llama2-final

January 3, 2024

1 AI Based Chatbot to answer Linux Questions

1.0.1 We intend to create LLM based AI Chatbot that would answer the questions about Linux

It is simple LLM bot that answers the questions only on the trained dataset unlike RAG. We will train the data on the corpus of few Linux books like Linux Bible, etc. The vectorstore to store the learnings is FAISS database.

1.0.2 Technology Used

1. **LLM**: meta-llama/Llama-2-7b-chat-hf [https://huggingface.co/meta-llama/Llama-2-7b-chat-hf]
2. **VectorStore**: ***FAISS*** => FAISS (Facebook AI Similarity Search) is a library that allows developers to quickly search for embeddings of multimedia documents that are similar to each other [https://ai.meta.com/tools/faiss/]
3. **Embeddings**: ***sentence-transformers/all-mpnet-base-v2*** [https://huggingface.co/sentence-transformers/all-mpnet-base-v2]

1.0.3 Installing dependencies using pip

[1]:

```
!pip install -i https://test.pypi.org/simple/ bitsandbytes
!pip install -r requirements.txt
```

```
Looking in indexes: https://test.pypi.org/simple/
Requirement already satisfied: bitsandbytes in /opt/conda/lib/python3.10/site-
packages (0.39.0)
Requirement already satisfied: pypdf in /opt/conda/lib/python3.10/site-packages
(from -r requirements.txt (line 1)) (3.17.4)
Requirement already satisfied: langchain in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 2)) (0.0.318)
Requirement already satisfied: torch in /opt/conda/lib/python3.10/site-packages
(from -r requirements.txt (line 3)) (2.1.2)
Requirement already satisfied: accelerate in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 4)) (0.21.0)
Requirement already satisfied: bitsandbytes in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 5)) (0.39.0)
Requirement already satisfied: transformers in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 6)) (4.31.0)
```

```
Requirement already satisfied: sentence_transformers in
/opt/conda/lib/python3.10/site-packages (from -r requirements.txt (line 7))
(2.2.2)
Requirement already satisfied: faiss_cpu in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 8)) (1.7.4)
Requirement already satisfied: chainlit in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 9)) (0.7.700)
Requirement already satisfied: huggingface_hub in
/opt/conda/lib/python3.10/site-packages (from -r requirements.txt (line 10))
(0.19.0)
Requirement already satisfied: unstructured in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 11)) (0.11.6)
Requirement already satisfied: xformers in /opt/conda/lib/python3.10/site-
packages (from -r requirements.txt (line 12)) (0.0.23.post1)
Requirement already satisfied: PyYAML>=5.3 in /opt/conda/lib/python3.10/site-
packages (from langchain->-r requirements.txt (line 2)) (6.0.1)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (1.4.49)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (3.8.6)
Requirement already satisfied: anyio<4.0 in /opt/conda/lib/python3.10/site-
packages (from langchain->-r requirements.txt (line 2)) (3.7.1)
Requirement already satisfied: async-timeout<5.0.0,>=4.0.0 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (4.0.3)
Requirement already satisfied: dataclasses-json<0.7,>=0.5.7 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (0.5.14)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (1.33)
Requirement already satisfied: langsmith<0.1.0,>=0.0.43 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (0.0.60)
Requirement already satisfied: numpy<2,>=1 in /opt/conda/lib/python3.10/site-
packages (from langchain->-r requirements.txt (line 2)) (1.26.0)
Requirement already satisfied: pydantic<3,>=1 in /opt/conda/lib/python3.10/site-
packages (from langchain->-r requirements.txt (line 2)) (1.10.13)
Requirement already satisfied: requests<3,>=2 in /opt/conda/lib/python3.10/site-
packages (from langchain->-r requirements.txt (line 2)) (2.31.0)
Requirement already satisfied: tenacity<9.0.0,>=8.1.0 in
/opt/conda/lib/python3.10/site-packages (from langchain->-r requirements.txt
(line 2)) (8.2.3)
Requirement already satisfied: filelock in /opt/conda/lib/python3.10/site-
packages (from torch->-r requirements.txt (line 3)) (3.13.1)
Requirement already satisfied: typing-extensions in
```

```
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (4.5.0)
Requirement already satisfied: sympy in /opt/conda/lib/python3.10/site-packages
(from torch->-r requirements.txt (line 3)) (1.12)
Requirement already satisfied: networkx in /opt/conda/lib/python3.10/site-
packages (from torch->-r requirements.txt (line 3)) (3.2.1)
Requirement already satisfied: jinja2 in /opt/conda/lib/python3.10/site-packages
(from torch->-r requirements.txt (line 3)) (3.1.2)
Requirement already satisfied: fsspec in /opt/conda/lib/python3.10/site-packages
(from torch->-r requirements.txt (line 3)) (2023.6.0)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.105)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.105)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.105)
Requirement already satisfied: nvidia-cudnn-cu12==8.9.2.26 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (8.9.2.26)
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.3.1)
