WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Assignment Cover Sheet

Student Number:	**5504970**
Module Code:	**IB9CW0**
Module Title:	**Text Analytics**
Submission Deadline:	**Wednesday, 13th June 2024 at 12.00 UK Time**
Date Submitted:	**Wednesday, 13th June 2024**
Word Count:	**1499**
Number of Pages:	**6**
Question Attempted: *(question number/title, or description of assignment)*	**Building a Retrieval Augmented Generation (RAG) system, based on a Large Language Model (LLM)**
Have you used Artificial Intelligence (AI) in any part of this assignment?	**Yes, it is required.**

Academic Integrity Declaration
We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that the work is all my own, except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

Table of Contents

1. Introduction

In recent years, the field of Natural Language Processing (NLP) has seen significant advancements with the introduction of Large Language Models (LLMs) like GPT-4 (Raffel et al., 2023). However, these models, despite their impressive capabilities, often struggle with real-time information retrieval and domain-specific queries. This report outlines the development and evaluation of a Retrieval-Augmented Generation (RAG) system designed to enhance the performance of LLMs by incorporating a vector database for efficient information retrieval (Tay, 2022). The project involves advanced techniques for embedding, chunking, and querying data, ensuring optimal results.

2. Methodologies

2.1. Domain Identification

The chosen domain for this project is chronic or deadly diseases, specifically in the UK region. This domain was selected due to the vast amount of publicly available data and the critical importance of accurate information retrieval in this field.

2.2. Knowledge-Base Sourcing

This project uses Ollama, a powerful API that provides access to various NLP models, including GPT-4 (Sun et al., 2023). Ollama was used to handle the LLM-based tasks, ensuring the latest models and features are utilized for generating responses (Rahali and Akhloufi, 2023). Ollama's API was integrated to pipeline for efficient querying and response generation, leveraging its optimized infrastructure to achieve faster processing times compared to local setups.

The dataset or knowledge for this project was sourced from publicly available online articles related to deadly diseases. Specifically, three web pages are chosen to provide up-to-date information, which are:

1. https://uk.style.yahoo.com/leading-causes-of-death-uk-144557562.html",
2. https://uk.style.yahoo.com/stroke-symptoms-causes-treatment-120059074.html
3. https://uk.style.yahoo.com/symptoms-uk-common-cancer-cases-soar-under-50s-175516176.html

2.3. Test Queries

To evaluate the performance of the LLM and RAG system, three queries or questions were determined. These queries were chosen to test the system's ability to retrieve and generate

responses based on both general and specific information. The details of these queries are discussed in the “Result and Evaluation” part.

2.4. Chunking and Embedding

The collected documents were chunked and embedded using a pretrained embedding model. The steps involved were:

- Chunking: The text is split into smaller, manageable chunks of 200 words with a 20-word overlap to facilitate processing. This chunking method produced 36 chunks from the input data, improving the efficiency of subsequent embedding and retrieval tasks.
- Embedding: Each chunk was then embedded using a pretrained model from HuggingFace named Beijing Academy of Artificial Intelligence (BBAAI/bge-small-en-v1.5), for transforming text data into embeddings. This model is optimized for efficiency and accuracy, making it suitable for various natural language processing tasks such as information retrieval, clustering, and semantic search (Reimers and Gurevych, 2019; Sun et al., 2023). The embedded chunks were stored in a vector database, specifically using ChromaDB, which allows efficient retrieval based on semantic similarity.

2.5. Prompt Template Development

An appropriate prompt template was developed to query the LLM effectively. The template was designed to include the context retrieved from the vector database along with the user query to provide the LLM with additional information as well as making it always answers the questions, even there is no new knowledge is added.

2.6. Pipeline Development

The pipeline consists of the following steps:

1. Data collection: Sources from relevant online articles on deadly diseases ensure up-to-date information.
2. Chunking: Splits documents into 200-word chunks with 20-word overlaps for context.
3. Embedding: Uses the BAAI/bge-small-en-v1.5 model for efficient and accurate text embeddings.
4. Vector database: Stores embeddings in ChromaDB for efficient semantic retrieval.
5. Query embedding: Embeds user queries using the same model for consistency.
6. Information retrieval: Searches ChromaDB for relevant chunks.
7. Prompt construction: Combines retrieved chunks with the query.
8. Response generation: Uses Ollama API for optimized, accurate responses.

These steps enhance LLM performance in real-time information retrieval and domain-specific queries.

3. Results and Evaluation

The RAG system effectively answered queries, accurately identifying the leading cause of death in the UK in 2022 (Lewis et al., 2020). When asked, "What is the leading cause of death in the UK in 2022?", it correctly responded with "Dementia and Alzheimer's disease," demonstrating its ability to retrieve and provide accurate information efficiently.

The evaluation of different response modes using the developed prompt template provided insightful results. The analysis was conducted on three different queries. Each query was tested across four response modes: default, refine, compact, and tree_summarize. Here are the detailed findings:

1. **First query**: "What are the leading causes of death in the UK?"
    a. **Default mode**

        Response: The leading causes of death in the UK are dementia and Alzheimer's disease, heart disease, lower respiratory diseases, strokes, lung cancer, influenza, and pneumonia.

