# DSC670-JyotiDave-Week5-Exercise

October 12, 2025

[1]:
```
# DSC670 - Week5 - Exercise  - Retrieval Augmented Generation
```

Create a chat like the one in the reading (Building Your First RAG System with Python and OpenAI) using any Wikipedia page. Be sure to ask your bot at least two separate questions. You do not need a user interface but are welcome to try one with Streamlit. Be sure to use the least expensive (usually the oldes) OpenAI model.

### 0.1 2023 Cricket World Cup

### 0.2 Load data from Wikipedia

We're going to first extract data from the 2023 Cricket World Cup.

[5]:
```
# Suppress all warnings
import warnings
warnings.filterwarnings("ignore")
```

[6]:
```
#install required libraries
#pip install langchain
```

[7]:
```
#pip install -U langchain langchain-community

#LangChain framework, contains the main logic for working with language models.
# langchain-community - Installs the community integrations, like document␣
 ↪loaders, retrievers, and vector stores (Chroma, FAISS, etc.)
```

[8]:
```
#pip install wikipedia  # requires to search wikipedia article
```

[9]:
```
#Use document loader integrations (like Wikipedia, CSV, PDF, etc.)
from langchain.document_loaders import WikipediaLoader

#Still part of the main LangChain package, so that line is fine.
from langchain.text_splitter import RecursiveCharacterTextSplitter
```

[10]:
```
# Define the topic to serach from wikipedia
search_term = "2023 Cricket World Cup"
# Pull the content directly from wikipedia, load_max_docs =1 will load only 1␣
 ↪document
docs = WikipediaLoader(query=search_term, load_max_docs=1).load()
```

[11]:
```
# We need to split big documents (like Wikipedia articles) into smaller chunks␣
 ↪that fit easily into the model.
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 200, # Each piece of text will be about 100 characters long.
    chunk_overlap  = 20, # Each chunk will share 20 characters with the␣
 ↪previous one.
    length_function = len, # Defines how to measure the chunk size - here, by␣
 ↪counting characters (not tokens).
    is_separator_regex = False, # Tells the splitter that your separators (like␣
 ↪"\n" or " ") are normal strings, not regex patterns.
)

# This takes your Wikipedia document(s) which might be thousands of characters␣
 ↪long and splits them into small overlapping chunks.
data = text_splitter.split_documents(docs)
data[:3] # return the first 3 items of the list.
```