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (11.0.2.54)
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (10.3.2.106)
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (11.4.5.107)
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.0.106)
Requirement already satisfied: nvidia-nccl-cu12==2.18.1 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (2.18.1)
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in
/opt/conda/lib/python3.10/site-packages (from torch->-r requirements.txt (line
3)) (12.1.105)
Requirement already satisfied: triton==2.1.0 in /opt/conda/lib/python3.10/site-
packages (from torch->-r requirements.txt (line 3)) (2.1.0)
Requirement already satisfied: nvidia-nvjitlink-cu12 in
/opt/conda/lib/python3.10/site-packages (from nvidia-cusolver-
cu12==11.4.5.107->torch->-r requirements.txt (line 3)) (12.3.101)
```

```
Requirement already satisfied: packaging>=20.0 in
/opt/conda/lib/python3.10/site-packages (from accelerate->-r requirements.txt
(line 4)) (23.2)
Requirement already satisfied: psutil in /opt/conda/lib/python3.10/site-packages
(from accelerate->-r requirements.txt (line 4)) (5.9.5)
Requirement already satisfied: regex!=2019.12.17 in
/opt/conda/lib/python3.10/site-packages (from transformers->-r requirements.txt
(line 6)) (2023.10.3)
Requirement already satisfied: tokenizers!=0.11.3,<0.14,>=0.11.1 in
/opt/conda/lib/python3.10/site-packages (from transformers->-r requirements.txt
(line 6)) (0.13.3)
Requirement already satisfied: safetensors>=0.3.1 in
/opt/conda/lib/python3.10/site-packages (from transformers->-r requirements.txt
(line 6)) (0.3.3)
Requirement already satisfied: tqdm>=4.27 in /opt/conda/lib/python3.10/site-
packages (from transformers->-r requirements.txt (line 6)) (4.66.1)
Requirement already satisfied: torchvision in /opt/conda/lib/python3.10/site-
packages (from sentence_transformers->-r requirements.txt (line 7))
(0.15.2a0+072ec57)
Requirement already satisfied: scikit-learn in /opt/conda/lib/python3.10/site-
packages (from sentence_transformers->-r requirements.txt (line 7)) (1.3.2)
Requirement already satisfied: scipy in /opt/conda/lib/python3.10/site-packages
(from sentence_transformers->-r requirements.txt (line 7)) (1.11.3)
Requirement already satisfied: nltk in /opt/conda/lib/python3.10/site-packages
(from sentence_transformers->-r requirements.txt (line 7)) (3.8.1)
Requirement already satisfied: sentencepiece in /opt/conda/lib/python3.10/site-
packages (from sentence_transformers->-r requirements.txt (line 7)) (0.1.99)
Requirement already satisfied: aiofiles<24.0.0,>=23.1.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (23.2.1)
Requirement already satisfied: asyncer<0.0.3,>=0.0.2 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.0.2)
Requirement already satisfied: click<9.0.0,>=8.1.3 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (8.1.7)
Requirement already satisfied: fastapi<0.101,>=0.100 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.100.1)
Requirement already satisfied: fastapi-socketio<0.0.11,>=0.0.10 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.0.10)
Requirement already satisfied: filetype<2.0.0,>=1.2.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (1.2.0)
Requirement already satisfied: httpx<0.25.0,>=0.23.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.24.1)
```

```
Requirement already satisfied: lazify<0.5.0,>=0.4.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.4.0)
Requirement already satisfied: nest-asyncio<2.0.0,>=1.5.6 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (1.5.8)
Requirement already satisfied: pyjwt<3.0.0,>=2.8.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (2.8.0)
Requirement already satisfied: python-dotenv<2.0.0,>=1.0.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (1.0.0)
Requirement already satisfied: python-graphql-client<0.5.0,>=0.4.3 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.4.3)
Requirement already satisfied: python-multipart<0.0.7,>=0.0.6 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.0.6)
Requirement already satisfied: syncer<3.0.0,>=2.0.3 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (2.0.3)
Requirement already satisfied: tomli<3.0.0,>=2.0.1 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (2.0.1)
Requirement already satisfied: uptrace<2.0.0,>=1.18.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (1.22.0)
Requirement already satisfied: uvicorn<0.24.0,>=0.23.2 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.23.2)
Requirement already satisfied: watchfiles<0.21.0,>=0.20.0 in
/opt/conda/lib/python3.10/site-packages (from chainlit->-r requirements.txt
(line 9)) (0.20.0)
Requirement already satisfied: chardet in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (5.2.0)
Requirement already satisfied: python-magic in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (0.4.27)
Requirement already satisfied: lxml in /opt/conda/lib/python3.10/site-packages
(from unstructured->-r requirements.txt (line 11)) (5.0.0)
Requirement already satisfied: tabulate in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (0.9.0)
