Machine Learning Homework 3

Lim You Xuan

7 January 2024

1 Introduction

The implementation of the Retrieval-Augmented Generation (RAG) framework in this project represents a significant advancement in the field of conversational artificial intelligence. Rooted in the premise that large language models (LLMs) can enhance their responses by retrieving facts from external knowledge bases, RAG combines the strengths of information retrieval with text generation. This integration allows LLMs to supplement their internal representation of information with contextually relevant and factual details during the generation process.

2 Project Overview

This project involves the integration of various technologies to create a conversational agent utilizing the RAG framework. The RAG framework enables the model to retrieve information from a preprocessed dataset, generating responses that go beyond the capabilities of traditional language models. The whole process is being categorized into the section below:

1. Data preprocessing
2. Text splitting
3. Text embedding and vector store
4. Conversational chain
5. Results and Demonstration

3 Data preprocessing

The original news dataset consists of 7 columns which are ID, Title, Description, Body, Keywords, Theme and Link. In this project, the data within the Body column will be used to provide information to the model. Due to the RateLimitError by OpenAI, the trim text function is being applied to limit the length of the "Body" column to 300 words. This is to ensure the Request Per Minute (RPM) does not exceed the given maximum amount during the text embedding process.

After that, we create a CSVLoader and load the dataset. This class converts each CSV row into a Document, where key/value pairs represent the row's content. By default, the source for each document is set to the file path value. However, customization is possible using the source column argument, linking each document's source to a specified CSV column.

Additionally, the metadata columns parameter allows the selection of specific columns as metadata, contributing to a more nuanced understanding of the loaded data.

4 Text Splitting

For the text splitting process, it is performed using the CharacterTextSplitter by dividing text into chunks of 1000 characters with a 200 character overlap. This process enhances the granularity of information retrieval, contributing to the effectiveness of the RAG framework.

5 Text embedding and vector store

For this project, we utilize OpenAIEmbeddings and Facebook AI Similarity Search (FAISS) to harness the power of text embeddings and efficient vector storage.

OpenAIEmbeddings, a crucial component, excels in quantifying the relatedness of text strings. By generating embeddings, essentially vectors of floating-point numbers, OpenAIEmbeddings encapsulates the semantic essence of textual information. The distance between these vectors becomes a metric for relatedness; smaller distances denote higher relatedness, while larger distances signify lower relatedness.

The integration of FAISS into the project is pivotal for efficient information retrieval during conversations. Storing and searching over unstructured data is optimized through embedding and storing resultant vectors. At query time, the unstructured query undergoes embedding, and FAISS retrieves embedding vectors that exhibit the highest similarity to the embedded query. In essence, FAISS serves as a dedicated vector store, seamlessly managing the storage of embedded data and executing vector searches with speed and precision.

6 Conversational chain

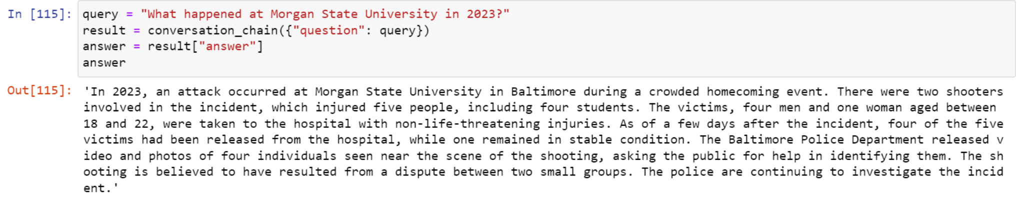
Afterwards, the ChatOpenAI module, which is the OpenAI Chat large language models API, is being configured with specific parameters and served as the language model. The model is selected to be gpt-4 in this project.

The conversation history is managed by the ConversationBufferMemory, enhancing the context-awareness of the model. This memory allows for storing messages and then extracts the messages in a variable.

The ConversationalRetrievalChain then handles the back-and-forth of conversations using chat history and new questions. The whole process has three steps: firstly, it creates a focused "standalone question" by picking out important details from the chat history and new question. Then, it sends this question to the retriever, which grabs the right documents for context. Lastly, a smart language model uses these documents to come up with a response, whether it's based on the new question or considers the original question and chat history. This helps us have smooth and relevant conversations by getting the right info at the right time.