[11]:
```
[Document(metadata={'title': '2023 Cricket World Cup', 'summary': "The 2023 ICC
Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup,
a quadrennial One Day International (ODI) cricket tournament organized by the
International Cricket Council (ICC). It was hosted from 5 October to 19 November
2023 across ten venues in India. This was the fourth World Cup held in India,
but the first where India was the sole host.\nThe tournament was contested by
ten national teams, maintaining the same format used in 2019. After six weeks of
round-robin matches, India, South Africa, Australia, and New Zealand finished as
the top four and qualified for the knockout stage. In the knockout stage, India
and Australia beat New Zealand and South Africa, respectively, to advance to the
final, played on 19 November at the Narendra Modi Stadium in Ahmedabad.
Australia beat India in the final by six wickets, winning their sixth Cricket
World Cup title.\nA total of 1,250,307 spectators attended the matches, the
highest number in any Cricket World Cup to date. The tournament set viewership
records in India, drawing 518 million viewers, with a peak of 59 million
streaming viewers during the final, which alone recorded a record-breaking
global audience of about 300 million viewers worldwide.", 'source':
'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup'}, page_content="The 2023
ICC Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World
Cup, a quadrennial One Day International (ODI) cricket tournament organized by
the International Cricket"),
 Document(metadata={'title': '2023 Cricket World Cup', 'summary': "The 2023 ICC
Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup,
a quadrennial One Day International (ODI) cricket tournament organized by the
International Cricket Council (ICC). It was hosted from 5 October to 19 November
2023 across ten venues in India. This was the fourth World Cup held in India,
but the first where India was the sole host.\nThe tournament was contested by
ten national teams, maintaining the same format used in 2019. After six weeks of
round-robin matches, India, South Africa, Australia, and New Zealand finished as
the top four and qualified for the knockout stage. In the knockout stage, India
```

```
and Australia beat New Zealand and South Africa, respectively, to advance to the
final, played on 19 November at the Narendra Modi Stadium in Ahmedabad.
Australia beat India in the final by six wickets, winning their sixth Cricket
World Cup title.\nA total of 1,250,307 spectators attended the matches, the
highest number in any Cricket World Cup to date. The tournament set viewership
records in India, drawing 518 million viewers, with a peak of 59 million
streaming viewers during the final, which alone recorded a record-breaking
global audience of about 300 million viewers worldwide.", 'source':
'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup'}, page_content='Cricket
Council (ICC). It was hosted from 5 October to 19 November 2023 across ten
venues in India. This was the fourth World Cup held in India, but the first
where India was the sole host.'),
 Document(metadata={'title': '2023 Cricket World Cup', 'summary': "The 2023 ICC
Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup,
a quadrennial One Day International (ODI) cricket tournament organized by the
International Cricket Council (ICC). It was hosted from 5 October to 19 November
2023 across ten venues in India. This was the fourth World Cup held in India,
but the first where India was the sole host.\nThe tournament was contested by
ten national teams, maintaining the same format used in 2019. After six weeks of
round-robin matches, India, South Africa, Australia, and New Zealand finished as
the top four and qualified for the knockout stage. In the knockout stage, India
and Australia beat New Zealand and South Africa, respectively, to advance to the
final, played on 19 November at the Narendra Modi Stadium in Ahmedabad.
Australia beat India in the final by six wickets, winning their sixth Cricket
World Cup title.\nA total of 1,250,307 spectators attended the matches, the
highest number in any Cricket World Cup to date. The tournament set viewership
records in India, drawing 518 million viewers, with a peak of 59 million
streaming viewers during the final, which alone recorded a record-breaking
global audience of about 300 million viewers worldwide.", 'source':
'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup'}, page_content='The
tournament was contested by ten national teams, maintaining the same format used
in 2019. After six weeks of round-robin matches, India, South Africa, Australia,
and New Zealand finished as the')]
```

### 0.3 Storing embeddings in ChromaDB

Next, let's store those chunks of text as embeddings in ChromaDB

[13]:
```
# Install ChromaDB, which is a vector database used in LangChain for storing␣
 ↪and retrieving text embeddings.
#pip install chromadb
```

[14]:
```
# library made by OpenAI for tokenizing text - that is, breaking text into␣
 ↪smaller pieces (tokens) that language models like GPT-4 understand.
#pip install tiktoken
```

[15]:
```
# A LangChain-specific package that provides pre-built classes for working with␣
 ↪OpenAI models (GPT-3.5, GPT-4, GPT-4o, ChatGPT, etc.)
#pip install -U langchain-openai
```

[16]:
```
from langchain.vectorstores import Chroma
from langchain_openai import OpenAIEmbeddings
```

[17]:
```
# Step 1  Create an OpenAI client with API key

## Load required libraries
from openai import OpenAI
import json
import os
from dotenv import load_dotenv

#Load variables from .env file into environment
load_dotenv()

#Create an OpenAI client with API key stored in env file
client = OpenAI(
    # Load the api key securely from env file.
    api_key=os.getenv("OPENAI_API_KEY")
)
embeddings = OpenAIEmbeddings()
```

[18]:
```
# Store document chunks in ChromaDB

store = Chroma.from_documents(
    data,  # text chunks (from RecursiveCharacterTextSplitter).
    embeddings,  # Converts each text chunk into a numeric vector.
    ids = [f"{item.metadata['source']}-{index}" for index, item in␣
 ↪enumerate(data)], # Optional unique IDs for each chunk. Here we combine the␣
 ↪source name and index.
    collection_name="CricketWorldCup-Embeddings", # Name of the vector␣
 ↪collection inside ChromaDB.
persist_directory='db', # Folder where Chroma stores the database on disk.
)
#store.persist()
```

### 0.4 Asking questions about 2023 Cricket World Cup

Now let's use OpenAI, augmented by ChromaDB, to ask some questions about the tournament.

