RAG-based Question-Answering System Development

Farjad Ahmed, Ghulam Abbas, Saad Ullah

IBA

Submitted to: Dr Sajjad Haider

Table of Contents

Abstract

This project focuses on the development of a Retrieval-Augmented Generation (RAG)-based question-answering system, combining the strengths of retrieval and LLM models. By integrating a powerful language model with an efficient document retriever, the system is designed to deliver accurate, context-aware responses. The report details the architecture, implementation, and evaluation of the system, emphasizing its ability to handle diverse knowledge domains. This approach demonstrates significant potential for enhancing question-answering capabilities in real-world applications

Platform Details

- Only Kaggle was used as the requirements of GPU

Data Details

- Two documents for Corpus of
    1. Laws for Legal Immigration in the 27 EU Member States (640 pages)
    2. IMMIGRATION POLICY (8 Pages)

Algorithms, Models, and Retrieval Methods.

- Initially we tried many embedding and LLM models but most of them didn't work the way we wanted it to work or it was requiring too much computational power such that we couldn't run it on Kaggle, following are the embedding and LLM model tried and failed many times over and over
    1. BAAI/bge-small-en-v1.5 (embedding)
    2. Alibaba-NLP/gte-Qwen2-7B-instruct (embedding)
    3. sentence-transformers/all-MiniLM-L12-v2 (embedding)
    4. hkunlp/instructor-xl (embedding)
    5. mistralai/Mistral-7B-v0.1 (HF_LLM)
    6. mistralai/Mistral-7B-v0.3 (HF_LLM)
    7. EleutherAI/gpt-neo-1.3B (HF_LLM)

    So above are all the failed models due to weird responses or too much computing power required so we selected following for our work to get better results

1. sentence-transformers/all-mpnet-base-v2 (embedding)
2. google/flan-t5-large (HF_LLM)

Additionally, at the end of my notebook I have run another model which work fine but is answering in such a way that it's hard to comprehend

1. facebook/bart-large

sentence-transformers/all-mpnet-base-v2 (embedding): The sentence-transformers/all-mpnet-base-v2 model was selected for its robust performance in generating high-quality embeddings for a wide range of text tasks, including document retrieval in the context of a Retrieval-Augmented Generation (RAG)-based question-answering system. This model, based on the MPNet architecture, has demonstrated superior ability to capture semantic meaning and relationships in text, making it ideal for encoding both query and document representations in the same vector space. It strikes a balance between computational efficiency and high-quality semantic embeddings, allowing for fast retrieval while maintaining relevance in answer generation. Additionally, the model's general-purpose nature, fine-tuned on diverse datasets, ensures its adaptability across various domains, making it a strong fit for the broad scope of queries expected in the question-answering system.

google/flan-t5-large: The google/flan-t5-large model was chosen for its advanced capabilities in natural language understanding and generation, making

it well-suited for a Retrieval-Augmented Generation (RAG)-based question-answering system. Built on the T5 (Text-to-Text Transfer Transformer) architecture, this model is fine-tuned with the FLAN (Fine-tuned Language Net) framework, which improves performance on a variety of instruction-following tasks. This ensures that the model can generate contextually appropriate and coherent answers even when provided with diverse and complex queries. The large-scale nature of the model, combined with its versatility across multiple domains, allows for high-quality response generation, leveraging both retrieved information and pre-trained knowledge. This makes it an ideal choice for enhancing the accuracy and fluency of answers in real-world question-answering applications.

Retrieval methods employed*.*

- In our Retrieval-Augmented Generation (RAG)-based question-answering system, **semantic search** was employed as the primary retrieval method. This approach was chosen for its ability to capture the underlying meaning of both queries and documents, as opposed to keyword-based search, which relies solely on matching exact words. Semantic search leverages embeddings generated by the **sentence-transformers/all-mpnet-base-v2** model, which encodes text into high-dimensional vector representations. These embeddings capture semantic relationships between words and phrases, enabling the system to retrieve relevant documents even when there is no exact match for query terms.
  The key advantage of semantic search over traditional keyword-based search is its ability to handle paraphrasing, synonyms, and context. For example, a query asking "How does machine learning work?" can retrieve relevant documents even if they mention "artificial intelligence" or "data-driven models" instead. This results in more accurate and contextually relevant answers.

Furthermore, semantic search allows the system to rank documents based on the similarity of their embeddings to the query, ensuring that the most relevant information is retrieved. This improves the overall performance of the RAG system by providing the generative model with high-quality, contextually relevant data to formulate precise answers.
In summary, the choice of semantic search was justified due to its ability to enhance retrieval accuracy and relevance, making it an ideal method for supporting the complex, varied queries expected in a question-answering system.