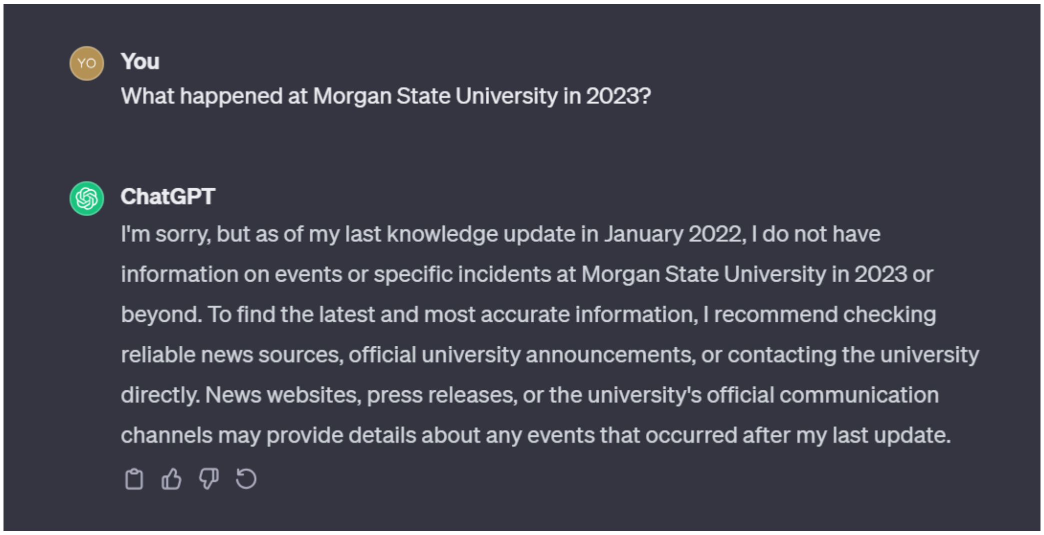
7 Results and Demonstration

An example query, "What happened at Morgan State University in 2023?" is presented to demonstrate the functionality of the RAG framework. The conversation chain processes the query, retrieves relevant information, and provides a coherent response, showcasing the effectiveness of the implemented solution.



```
In [115]: query = "What happened at Morgan State University in 2023?"
          result = conversation_chain({"question": query})
          answer = result["answer"]
          answer

Out[115]: 'In 2023, an attack occurred at Morgan State University in Baltimore during a crowded homecoming event. There were two shooters
          involved in the incident, which injured five people, including four students. The victims, four men and one woman aged between
          18 and 22, were taken to the hospital with non-life-threatening injuries. As of a few days after the incident, four of the five
          victims had been released from the hospital, while one remained in stable condition. The Baltimore Police Department released v
          ideo and photos of four individuals seen near the scene of the shooting, asking the public for help in identifying them. The sh
          ooting is believed to have resulted from a dispute between two small groups. The police are continuing to investigate the incid
          ent.'
```

To demonstrate the powerful of the RAG framework, we have tested the same prompt using ChatGPT. As we all know, the GPT model is only trained up until information around 2022 January, hence it is not able to answer the question effectively.



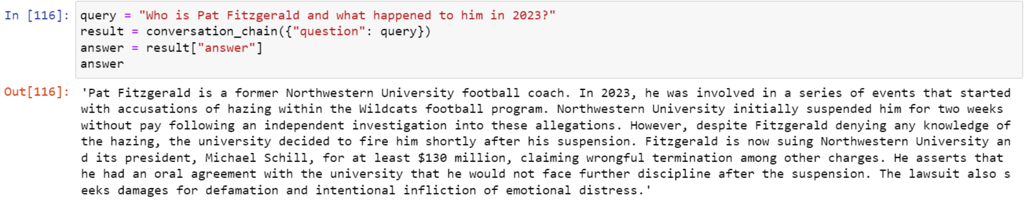
You

What happened at Morgan State University in 2023?

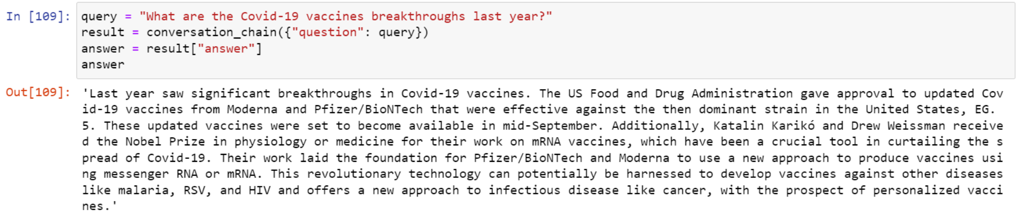
ChatGPT

I'm sorry, but as of my last knowledge update in January 2022, I do not have information on events or specific incidents at Morgan State University in 2023 or beyond. To find the latest and most accurate information, I recommend checking reliable news sources, official university announcements, or contacting the university directly. News websites, press releases, or the university's official communication channels may provide details about any events that occurred after my last update.

