7,>=0.5.7->langchain) (0.9.0)
Requirement already satisfied: filelock in c:\users\mainp\appdata\local\programs\python\python
311\lib\site-packages (from huggingface-hub>=0.4.0->sentence-transformers) (3.12.2)
Requirement already satisfied: fsspec>=2023.5.0 in c:\users\mainp\appdata\local\programs\pytho
n\python311\lib\site-packages (from huggingface-hub>=0.4.0->sentence-transformers) (2023.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in c:\users\mainp\appdata\local\prog
rams\python\python311\lib\site-packages (from huggingface-hub>=0.4.0->sentence-transformers)
(4.6.3)
Requirement already satisfied: packaging>=20.9 in c:\users\mainp\appdata\local\programs\python
\python311\lib\site-packages (from huggingface-hub>=0.4.0->sentence-transformers) (23.1)
Requirement already satisfied: jsonpointer>=1.9 in c:\users\mainp\appdata\local\programs\pytho
n\python311\lib\site-packages (from jsonpatch<2.0,>=1.33->langchain) (2.4)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from pydantic<3,>=1->langchain) (0.6.0)
Requirement already satisfied: pydantic-core==2.14.5 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from pydantic<3,>=1->langchain) (2.14.5)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\mainp\appdata\local\progra
ms\python\python311\lib\site-packages (from requests<3,>=2->langchain) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\mainp\appdata\local\programs\pyt
hon\python311\lib\site-packages (from requests<3,>=2->langchain) (2.1.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\mainp\appdata\local\programs\pyt
hon\python311\lib\site-packages (from requests<3,>=2->langchain) (2023.11.17)
Requirement already satisfied: greenlet!=0.4.17 in c:\users\mainp\appdata\local\programs\pytho
n\python311\lib\site-packages (from SQLAlchemy<3,>=1.4->langchain) (3.0.1)
Requirement already satisfied: sympy in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (from torch>=1.6.0->sentence-transformers) (1.12)
Requirement already satisfied: networkx in c:\users\mainp\appdata\local\programs\python\python
311\lib\site-packages (from torch>=1.6.0->sentence-transformers) (3.1)
Requirement already satisfied: jinja2 in c:\users\mainp\appdata\local\programs\python\python31
1\lib\site-packages (from torch>=1.6.0->sentence-transformers) (3.1.2)
Requirement already satisfied: colorama in c:\users\mainp\appdata\local\programs\python\python
311\lib\site-packages (from tqdm->sentence-transformers) (0.4.6)
Requirement already satisfied: regex!=2019.12.17 in c:\users\mainp\appdata\local\programs\pyth
on\python311\lib\site-packages (from transformers<5.0.0,>=4.6.0->sentence-transformers) (2023.
10.3)
Requirement already satisfied: tokenizers<0.19,>=0.14 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from transformers<5.0.0,>=4.6.0->sentence-transformers)
(0.15.0)
Requirement already satisfied: safetensors>=0.3.1 in c:\users\mainp\appdata\local\programs\pyt
hon\python311\lib\site-packages (from transformers<5.0.0,>=4.6.0->sentence-transformers) (0.4.
1)
Requirement already satisfied: click in c:\users\mainp\appdata\local\programs\python\python311
\lib\site-packages (from nltk->sentence-transformers) (8.1.7)
Requirement already satisfied: joblib in c:\users\mainp\appdata\local\programs\python\python31
1\lib\site-packages (from nltk->sentence-transformers) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\mainp\appdata\local\programs\p
ython\python311\lib\site-packages (from scikit-learn->sentence-transformers) (3.1.0)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from torchvision->sentence-transformers) (9.5.0)
Requirement already satisfied: mypy-extensions>=0.3.0 in c:\users\mainp\appdata\local\programs
\python\python311\lib\site-packages (from typing-inspect<1,>=0.4.0->dataclasses-json<0.7,>=0.
5.7->langchain) (1.0.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\mainp\appdata\local\programs\python
\python311\lib\site-packages (from jinja2->torch>=1.6.0->sentence-transformers) (2.1.3)
Requirement already satisfied: mpmath>=0.19 in c:\users\mainp\appdata\local\programs\python\py
thon311\lib\site-packages (from sympy->torch>=1.6.0->sentence-transformers) (1.3.0)
```

## Code:

```
In [ ]: # importing required libraries
        import os
        from langchain.document_loaders import PyPDFLoader
        from langchain.embeddings.openai import OpenAIEmbeddings
from langchain.text_splitter import CharacterTextSplitter
from langchain.vectorstores import FAISS
from langchain.chains.question_answering import load_qa_chain
from langchain import HuggingFaceHub
from langchain.embeddings import HuggingFaceEmbeddings
```

In [ ]:

```
# setting up the huggingfacehub api token
os.environ["HUGGINGFACEHUB_API_TOKEN"] = "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" # this is a
```

In [ ]:

```
# Loading the pdf document using pyPDF and the document Loader
pdf_loader = PyPDFLoader("https://scholarworks.calstate.edu/downloads/vq27zt20r")
pdf_document = pdf_loader.load()
```

In [ ]:

```
# converting the pdf document to raw text
pdf_text = ''

for i, page in enumerate(pdf_document):
    page_text = page.page_content
    if(page_text != None):
        pdf_text += page_text
```

In [ ]:

```
# splitting the document into chunks using RecursiveTextSplitter
text_splitter = CharacterTextSplitter(
    separator="\n",
    chunk_size=800,
    chunk_overlap=200,
    length_function=len,

)

# splitting the document into chunks
chunks = text_splitter.split_text(pdf_text)
```

In [ ]:

```
# Loading the openai embeddings
embeddings = HuggingFaceEmbeddings()
```

In [ ]:

```
# creating the vector store
document_search = FAISS.from_texts(chunks, embeddings)
```

In [ ]:

```
# Loading the Flan-T5 XL model from the huggingface hub

model = HuggingFaceHub(repo_id="google/flan-t5-xl",
                       model_kwargs={"temperature": 1, "max_length": 1000000})
```

```
c:\Users\mainp\AppData\Local\Programs\Python\Python311\Lib\site-packages\huggingface_hub\utils
\_deprecation.py:127: FutureWarning: '__init__' (from 'huggingface_hub.inference_api') is depr
ecated and will be removed from version '1.0'. `InferenceApi` client is deprecated in favor of
the more feature-complete `InferenceClient`. Check out this guide to learn how to convert your
script to use it: https://huggingface.co/docs/huggingface_hub/guides/inference#legacy-inferenc
eapi-client.
  warnings.warn(warning_message, FutureWarning)
```

In [ ]:

```
# Loading the question answering chain
chain = load_qa_chain(model, chain_type="stuff")
```

In [ ]:

```
query = "What are the algorithms used in the paper?"
docs = document_search.similarity_search(query)
answer = chain.run(input_documents=docs, question=query)
print(answer)
```

```
DQN [11] and A2C [21]
```