[20]:
```
from langchain.chains import RetrievalQA
from langchain.prompts import PromptTemplate
from langchain_openai import ChatOpenAI
import pprint
```

[21]:
```
# Instead of hardcoding text every time, you define a template with␣
 ↪placeholders, template is a Contains placeholders for variables.
template = """You are a bot that answers questions about World Cup Cricket␣
 ↪2023, using only the context provided.
If you don't know the answer, simply state that you don't know.

{context}

Question: {question}"""

# Creat a custom prompt for the RetrievalQA or LLM chain,question: The user's␣
 ↪query, context: The text chunk(s) retrieved by vector store.
PROMPT = PromptTemplate(
    template=template, input_variables=["context", "question"]
)
```

[22]:
```
# creating a RetrievalQA chain with a custom prompt and GPT-4o-mini ,␣
 ↪temperature makes the answers deterministic (less randomness).
llm = ChatOpenAI(temperature=0, model="gpt-4o-mini")
```

### 0.5 Create RetrievalQA chain

[24]:
```
#qa_with_source is used for viewing both the answer and source documents␣
 ↪returned by return_source_documents=True
qa_with_source = RetrievalQA.from_chain_type(
    llm=llm, # language model "llm" is getting used here
    chain_type="stuff", # all retrieved documents are combined into one prompt␣
 ↪before sending to the model.
    retriever=store.as_retriever(), # Use ChromaDB vector store to find␣
 ↪relevant chunks.
    chain_type_kwargs={"prompt": PROMPT, }, # Use  custom prompt template
    return_source_documents=True, # Return the documents that were used to␣
 ↪answer the query.
)
```

#### 0.5.1 Trying the first quesiton

[26]:
```
#Display the result of the RetrievalQA query
pprint.pprint(
    qa_with_source.invoke({"query": "When and where was World Cup Cricket 2023␣
 ↪held?"}))
```

```
{'query': 'When and where was World Cup Cricket 2023 held?',
 'result': 'The World Cup Cricket 2023 was held from 5 October to 19 November '
           '2023 across ten venues in India.',
 'source_documents': [Document(metadata={'summary': "The 2023 ICC Men's Cricket
World Cup was the 13th edition of the ICC Men's Cricket World Cup, a quadrennial
```

One Day International (ODI) cricket tournament organized by the International Cricket Council (ICC). It was hosted from 5 October to 19 November 2023 across ten venues in India. This was the fourth World Cup held in India, but the first where India was the sole host.\nThe tournament was contested by ten national teams, maintaining the same format used in 2019. After six weeks of round-robin matches, India, South Africa, Australia, and New Zealand finished as the top four and qualified for the knockout stage. In the knockout stage, India and Australia beat New Zealand and South Africa, respectively, to advance to the final, played on 19 November at the Narendra Modi Stadium in Ahmedabad. Australia beat India in the final by six wickets, winning their sixth Cricket World Cup title.\nA total of 1,250,307 spectators attended the matches, the highest number in any Cricket World Cup to date. The tournament set viewership records in India, drawing 518 million viewers, with a peak of 59 million streaming viewers during the final, which alone recorded a record-breaking global audience of about 300 million viewers worldwide.", 'title': '2023 Cricket World Cup', 'source': 'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup'}, page_content='Cricket Council (ICC). It was hosted from 5 October to 19 November 2023 across ten venues in India. This was the fourth World Cup held in India, but the first where India was the sole host.'),
Document(metadata={'title': '2023 Cricket World Cup', 'source': 'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup', 'summary': "The 2023 ICC Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup, a quadrennial One Day International (ODI) cricket tournament organized by the International Cricket Council (ICC). It was hosted from 5 October to 19 November 2023 across ten venues in India. This was the fourth World Cup held in India, but the first where India was the sole host.\nThe tournament was contested by ten national teams, maintaining the same format used in 2019. After six weeks of round-robin matches, India, South Africa, Australia, and New Zealand finished as the top four and qualified for the knockout stage. In the knockout stage, India and Australia beat New Zealand and South Africa, respectively, to advance to the final, played on 19 November at the Narendra Modi Stadium in Ahmedabad. Australia beat India in the final by six wickets, winning their sixth Cricket World Cup title.\nA total of 1,250,307 spectators attended the matches, the highest number in any Cricket World Cup to date. The tournament set viewership records in India, drawing 518 million viewers, with a peak of 59 million streaming viewers during the final, which alone recorded a record-breaking global audience of about 300 million viewers worldwide."}, page_content="The 2023 ICC Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup, a quadrennial One Day International (ODI) cricket tournament organized by the International Cricket"),
Document(metadata={'source': 'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup', 'title': '2023 Cricket World Cup', 'summary': "The 2023 ICC Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup, a quadrennial One Day International (ODI) cricket tournament organized by the International Cricket Council (ICC). It was hosted from 5 October to 19 November 2023 across ten venues in India. This was the fourth World Cup held in India, but the first where India was the sole host.\nThe tournament was contested by ten national teams, maintaining the