Here are some additional examples that showcase the effectiveness of the RAG framework:

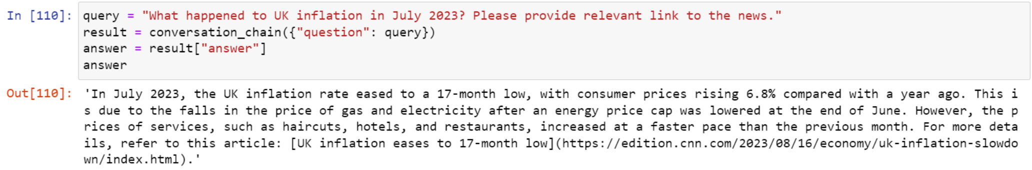


```
In [116]: query = "Who is Pat Fitzgerald and what happened to him in 2023?"
          result = conversation_chain({"question": query})
          answer = result["answer"]
          answer

Out[116]: 'Pat Fitzgerald is a former Northwestern University football coach. In 2023, he was involved in a series of events that started
          with accusations of hazing within the Wildcats football program. Northwestern University initially suspended him for two weeks
          without pay following an independent investigation into these allegations. However, despite Fitzgerald denying any knowledge of
          the hazing, the university decided to fire him shortly after his suspension. Fitzgerald is now suing Northwestern University an
          d its president, Michael Schill, for at least $130 million, claiming wrongful termination among other charges. He asserts that
          he had an oral agreement with the university that he would not face further discipline after the suspension. The lawsuit also s
          eeks damages for defamation and intentional infliction of emotional distress.'
```

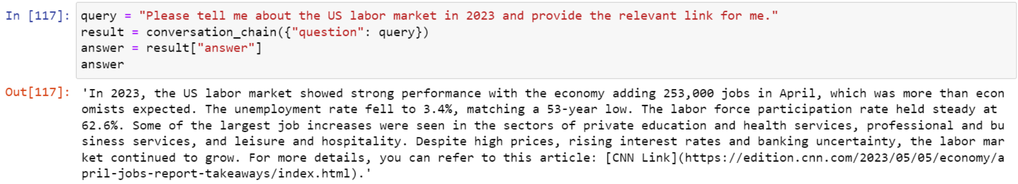


```
In [109]: query = "What are the Covid-19 vaccines breakthroughs last year?"
          result = conversation_chain({"question": query})
          answer = result["answer"]
          answer

Out[109]: 'Last year saw significant breakthroughs in Covid-19 vaccines. The US Food and Drug Administration gave approval to updated Cov
          id-19 vaccines from Moderna and Pfizer/BioNTech that were effective against the then dominant strain in the United States, EG.
          5. These updated vaccines were set to become available in mid-September. Additionally, Katalin Karikó and Drew Weissman receive
          d the Nobel Prize in physiology or medicine for their work on mRNA vaccines, which have been a crucial tool in curtailing the s
          pread of Covid-19. Their work laid the foundation for Pfizer/BioNTech and Moderna to use a new approach to produce vaccines usi
          ng messenger RNA or mRNA. This revolutionary technology can potentially be harnessed to develop vaccines against other diseases
          like malaria, RSV, and HIV and offers a new approach to infectious disease like cancer, with the prospect of personalized vacci
          nes.'
```

With the RAG framework implemented, we can even extract the source of the answer, which is the link of the actual news in our case. This is a cornerstone in the world of LLM as we need to know if the provided answers are sourced from where and determine the reliability of the answer.



```
In [110]: query = "What happened to UK inflation in July 2023? Please provide relevant link to the news."
          result = conversation_chain({"question": query})
          answer = result["answer"]
          answer

Out[110]: 'In July 2023, the UK inflation rate eased to a 17-month low, with consumer prices rising 6.8% compared with a year ago. This i
          s due to the falls in the price of gas and electricity after an energy price cap was lowered at the end of June. However, the p
          rices of services, such as haircuts, hotels, and restaurants, increased at a faster pace than the previous month. For more deta
          ils, refer to this article: [UK inflation eases to 17-month low](https://edition.cnn.com/2023/08/16/economy/uk-inflation-slowdo
          wn/index.html).'
```



```
In [117]: query = "Please tell me about the US labor market in 2023 and provide the relevant link for me."
          result = conversation_chain({"question": query})
          answer = result["answer"]
          answer

Out[117]: 'In 2023, the US labor market showed strong performance with the economy adding 253,000 jobs in April, which was more than econ
          omists expected. The unemployment rate fell to 3.4%, matching a 53-year low. The labor force participation rate held steady at
          62.6%. Some of the largest job increases were seen in the sectors of private education and health services, professional and bu
          siness services, and leisure and hospitality. Despite high prices, rising interest rates and banking uncertainty, the labor mar
          ket continued to grow. For more details, you can refer to this article: [CNN Link](https://edition.cnn.com/2023/05/05/economy/a
          pril-jobs-report-takeaways/index.html).'
```

8 Conclusion

In conclusion, this project successfully implements the RAG framework, showcasing its potential in enhancing conversational AI. The combination of advanced language models, efficient data preprocessing, and effective information retrieval contributes to the project's success in achieving contextually coherent responses.