Requirement already satisfied: beautifulsoup4 in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (4.12.2)
Requirement already satisfied: emoji in /opt/conda/lib/python3.10/site-packages
(from unstructured->-r requirements.txt (line 11)) (2.9.0)
Requirement already satisfied: python-iso639 in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (2024.1.2)
Requirement already satisfied: langdetect in /opt/conda/lib/python3.10/site-
```

```
packages (from unstructured->-r requirements.txt (line 11)) (1.0.9)
Requirement already satisfied: rapidfuzz in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (3.6.1)
Requirement already satisfied: backoff in /opt/conda/lib/python3.10/site-
packages (from unstructured->-r requirements.txt (line 11)) (2.2.1)
Requirement already satisfied: unstructured-client in
/opt/conda/lib/python3.10/site-packages (from unstructured->-r requirements.txt
(line 11)) (0.6.0)
Requirement already satisfied: wrapt in /opt/conda/lib/python3.10/site-packages
(from unstructured->-r requirements.txt (line 11)) (1.15.0)
Requirement already satisfied: attrs>=17.3.0 in /opt/conda/lib/python3.10/site-
packages (from aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2))
(23.1.0)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in
/opt/conda/lib/python3.10/site-packages (from
aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2)) (3.3.2)
Requirement already satisfied: multidict<7.0,>=4.5 in
/opt/conda/lib/python3.10/site-packages (from
aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2)) (6.0.4)
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/site-
packages (from aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2))
(1.9.2)
Requirement already satisfied: frozenlist>=1.1.1 in
/opt/conda/lib/python3.10/site-packages (from
aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2)) (1.4.0)
Requirement already satisfied: aiosignal>=1.1.2 in
/opt/conda/lib/python3.10/site-packages (from
aiohttp<4.0.0,>=3.8.3->langchain->-r requirements.txt (line 2)) (1.3.1)
Requirement already satisfied: idna>=2.8 in /opt/conda/lib/python3.10/site-
packages (from anyio<4.0->langchain->-r requirements.txt (line 2)) (3.4)
Requirement already satisfied: sniffio>=1.1 in /opt/conda/lib/python3.10/site-
packages (from anyio<4.0->langchain->-r requirements.txt (line 2)) (1.3.0)
Requirement already satisfied: exceptiongroup in /opt/conda/lib/python3.10/site-
packages (from anyio<4.0->langchain->-r requirements.txt (line 2)) (1.1.3)
Requirement already satisfied: marshmallow<4.0.0,>=3.18.0 in
/opt/conda/lib/python3.10/site-packages (from dataclasses-
json<0.7,>=0.5.7->langchain->-r requirements.txt (line 2)) (3.20.1)
Requirement already satisfied: typing-inspect<1,>=0.4.0 in
/opt/conda/lib/python3.10/site-packages (from dataclasses-
json<0.7,>=0.5.7->langchain->-r requirements.txt (line 2)) (0.9.0)
Requirement already satisfied: starlette<0.28.0,>=0.27.0 in
/opt/conda/lib/python3.10/site-packages (from
fastapi<0.101,>=0.100->chainlit->-r requirements.txt (line 9)) (0.27.0)
Requirement already satisfied: python-socketio>=4.6.0 in
/opt/conda/lib/python3.10/site-packages (from fastapi-
socketio<0.0.11,>=0.0.10->chainlit->-r requirements.txt (line 9)) (5.10.0)
Requirement already satisfied: certifi in /opt/conda/lib/python3.10/site-
packages (from httpx<0.25.0,>=0.23.0->chainlit->-r requirements.txt (line 9))
```

```
(2023.7.22)
Requirement already satisfied: httpcore<0.18.0,>=0.15.0 in
/opt/conda/lib/python3.10/site-packages (from
httpx<0.25.0,>=0.23.0->chainlit->-r requirements.txt (line 9)) (0.17.3)
Requirement already satisfied: jsonpointer>=1.9 in
/opt/conda/lib/python3.10/site-packages (from
jsonpatch<2.0,>=1.33->langchain->-r requirements.txt (line 2)) (2.4)
Requirement already satisfied: websockets>=5.0 in
/opt/conda/lib/python3.10/site-packages (from python-graphql-
client<0.5.0,>=0.4.3->chainlit->-r requirements.txt (line 9)) (12.0)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/opt/conda/lib/python3.10/site-packages (from requests<3,>=2->langchain->-r
requirements.txt (line 2)) (1.26.18)
Requirement already satisfied: greenlet!=0.4.17 in
/opt/conda/lib/python3.10/site-packages (from SQLAlchemy<3,>=1.4->langchain->-r
requirements.txt (line 2)) (3.0.1)
Requirement already satisfied: opentelemetry-api~=1.22 in
/opt/conda/lib/python3.10/site-packages (from
uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9)) (1.22.0)
Requirement already satisfied: opentelemetry-exporter-otlp~=1.22 in
/opt/conda/lib/python3.10/site-packages (from
uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9)) (1.22.0)
Requirement already satisfied: opentelemetry-instrumentation~=0.43b0 in
/opt/conda/lib/python3.10/site-packages (from
uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9)) (0.43b0)
Requirement already satisfied: opentelemetry-sdk~=1.22 in
/opt/conda/lib/python3.10/site-packages (from
uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9)) (1.22.0)
Requirement already satisfied: h11>=0.8 in /opt/conda/lib/python3.10/site-
packages (from uvicorn<0.24.0,>=0.23.2->chainlit->-r requirements.txt (line 9))
(0.14.0)
Requirement already satisfied: soupsieve>1.2 in /opt/conda/lib/python3.10/site-
