

```
%pip install faiss-cpu
```

```
Requirement already satisfied: faiss-cpu in /usr/local/lib/python3.12/dist-packages (1.12.0)
Requirement already satisfied: numpy<3.0,>=1.25.0 in /usr/local/lib/python3.12/dist-packages (from faiss-cpu) (2.0.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.12/dist-packages (from faiss-cpu) (25.0)
```

```
#step 1: required libraries to download
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from langchain.chains import RetrievalQA
from langchain.memory import ConversationBufferMemory
from langchain.prompts import PromptTemplate
from transformers import AutoTokenizer, AutoModelForCausalLM, pipeline
from langchain_community.llms import HuggingFacePipeline
```

```
#step 2:
# Load PDF and extract clean text
loader = PyPDFLoader("/content/sample (3).pdf")
documents = loader.load()
```

```
# Step 3: Semantic chunking
splitter = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=200)
chunks = splitter.split_documents(documents)
```

```
#Step 4: Batch embedding creation & caching(temporal memory)
embedding_model = HuggingFaceEmbeddings(model_name="sentence-transformers/all-MiniLM-L6-v2")

cache = {}

def get_embedding(text):
    if text in cache:
        return cache[text]
    emb = embedding_model.embed_query(text)
    cache[text] = emb
    return emb

#batch embedding of all chunks
embeddings = [get_embedding(chunk.page_content) for chunk in chunks]
```

```
# Step 5: Store embeddings in FAISS vector store
vector_store = FAISS.from_texts([chunk.page_content for chunk in chunks], embedding_model)
```

```
# Step 6 : Set up retriever and retrieval with reranking
retriever = vector_store.as_retriever(search_type="similarity", search_kwargs={"k":2})
```

Step 7: prompt template  
prompt\_template = """  
Use the following context to answer the question. If you don't know the answer, say so.

Context:  
{context}

Question:  
{question}

```
prompt = PromptTemplate(input_variables=["context", "question"], template=prompt_template)
```

```
#step 8 : load model
```

```
model_name = "gpt2"

# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_name)

# Load model
model = AutoModelForCausalLM.from_pretrained(model_name)

# Create a text generation pipeline
pipe = pipeline("text-generation", model=model, tokenizer=tokenizer, max_new_tokens=100)
llm = HuggingFacePipeline(pipe=pipe)
```

Device set to use cpu

```
#Step 9: conversation memory
memory = ConversationBufferMemory(memory_key="chat_history", output_key='result')
```

```
#Step 10: Build RetrievalQA
qa_chain = RetrievalQA.from_chain_type(llm=llm,
                                         retriever=retriever,
                                         return_source_documents=True,
                                         chain_type_kwargs={"prompt": prompt},
                                         memory=memory)
```

```
# query
query = "Phases of NLP"
result = qa_chain.invoke({"query": query})
print("Answer:", result['result'])
```

Setting `pad\_token\_id` to `eos\_token\_id`:50256 for open-end generation.  
 Answer:  
 Use the following context to answer the question. If you don't know the answer, say so.

Context:  
 process language and each phase helps in understanding structure and meaning of  
 human language. In this article, we will understand these phases.

Phases of NLP  
 1. Lexical and Morphological Analysis  
 Lexical Analysis  
 It focuses on identifying and processing words (or lexemes) in a text. It breaks down  
 the input text into individual tokens that are meaningful units of language such as words  
 or phrases.  
 Key tasks in Lexical analysis:  
 1. Tokenization: Process of dividing a text into smaller chunks called tokens. For example  
 the sentence "I love programming" would be tokenized into ["I", "love",  
 "programming"].  
 2. Part-of-Speech Tagging: Assigning parts of speech such as noun, verb, adjective to  
 each token in the sentence. This helps us to understand grammatical roles of words in  
 the context.  
 Example: Consider the sentence: "I am reading a book."

diversity of human language. Let's discuss 10 major challenges in NLP:

1. Language differences  
 The human language and understanding is rich and intricate and there are many languages  
 spoken by humans. Human language is diverse and thousands of human languages  
 spoken around the world with having its own grammar, vocabulary and cultural nuances.  
 Human cannot understand all the languages and the productivity of human language is  
 high. There is ambiguity in natural language since same words and phrases can have  
 different meanings and different contexts. This is the major challenges in understanding  
 of natural language.

There are complex syntactic structures and grammatical rules of natural languages.  
 The rules are such as word order, verb, conjugation, tense, aspect and agreement. There  
 is rich semantic content in human language that allows speaker to convey a wide range  
 of meaning through words and sentences. Natural Language is pragmatics which means

Question:  
 Phases of NLP  
 1. Syntax Analysis  
 The syntactic analysis is part of the process of developing a sentence and then forming the sentence  
 into a sentence.

Key tasks in Syntax analysis:

1. Language Differences  
 The human language is all about syntax.

The human language has a special syntax that is very strong.

There are many syntactic structures and grammatical rules  
 that can be used in language analysis.

Start coding or generate with AI.