Relevance:

```
Question: What are the general immigration requirements for newcomers in Germany?
Answer: • Proof of sufficient financial means • Health insurance • Basic knowledge of the German language • Third-country national must sign a declaration that he is aware
that the permit can be refused and expulsion can be ordered if false information is provided • Other documents necessary to prove the purpose of immigration

Question: How long does the visa application process take in Germany?
Answer: three months

Question: What are the family reunification policies in Germany?
Answer: Spouses and minor children of German citizens may be issued a residence permit for the purpose of family reunification, as may the parents of minor unmarried German
s for the purpose of care

Question: Are there any specific language requirements to migrate to Germany?
Answer: Knowledge of the German language is not required prior to entry

Question: What types of work visas are available in Germany?
Answer: seasonal work

Question: Can I apply for permanent residency in Germany after a certain period?
Answer: 2.3.2.2 Permanent Residence

Question: what is capital of pakistan
Answer: Pakistan
```

Above are the answer provided by RAG system we developed and below is from internet

What are the general immigration requirements for newcomers in Germany?

- To immigrate to Germany, newcomers generally need to fulfill the following requirements:
  - A valid passport or travel document.
  - Proof of sufficient financial resources (either through employment or other means).
  - Health insurance.
  - A valid visa or residence permit for the intended purpose (e.g., work, study, or family reunification).
  - Meeting specific requirements depending on the type of visa (e.g., skilled worker, student, or family reunification).

How long does the visa application process take in Germany?

- The visa application process in Germany can take anywhere from 2 weeks to several months, depending on the type of visa and the applicant's specific circumstances. For example, work visas may take longer due to the need for approval from the German employment agency.

What are the family reunification policies in Germany?

- Germany allows family reunification for spouses, children, and other close family members of immigrants who have valid residence permits. Applicants must meet certain financial requirements and provide proof of accommodation. The reunification process usually takes several months, and applicants may need to demonstrate basic proficiency in the German language.

Are there any specific language requirements to migrate to Germany?

- For most visa types, applicants are required to demonstrate proficiency in the German language, particularly for family reunification and long-term residency permits. A minimum of A1 level proficiency is typically required for spouses or other family members. For skilled workers and students, the language requirements may vary depending on the nature of the job or course.

What types of work visas are available in Germany?

- Germany offers several work visas, including:
    - **Skilled Worker Visa**: For individuals with qualifications in high-demand occupations.
    - **EU Blue Card**: For highly skilled workers with a university degree and a job offer that meets a certain salary threshold.
    - **Job Seeker Visa**: Allows individuals to enter Germany and search for a job.
    - **Startup Visa**: For entrepreneurs who wish to start a business in Germany.
    - **Seasonal Work Visa**: For temporary, seasonal jobs.

Can I apply for permanent residency in Germany after a certain period?

- Yes, you can apply for permanent residency (also known as a settlement permit) after living in Germany for 5 years on a valid residence permit, though the required period can be shorter under certain circumstances (e.g., for skilled workers or if you speak German at a high level). During this time, you must have stable employment and demonstrate integration into German society.

What is the capital of Pakistan?

- The capital of Pakistan is **Islamabad**.

Evaluation

- for answers which are actually in model we can see retrieval is quite accurate and on a scale from 1-5 evaluation rate: 4
- for irrelevant question in document evaluation rate: 1 (it tried but failed completely)

Reproducibility.

Dependencies:

```
!pip install pypd
!pip install sentence_transformers
!pip install faiss-gpu
!pip install langchain-huggingface
!pip install rag-evaluator
!pip install langchain-community langchain-core
!pip install langchain-community
!pip install transformers
!pip install transformers_stream_generator
!pip install tiktoken
```

Libraries

```
import os
from urllib.request import urlretrieve
import numpy as np
from langchain_community.embeddings import HuggingFaceBgeEmbeddings
from langchain_community.llms import HuggingFacePipeline
from langchain_community.document_loaders import PyPDFLoader
from langchain_community.document_loaders import PyPDFDirectoryLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain_community.vectorstores import FAISS
from langchain.chains import RetrievalQA
from langchain.prompts import PromptTemplate
from langchain_huggingface import HuggingFaceEndpoint
from langchain.llms import HuggingFaceHub
from langchain.document_loaders import TextLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.embeddings import HuggingFaceEmbeddings
from langchain.vectorstores import FAISS
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer, pipeline
```

References

There are no sources in the current document.