packages (from beautifulsoup4->unstructured->-r requirements.txt (line 11))
(2.5)
Requirement already satisfied: MarkupSafe>=2.0 in
/opt/conda/lib/python3.10/site-packages (from jinja2->torch->-r requirements.txt
(line 3)) (2.1.3)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages
(from langdetect->unstructured->-r requirements.txt (line 11)) (1.16.0)
Requirement already satisfied: joblib in /opt/conda/lib/python3.10/site-packages
(from nltk->sentence_transformers->-r requirements.txt (line 7)) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/opt/conda/lib/python3.10/site-packages (from scikit-
learn->sentence_transformers->-r requirements.txt (line 7)) (3.2.0)
Requirement already satisfied: mpmath>=0.19 in /opt/conda/lib/python3.10/site-
packages (from sympy->torch->-r requirements.txt (line 3)) (1.3.0)
Requirement already satisfied: pillow!=8.3.*,>=5.3.0 in
/opt/conda/lib/python3.10/site-packages (from
torchvision->sentence_transformers->-r requirements.txt (line 7)) (9.5.0)
Requirement already satisfied: jsonpath-python>=1.0.6 in
/opt/conda/lib/python3.10/site-packages (from unstructured-
client->unstructured->-r requirements.txt (line 11)) (1.0.6)
Requirement already satisfied: marshmallow-enum>=1.5.1 in
/opt/conda/lib/python3.10/site-packages (from unstructured-
client->unstructured->-r requirements.txt (line 11)) (1.5.1)
Requirement already satisfied: mypy-extensions>=0.4.3 in
/opt/conda/lib/python3.10/site-packages (from unstructured-
client->unstructured->-r requirements.txt (line 11)) (1.0.0)
Requirement already satisfied: pyparsing>=3.0.9 in
/opt/conda/lib/python3.10/site-packages (from unstructured-
client->unstructured->-r requirements.txt (line 11)) (3.1.1)
Requirement already satisfied: python-dateutil>=2.8.2 in
/opt/conda/lib/python3.10/site-packages (from unstructured-
client->unstructured->-r requirements.txt (line 11)) (2.8.2)
Requirement already satisfied: deprecated>=1.2.6 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-
api~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.2.14)
Requirement already satisfied: importlib-metadata<7.0,>=6.0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-
api~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(6.8.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-grpc==1.22.0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.22.0)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-http==1.22.0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.22.0)
Requirement already satisfied: googleapis-common-protos~=1.52 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-otlp-proto-
grpc==1.22.0->opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.61.0)
Requirement already satisfied: grpcio<2.0.0,>=1.0.0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-otlp-proto-
grpc==1.22.0->opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.54.3)
Requirement already satisfied: opentelemetry-exporter-otlp-proto-common==1.22.0
in /opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-otlp-
proto-grpc==1.22.0->opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.22.0)
Requirement already satisfied: opentelemetry-proto==1.22.0 in
```

```
/opt/conda/lib/python3.10/site-packages (from opentelemetry-exporter-otlp-proto-
grpc==1.22.0->opentelemetry-exporter-
otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(1.22.0)
Requirement already satisfied: protobuf<5.0,>=3.19 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-
proto==1.22.0->opentelemetry-exporter-otlp-proto-grpc==1.22.0->opentelemetry-
exporter-otlp~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line
9)) (4.21.12)
Requirement already satisfied: setuptools>=16.0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-
instrumentation~=0.43b0->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt
(line 9)) (68.2.2)
Requirement already satisfied: opentelemetry-semantic-conventions==0.43b0 in
/opt/conda/lib/python3.10/site-packages (from opentelemetry-
sdk~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(0.43b0)
Requirement already satisfied: bidict>=0.21.0 in /opt/conda/lib/python3.10/site-
packages (from python-socketio>=4.6.0->fastapi-
socketio<0.0.11,>=0.0.10->chainlit->-r requirements.txt (line 9)) (0.22.1)
Requirement already satisfied: python-engineio>=4.8.0 in
/opt/conda/lib/python3.10/site-packages (from python-socketio>=4.6.0->fastapi-
socketio<0.0.11,>=0.0.10->chainlit->-r requirements.txt (line 9)) (4.8.1)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.10/site-
packages (from importlib-metadata<7.0,>=6.0->opentelemetry-
api~=1.22->uptrace<2.0.0,>=1.18.0->chainlit->-r requirements.txt (line 9))
(3.17.0)
Requirement already satisfied: simple-websocket>=0.10.0 in
/opt/conda/lib/python3.10/site-packages (from python-engineio>=4.8.0->python-
socketio>=4.6.0->fastapi-socketio<0.0.11,>=0.0.10->chainlit->-r requirements.txt
(line 9)) (1.0.0)
Requirement already satisfied: wsproto in /opt/conda/lib/python3.10/site-
packages (from simple-websocket>=0.10.0->python-engineio>=4.8.0->python-
socketio>=4.6.0->fastapi-socketio<0.0.11,>=0.0.10->chainlit->-r requirements.txt
(line 9)) (1.2.0)
```

