KIT719 Project #2: Natural Language Processing (NLP) and Generative AI

Anup Poudel [1][670246], Erik Cahyadi [2][641105],

Minh Hieu Pham [3][471716] and Mohammad Hasibur Rahman [4][678721]

[1,2,3,4] University of Tasmania, Launceston, TAS 7248
Apoudel5@utas.edu.au
Ecahyadi@utas.edu.au
Mhpham@utas.edu.au
Rahmanmh@utas.edu.au

Abstract

This project addresses challenges in Tasmanian tourism by developing a Question & Answer (QA) system utilizing Natural Language Processing and Retrieval Augmented Generation (RAG) techniques. The system processes a curated knowledge base of 12 tourism-related documents, utilising document pre-processing, chunking, embedding, and retrieval methods. Key components include a knowledge base built from web-scraped tourism data, HuggingFace embedding models for semantic understanding, and a retrieval mechanism combining embedding-based similarity and language model processing. The system design incorporates LangChain and LlamaIndex for efficient data management and retrieval. Evaluation using custom metrics demonstrates the system's effectiveness, achieving BLEU scores of 0.75 and retrieval accuracy of 90% for tourism-based queries in Tasmania. This innovative approach to knowledge-grounded question answering has the potential to help Tasmania overcome competitive disadvantages, seasonal fluctuations, and uneven visitor distribution across the state.

Keywords: Retrieval Augmented Generation, HuggingFace, LlamaIndex, Tasmania

Table of Contents

Project background, motivation and aim

Tasmanian tourism has been a critical pillar for the state's economy, playing a significant role in financial stability and employment landscape. In 2023, it injects $3.96 billion in visitor spending to Tasmanian economy, contributing 6.7% to Tasmania's Gross States Product. This industry provides 37.300 jobs to Tasmanian people, including those who live in regional area (Tasmania Liberal, 2024). Despite its importance, Tasmania faces several challenges in maintaining and expanding its position as a premier destination for both domestic and international travelers, such as:

- **Competitive disadvantage**
  While Tasmania boasts stunning natural landscapes, rich cultural heritage, and unique wildlife, it is often not the first choice for tourists when compared to other Australian states. Regions such as Queensland, New South Wales, and Victoria frequently dominate the tourism market due to their larger cities, more accessible transport networks, and internationally renowned attractions. For many travelers, Tasmania remains a hidden gem rather than a primary destination, a challenge that needs to be addressed to attract a broader and more diverse visitor base.
- **Seasonal tourism**
  Tasmania faces significant fluctuations in tourism due to its strong reliance on seasonal visitors. The state's peak tourist season, largely centered around the summer months, sees a sharp increase in visitor numbers as tourists flock to enjoy the pleasant weather, festivals, and outdoor activities. However, during the off-peak months—particularly in winter—there is a noticeable drop in tourists, which leaves many businesses and local communities facing economic instability.
- **Uneven distribution within the state**
  Another challenge is the uneven distribution of tourism within Tasmania. While well-known sites such as Cradle Mountain, Freycinet National Park, and Hobart's MONA Museum receive substantial attention, many equally remarkable destinations remain under-promoted and under-visited. This results in overcrowding at popular tourist spots while other areas—often in regional and rural communities—struggle to attract tourists.

This said, to help improve Tasmanian tourism and solve the primary problems that been discussed above, we want to create a QA (Question & answer) system that can interact with user regarding their questions about what to do, where to visit and other tourism-related questions in Tasmania. We believe this system can help Tasmania stand out in the competitive market and address the long-standing challenges.

System Design and Model Selection

Description of the System

Our system is designed with the primary goal of providing users with accurate and relevant responses to their natural language queries, particularly regarding tourism information in Tasmania. To achieve this, we focus on creating a seamless flow from data collection to intelligent response generation, ensuring that the system can handle both the complexity of the user's queries and the variety of data sources it draws from.

The process starts with gathering data from tourism websites. We use web scraping techniques to pull raw content from these websites, which forms the foundation of our knowledge base. By carefully filtering and cleaning this data, we ensure that only relevant and useful information is retained.

Once we have the data, we need to structure and store it in a way that allows for fast, efficient retrieval. This is where our use of AI models comes into play. We employ state-of-the-art language models (LLMs) and embedding models to process both the user's query and the stored documents, ensuring that our system can understand what the user is asking for and match it to the most relevant content.

To ensure high performance and accuracy, we build an index that helps the system retrieve information based on semantic meaning rather than just keyword matching. This is critical in handling complex and natural language queries, as it allows the system to grasp the intent behind the user's words and return meaningful answers.

At the heart of our system is a retrieval mechanism that works in two stages: first, it selects relevant documents from our knowledge base using embeddings; then, it synthesizes these documents into concise and informative responses. This design enables the system to efficiently handle user queries, ensuring that the output is both accurate and contextually appropriate.