same format used in 2019. After six weeks of round-robin matches, India, South Africa, Australia, and New Zealand finished as the top four and qualified for the knockout stage. In the knockout stage, India and Australia beat New Zealand and South Africa, respectively, to advance to the final, played on 19 November at the Narendra Modi Stadium in Ahmedabad. Australia beat India in the final by six wickets, winning their sixth Cricket World Cup title.\nA total of 1,250,307 spectators attended the matches, the highest number in any Cricket World Cup to date. The tournament set viewership records in India, drawing 518 million viewers, with a peak of 59 million streaming viewers during the final, which alone recorded a record-breaking global audience of about 300 million viewers worldwide."}, page_content='On 11 December 2017, India was announced by the ICC as hosts of the 2023 Cricket World Cup; while India had served as a co-host during three previous tournaments (most recently in 2011, which it'),
 Document(metadata={'source': 'https://en.wikipedia.org/wiki/2023_Cricket_World_Cup', 'title': '2023 Cricket World Cup', 'summary': "The 2023 ICC Men's Cricket World Cup was the 13th edition of the ICC Men's Cricket World Cup, a quadrennial One Day International (ODI) cricket tournament organized by the International Cricket Council (ICC). It was hosted from 5 October to 19 November 2023 across ten venues in India. This was the fourth World Cup held in India, but the first where India was the sole host.\nThe tournament was contested by ten national teams, maintaining the same format used in 2019. After six weeks of round-robin matches, India, South Africa, Australia, and New Zealand finished as the top four and qualified for the knockout stage. In the knockout stage, India and Australia beat New Zealand and South Africa, respectively, to advance to the final, played on 19 November at the Narendra Modi Stadium in Ahmedabad. Australia beat India in the final by six wickets, winning their sixth Cricket World Cup title.\nA total of 1,250,307 spectators attended the matches, the highest number in any Cricket World Cup to date. The tournament set viewership records in India, drawing 518 million viewers, with a peak of 59 million streaming viewers during the final, which alone recorded a record-breaking global audience of about 300 million viewers worldwide."}, page_content='in 2011, which it co-hosted with Sri Lanka and Bangladesh), it would mark the first Cricket World Cup to be hosted solely by India.')]}

### 0.5.2 Now print just the answers without displaying the source document

[28]:
```
query = "What countires participated in the final match?"

answer = qa_with_source.invoke({"query": query})["result"]

print(answer)
```

The countries that participated in the final match were Australia and India.

[29]:
```
query = "How many spectators saw the match?"

answer = qa_with_source.invoke({"query": query})["result"]
```

```
print(answer)
```

```
A total of 1,250,307 spectators attended the matches in the World Cup Cricket
2023.
```

[30]:
```
query = "Where was the final match played?"

answer = qa_with_source.invoke({"query": query})["result"]

print(answer)
```

```
The final match was played at the Narendra Modi Stadium in Ahmedabad.
```

[31]:
```
query = "Who won the final match?"

answer = qa_with_source.invoke({"query": query})["result"]

print(answer)
```

```
Australia won the final match.
```

[32]:
```
query = "How many times Australia won the tital?"

answer = qa_with_source.invoke({"query": query})["result"]

print(answer)
```

```
Australia won the Cricket World Cup title six times.
```