1.0.4 Necessary Imports

[14]:

```python
import warnings
import os
import huggingface_hub
import torch
from torch import cuda, bfloat16
from transformers import BitsAndBytesConfig
from transformers import AutoTokenizer, AutoModelForCausalLM
from transformers import pipeline
from langchain.llms import HuggingFacePipeline
from langchain.document_loaders import PyPDFLoader, DirectoryLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from langchain import PromptTemplate
from langchain.chains import RetrievalQA
from langchain.globals import set_debug, set_verbose
import logging
set_debug(False)
set_verbose(False)
logging.getLogger('langchain').setLevel(logging.DEBUG)
```

1.0.5 Loading the model

[3]:

```
def load_llama(model_id):
    device = f'cuda:{cuda.current_device()}' if cuda.is_available() else 'cpu'
    bnb_config = BitsAndBytesConfig(
        load_in_4bit=True,
        bnb_4bit_quant_type='nf4',
        bnb_4bit_use_double_quant=True,
        bnb_4bit_compute_dtype=bfloat16
    )
    if torch.cuda.is_available():
            torch.set_default_tensor_type(torch.cuda.HalfTensor)
    print(f"Loading the model {model_id}")
    tokenizer = AutoTokenizer.from_pretrained(model_id)
    model = AutoModelForCausalLM.from_pretrained(model_id,␣
 ↪pad_token_id=tokenizer.eos_token_id)
    llama_pipeline = pipeline(
        model=model,
        tokenizer=tokenizer,
        return_full_text=True,
        max_new_tokens=512,
        #quantization_config=bnb_config,
        temperature=0.7,
        task="text-generation",  # LLM task
        torch_dtype=torch.float16,
        device_map="auto",
    )
    llm = HuggingFacePipeline(pipeline=llama_pipeline)
    return llm
```

1.0.6 Logging in to HuggingFace repo

[4]:
```
def set_hf_token():
     os.environ["HF_TOKEN"] = "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"
     print("------------------------------------")
     print("Huggingface login")
     huggingface_hub.login(os.environ["HF_TOKEN"])
     print("------------------------------------\n")
```

1.0.7 Loading the documents

[5]:
```
def load_documents(local_directory_path):
    # For PDF files
    print(f"\n------------------------------------")
    print(f"Loading PDFs from {local_directory_path}")
    loader = DirectoryLoader(local_directory_path,
                                glob='*.pdf',
                                loader_cls=PyPDFLoader)
    print(loader)
    documents = loader.load()
    print(documents)
    print(f"Documents Loaded")
    print(f"------------------------------------\n")
    return documents
```

1.0.8 Processing the documents

[6]:
```
def process_documents(documents):
    print(f"\n------------------------------------")
    print(f"Processing the documents")
    text_splitter = RecursiveCharacterTextSplitter(chunk_size=1000,
                                                    chunk_overlap=50)
    texts = text_splitter.split_documents(documents)
    print(f"Documents processed")
    print(f"------------------------------------\n")
    return texts
```

1.0.9 Setting the embeddings

[7]:
```
def set_embeddings(model_name):
     print(f"\n------------------------------------")
     print(f"Setting the embeddings")
     embeddings = HuggingFaceEmbeddings(model_name=model_name)
     print(f"Embeddings set successfully")
     print(f"------------------------------------\n")
     return embeddings
```