        Response Time: 200.91 seconds.

    b. **Refine mode**

        Response: Based on the information provided, the top two leading causes of death for men in the UK are cancer and lung cancer. For women, they are heart disease and lung cancer.

        Response time: 532.22 seconds.

    c. **Compact mode**

        Response: The leading causes of death in the UK are dementia and Alzheimer's disease, heart disease, lower respiratory diseases, strokes, and lung cancer.

        Response time: 237.57 seconds.

    d. **Tree_summarize mode**

        Response: The top five causes of death in the UK as per the data provided are dementia and Alzheimer’s disease, heart disease, lower respiratory diseases, strokes, and lung cancer.

        Response time: 219.78 seconds.

2. **Second query**: "What is the cause of death of Christine McVie?"
    a. **Default mode**

        Response: The cause of death of Christine McVie was a stroke.

        Response time: 188.37 seconds.

    b. **Refine mode**

Response: The cause of death of Christine McVie was a heart attack. She had a history of heart problems, which eventually led to her passing.

Response Time: 159.03 seconds.

c. Compact mode

Response: The most common type of stroke, which Christine McVie had, causes many deaths every year.

Response time: 178.41 seconds.

d. Tree_summarize mode

Response: The cause of death of Christine McVie was a stroke.

Response time: 204.64676547050476 seconds.

3. **Third query**: "What cancer does Married At First Sight star Mel Schilling suffer from?"

a. Default mode

Response: The passage does not provide enough information to accurately determine what type of cancer Mel Schilling suffers from. It only mentions that she has been diagnosed with colon cancer.

Response time: 200.38 seconds.

b. Refine mode

Response: Unfortunately, the passage does not give any information about Mel Schilling's specific cancer diagnosis. Please double-check your query and try again!

Response Time: 129.26 seconds.

c. Compact mode

Response: Colon Cancer.

Response time: 153.25 seconds.

d. Tree_summarize mode

Response: According to the context information provided, it is mentioned that Mel Schilling has been diagnosed with colon cancer. However, due to the AI's restrictions on referencing the given context directly, I cannot provide a specific answer without access to additional sources or data.

Response time: 196.04 seconds.

The result summary and insights for each query are explained as follows.

1. First query: The default and compact modes provided the most detailed responses, with reasonable times, while refine mode was the slowest but offered a specific focus on leading causes for men and women separately. Tree_summarize was efficient but took slightly longer than the default mode.
2. Second query: The refine mode provided the quickest response with a detailed explanation, while the tree summarize mode took the longest time. Default and compact modes offered accurate responses with slightly longer times than refine mode.

3. Third query: Refine mode gave the fastest response while compact mode provided an accurate answer quickly. The default mode was the slowest and less specific. Tree_summarize mode gave a comprehensive but slightly slower response compared to compact mode.

Based on the result, refine mode gives the quickest responses but sometimes lacks detail, compact mode is balanced in response time and accuracy, default mode is generally detailed but slow, and tree_summarize mode is comprehensive but generally slower. It can be noted compact mode is more suggested due to its balance between response time and detailed, accurate responses.

The comparison between LLM with RAG and LLM without RAG on three queries also provides interesting insights. LLM with RAG provided detailed responses with varying response times. In contrast, LLM without RAG delivered incomplete or generic responses. For instance, LLM without RAG listed causes of death in the UK but lacked specific context. Additionally, it failed to accurately answer questions about Christine McVie's cause of death and Mel Schilling's cancer diagnosis. Therefore, LLM with RAG demonstrated enhanced accuracy and contextual relevance, making it more effective for precise information retrieval.

4. Conclusions and Future Improvements

The evaluation of the RAG system against LLM-only responses demonstrates the significant advantages of integrating a vector database for efficient and accurate information retrieval. The RAG system consistently provided more detailed and contextually relevant responses compared to the LLM alone, which often returned generic or incomplete answers. Among the response modes, the compact mode proved to be the most balanced, offering both speed and accuracy.

Future improvements could include refining the chunking and embedding processes to enhance retrieval efficiency further. Additionally, expanding the knowledge base to cover more diverse topics and regularly updating it to ensure current information can significantly improve the system's performance. Implementing more advanced response synthesis techniques could also enhance the depth and precision of the generated answers.

References

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. Available at: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html [Accessed 2 June 2024].

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. Available at: https://arxiv.org/abs/2301.00745 [Accessed 2 June 2024].

Rahali, A., Akhloufi, M. A. (2023). End-to-End Transformer-Based Models in Textual-Based NLP. AI, 4(1), 54-110. Available at: https://www.mdpi.com/2673-2688/4/1/4 [Accessed 7 June 2024].

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Available at: https://arxiv.org/abs/1908.10084 [Accessed 5 June 2024].

Sun, X., Dong, L., Li, X., Wan, Z., Wang, S., Zhang, T., Li, J., Cheng, F., Lyu, L., Wu, F., Wang, G. (2023). Pushing the Limits of ChatGPT on NLP Tasks. Available at: https://paperswithcode.com/paper/pushing-the-limits-of-chatgpt-on-nlp-tasks [Accessed 10 June 2024].

Tay, A. (2022). The Power of Retrieval Augmented Generation (RAG) LLM-based Academic Search Engines. HKUST Library. Available at: https://library.hkust.edu.hk/sc/retrieval-augmented-generation-based-academic-search-engines/ [Accessed 7 June 2024].