1.0.10 Saving to FAISS Vectorstore

[8]:
```
def save_to_vectorstore(texts, embeddings, vectorestore_path):
     print(f"\n-------------------------------------")
     print(f"Saving the vectorestore to {vectorestore_path}")
     vectorstore = FAISS.from_documents(texts, embeddings)
     vectorstore.save_local(vectorestore_path)
     print(f"Vectore DB stored at {vectorestore_path}")
     print(f"-------------------------------------\n")
     return vectorstore
```

1.0.11 Setting the custom prompt template

[9]:
```
custom_prompt_template = """Use the following information to answer the user's␣
 ↪question.
In case you don't know the answer, just say that you don't know, don't try to␣
 ↪make up an answer.

Context: {context}
Question: {question}

Only return the helpful answer below and nothing else.
Helpful answer:
"""

def set_custom_prompt():
    """
    Prompt template for QA retrieval for each vectorstore
    """
    print("Setting the custom prompt")
    prompt = PromptTemplate(template=custom_prompt_template,
                            input_variables=['context', 'question'])
    return prompt
```

1.0.12 Retrieval QA Chain function

The RetrievalQAChain is a chain that combines a Retriever and a QA chain (described above). It is used to retrieve documents from a Retriever and then use a QA chain to answer a question based on the retrieved documents. Read this for more info on RetrievalQA

[10]:
```
def retrieval_qa_chain(llm, vectorstore):
     chain = RetrievalQA.from_chain_type(llm=llm,
                                         chain_type='stuff',
                                         retriever=vectorstore.
 ↪as_retriever(search_kwargs={'k': 2}),
                                         return_source_documents=True,
                                          chain_type_kwargs={'prompt':␣
↪set_custom_prompt()},
                                          verbose=False
                                          )
    return chain
```

1.0.13 Chatbot Query

[11]:
```
def chatbot(llm, vectorstore):
    chain = retrieval_qa_chain(llm, vectorstore)

    while True:
        user_input = input("User: ")
        if user_input.lower() == "exit":
            print("Chatbot: Thanks!")
            break
        result = chain({"query": user_input})

        response = result["result"]

        print("Chatbot: ", response)
        print("--------------------------------------------------\n\n")
```

1.0.14 Sample testing. Having the debugger output on for understanding the flow of responses

[12]:
```
if __name__ == "__main__":
    set_hf_token()
    documents = load_documents("/home/sagemaker-user/content/corpus")
    texts = process_documents(documents)
    embeddings = set_embeddings('sentence-transformers/all-mpnet-base-v2')
    vectorstore = save_to_vectorstore(texts, embeddings, "/home/sagemaker-user/
↪content/vectorstore/")
    chatbot(load_llama("meta-llama/Llama-2-7b-chat-hf"), vectorstore)
```

```
------------------------------------
Huggingface login
Token will not been saved to git credential helper. Pass
`add_to_git_credential=True` if you want to set the git credential as well.
Token is valid (permission: write).
Your token has been saved to /home/sagemaker-user/.cache/huggingface/token
Login successful
------------------------------------



------------------------------------
```

```
Loading PDFs from /home/sagemaker-user/content/corpus
<langchain.document_loaders.directory.DirectoryLoader object at 0x7f02e2b4df00>

IOPub data rate exceeded.
The Jupyter server will temporarily stop sending output
to the client in order to avoid crashing it.
To change this limit, set the config variable
`--ServerApp.iopub_data_rate_limit`.

Current values:
ServerApp.iopub_data_rate_limit=1000000.0 (bytes/sec)
ServerApp.rate_limit_window=3.0 (secs)


Documents processed
------------------------------------



------------------------------------
Setting the embeddings
Embeddings set successfully
------------------------------------



------------------------------------
Saving the vectorestore to /home/sagemaker-user/content/vectorstore/
Vectore DB stored at /home/sagemaker-user/content/vectorstore/
------------------------------------

Loading the model meta-llama/Llama-2-7b-chat-hf

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]

Setting the custom prompt

User:  What is linux

[chain/start] [1:chain:RetrievalQA] Entering Chain run with

input:

{
  "query": "What is linux"
}
[chain/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain] Entering Chain run with input:

[inputs]
```

```
[chain/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain] Entering Chain run with input:

{
  "question": "What is linux",
  "context": "other hand, was developed in a different context. Linux is a PC
version of the Unix operating system that has been used for decades on
mainframes and minicomputers and is currently the system of choice for network
servers and workstations. Linux brings the \nspeed, efficiency, scalability, and
flexibility of Unix to your PC, taking advantage of all the \ncapabilities that
PCs can now provide.\nTechnically, Linux consists of the operating system
program, referred to as the kernel,\n\nLinux
   \n \n         \n \n      "
}
[llm/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:HuggingFacePipeline]

Entering LLM run with input:

{
  "prompts": [
    "Use the following information to answer the user's question.\nIn case you
don't know the answer, just say that you don't know, don't try to make up an
answer.\n\nContext: other hand, was developed in a different context. Linux is a
PC version of the Unix operating system that has been used for decades on
mainframes and minicomputers and is currently the system of choice for network
servers and workstations. Linux brings the \nspeed, efficiency, scalability, and
flexibility of Unix to your PC, taking advantage of all the \ncapabilities that
PCs can now provide.\nTechnically, Linux consists of the operating system
program, referred to as the kernel,\n\nLinux
   \n \n         \n \n      \nQuestion: What is linux\n\nOnly
return the helpful answer below and nothing else.\nHelpful answer:"
  ]
}
[llm/end] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:HuggingFacePipeline]

[17.03s] Exiting LLM run with output:

{
  "generations": [
    [
      {
        "text": "Linux is an open-source operating system that is based on the
Unix operating system and is designed to be highly scalable, flexible, and
efficient. It is commonly used on servers and workstations, but can also be used
on personal computers.",
        "generation_info": null
      }
    ]
  ],
  "llm_output": null,
  "run": null
}
[chain/end] [1:chain:RetrievalQA >
3:chain:StuffDocumentsChain > 4:chain:LLMChain] [17.03s] Exiting Chain run with
output:
{
  "text": "Linux is an open-source operating system that is based on the Unix
operating system and is designed to be highly scalable, flexible, and efficient.
It is commonly used on servers and workstations, but can also be used on
personal computers."
}
[chain/end] [1:chain:RetrievalQA >
3:chain:StuffDocumentsChain] [17.03s] Exiting Chain run with output:
{
  "output_text": "Linux is an open-source operating system that is based on the
Unix operating system and is designed to be highly scalable, flexible, and
efficient. It is commonly used on servers and workstations, but can also be used
on personal computers."
}
[chain/end] [1:chain:RetrievalQA] [17.07s] Exiting Chain
run with output:
[outputs]
Chatbot:  Linux is an open-source operating system that is based on the Unix
operating system and is designed to be highly scalable, flexible, and efficient.
It is commonly used on servers and workstations, but can also be used on
personal computers.
--------------------------------------------------



User:  bye
[chain/start] [1:chain:RetrievalQA] Entering Chain run with
input:
{
  "query": "bye"
}
```

```
[chain/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain] Entering Chain run with input:

[inputs]
[chain/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain] Entering Chain run with input:

{
  "question": "bye",
  "context": "114 Part II: The Linux Shell and File Structure\nAs with .login ,
you can add your own shell commands to the .logout  file. Using the Vi \neditor,
you can change the farewell message or add other operations. In the next
example, the user has a \nclear  and an echo  command in the .logout  file. When
the user logs out, the \nclear  command will clear the screen, and echo  will
display the message "Good-bye for \nnow".\n.logout\nclear\necho \"Good-bye for
now\"\n\nexit(0) A;[Enter]}[Esc] exit(0);\n  }/\ndas76205_Ch05_122-155.indd
128das76205_Ch05_122-155.indd   128 12/13/11   10:44 AM12/13/11   10:44 AM"
}
[llm/start] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:HuggingFacePipeline]

Entering LLM run with input:

{
  "prompts": [
    "Use the following information to answer the user's question.\nIn case you
don't know the answer, just say that you don't know, don't try to make up an
answer.\n\nContext: 114 Part II: The Linux Shell and File Structure\nAs with
.login , you can add your own shell commands to the .logout  file. Using the Vi
\neditor, you can change the farewell message or add other operations. In the
next example, the user has a \nclear  and an echo  command in the .logout  file.
When the user logs out, the \nclear  command will clear the screen, and echo
will display the message "Good-bye for \nnow".\n.logout\nclear\necho \"Good-bye
for now\"\n\nexit(0) A;[Enter]}[Esc] exit(0);\n  }/\ndas76205_Ch05_122-155.indd
128das76205_Ch05_122-155.indd   128 12/13/11   10:44 AM12/13/11   10:44
AM\nQuestion: bye\n\nOnly return the helpful answer below and nothing
else.\nHelpful answer:"
  ]
}
[llm/end] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain > 5:llm:HuggingFacePipeline]

[6.14s] Exiting LLM run with output:

{
  "generations": [
    [
      {
        "text": "The command to logout in Linux is \"exit\".",
        "generation_info": null
      }
    ]
  ],
  "llm_output": null,
  "run": null
}
[chain/end] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain > 4:chain:LLMChain] [6.14s] Exiting Chain run with

output:

{
  "text": "The command to logout in Linux is \"exit\"."
}
[chain/end] [1:chain:RetrievalQA >

3:chain:StuffDocumentsChain] [6.14s] Exiting Chain run with output:

{
  "output_text": "The command to logout in Linux is \"exit\"."
}
[chain/end] [1:chain:RetrievalQA] [6.17s] Exiting Chain run

with output:

[outputs]
Chatbot:  The command to logout in Linux is "exit".
--------------------------------------------------



User:  exit

Chatbot: Thanks!
```

1.1 Final testing

[15]:

```python
chatbot(load_llama("meta-llama/Llama-2-7b-chat-hf"), vectorstore)
```

```
Loading the model meta-llama/Llama-2-7b-chat-hf

Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]

Setting the custom prompt

User:  What is linux

Chatbot:  Linux is an open-source operating system that is based on the Unix
operating system and is designed to be fast, efficient, scalable, and flexible.
It is commonly used on servers and workstations, and is known for its stability
and security.
```

User: what is difference between sudo -i and sudo su -?

Chatbot: sudo -i and sudo su - are both used to switch to a different user, but they have slightly different behaviors:

* sudo -i runs the command as the specified user, but does not prompt for a password. Instead, it uses the user's login credentials from the session.
* sudo su -, on the other hand, runs the command as the specified user and prompts for the user's password before executing the command.

So, the main difference between the two is that sudo -i does not require a password, while sudo su - does.

User: what is devops

Chatbot: DevOps is defined as a set of principles and practices that bring both developers and operations teams together from the start of a software system development to its end-user delivery, with the goal of faster, quicker, and more efficient delivery of software systems to end customers, while reducing time to market and gaining competitive advantage.

User: What should be career path for devops

Chatbot: The author suggests that the best career path for a DevOps engineer is to be involved in the entire SDLC (Software Development Life Cycle), solving communication problems and automating tasks to reduce human intervention. They also mention that it's important to keep up to date with the latest tools and technologies in the field.

User: How to start learning linux?

Chatbot: If you know absolutely nothing about Unix, we recommend you read a basic guide (introductory chapters in the O'Reilly books Learning Red Hat Enterprise Linux and Fedora, by Bill McCarty, and Running Linux, mentioned previously) to get started. This will give you a good understanding of the common utilities run from the shell prompt and how to use them.

User: Suggest linux learning path for beginner

Chatbot: If you are a beginner, I would suggest starting with one of the distributions listed in the text, such as KNOPPIX, Ubuntu, or Damn Small Linux. These distributions are designed to be easy to use and provide a gentle introduction to Linux. Additionally, you can try out the live CDs from the media that comes with this book to get a feel for different distributions and see which one you like best.

User: list 20 important commands for linux

Chatbot:
Here are 20 important commands in Linux:

1. `cd` - change directory
2. `ls` - list files and directories
3. `mkdir` - make a new directory
4. `rm` - remove a file or directory
5. `cp` - copy a file
6. `mv` - move or rename a file
7. `pwd` - print the current working directory
8. `man` - display manual pages
9. `chmod` - change file permissions
10. `chown` - change file ownership
11. `find` - search for files
12. `grep` - search for a pattern in a file
13. `sort` - sort files
14. `uniq` - display only unique lines in a file
15. `cat` - concatenate and display files
16. `nano` - a simple text editor
17. `echo` - output text to a file
18. `sudo` - execute a command with superuser privileges
19. `tar` - create an archive of files
20. `gzip` - compress a file

These are just a few of the many commands available in Linux, but they are some of the most commonly used and useful ones to know.

User: what is ci cd?

Chatbot: CI/CD is an umbrella term that encompasses both Continuous Integration (CI) and Continuous Delivery (CD). It refers to the practice of automating the build, test, and deployment of software, with the goal of delivering high-quality software faster and more reliably. CI/CD involves integrating code

```
changes, running automated tests, and deploying software to production in a
repeatable and efficient manner.
--------------------------------------------------



User:  list tools for CI/CD

Chatbot:  There are several tools available for CI/CD, including:

1. Jenkins
2. Travis CI
3. CircleCI
4. Codeship
5. GoCD
6. TeamCity
7. CruiseControl
8. GitLab CI/CD
9. AWS CodePipeline
10. Azure DevOps

Please note that this is not an exhaustive list and there are many other tools
available.
--------------------------------------------------



User:  what is travis ci?

Chatbot:  Travis CI is one of the second-generation CI/CD tools mentioned in the
passage. It is a cloud-based service that automates the build, test, and
deployment process for software projects. Travis CI is known for its simplicity
and ease of use, and it is often used for small to medium-sized projects.
--------------------------------------------------



User:  Exit

Chatbot: Thanks!
```