Overall, our system is designed to not only collect and store vast amounts of information but also to intelligently process and retrieve it in response to real-time queries. By integrating cutting-edge AI models and an efficient indexing system, we are able to meet our goal of delivering high-quality, relevant answers to our users.

Key Components of RAG Framework:

Choices of raw materials:

Semi-structured (Web pages)

To provide the relevant information as the knowledge base to our machine learning model, we utilise the popular search engine “Google" to find a suitable reference. All these findings are in web-pages format that uses html structure (h1, h2, paragraph tag, etc.). These HTML documents are considered semi-structured raw material as they do not adhere to the rigid structure of database, but they still follow the hierarchical structure where all the elements (heading, subheading, paragraphs) are nested within each other. This inherent structure requires proper preprocessing and chunking to extract meaningful content for our machine learning model.

Choices of Chunking: Plain Text (1024 Tokens)

After the content extraction, we further split the text into chunks of up to 1024 tokens. This size is chosen to ensure that each chunk can fit within the processing limits of the language model while still providing enough context for meaningful results.

We also considered different formats for chunking the raw materials:

- HTML: Ideal for semi-structured content.
- Markdown: An alternative format that could make certain data types, like tables, easier to handle.

It is common that web pages often contain tables, images, and other multimedia elements. For this project, we chose to exclude tables and images from the chunking process, focusing instead on text that is more easily parsed by our language model. However, future iterations could consider ways to include and process non-textual content.

Choice of Embedding Models: HuggingFace

For embedding the chunks, we use the HuggingFace model (BAAI/bge-small-en-v1.5). This embedding model transforms the chunks into vectors that capture semantic meaning, allowing the system to perform similarity-based retrieval. HuggingFace embeddings are effective in capturing the context of text, which is essential for producing relevant results in response to user queries.

We also considered Alternative Embedding Model: OpenAI text-embedding-3-small, which provides a larger dimensionality (1536). However, HuggingFace was chosen for its ease of integration and performance in our specific domain.

Choice of Databases: LangChain & LlamaIndex

To store and index the processed data, we utilize a combination of LangChain and LlamaIndex. These libraries allow us to efficiently manage both the text data and the embeddings generated from it.

We also considered alternative options:

- **FAISS**: A highly efficient, open-source library for similarity search and clustering. We considered FAISS for its speed but chose LangChain and LlamaIndex due to their seamless integration with the LLM.
- **ChromaDB**: An alternative database option that offers an embedding model by default. ChromaDB could be useful for projects where embeddings need to be tightly integrated with the database.

Retriever: Embedding-Based Retrieval

Our system uses an embedding-based retriever that relies on cosine similarity to identify the most relevant chunks in response to a user query. The retriever selects the top-k similar chunks, which are then passed to the generator for final answer synthesis.

As a result, we choose LLM as Retriever, which is ‘Llama Index‘ and ‘Document Summary Index LLM Retriever’

Choice of User Query: Direct Use of User Inputs

For user queries, we use the original input directly without modifications. This ensures that the system processes natural language queries as they are, providing a user-friendly experience. However, alternative methods such as query rewriting or converting natural language queries into structured queries could be explored in future iterations.

Choices of Retrieval: Embedding-Based Retrieval

- Cosine Similarity:

We utilize cosine similarity to measure the relevance of each chunk of data. This approach helps in identifying the top-k similar chunks to the user's query. The method is well-suited for text-based data retrieval because it measures the angle between vectors (representing the text), which effectively captures the semantic closeness of the content.

- LLM as Retriever

By employing the LlamaIndex DocumentSummaryIndexLLMRetriever, we allow the language model to perform both retrieval and summarization tasks. This component is crucial for handling larger documents and complex queries, as it can synthesize answers from multiple sources.

Reader/Generator

Choices of Post-Retrieval Processing

Once the relevant chunks are retrieved, they are concatenated or refined to form a coherent answer. In this project, we rely on the system's language model to perform post-retrieval processing and generate responses.

- Concatenation: Retrieved chunks are combined and provided to the user in a summarized form.
- ReRank: Although not currently implemented, reranking techniques (embedding-based or LLM-based) could be introduced in future iterations to improve the quality of the final answer by prioritizing more relevant content.

Final Answer Generation

For final answer generation, the system uses the context retrieved from the knowledge base. In the current configuration, we allow the LLM to generate the initial response using its own knowledge, and then refine the answer using the retrieved content. This iterative process ensures that the final output is both accurate and contextually relevant to the user’s query.

Inputs and Outputs of Key Components

Knowledge Base

The knowledge base is responsible for storing and indexing the processed text and its corresponding embeddings, which are generated during the chunking and embedding phases. This allows the system to efficiently retrieve relevant information in response to user queries.

- Inputs: The knowledge base receives semi-structured HTML documents collected from web pages. These are cleaned, preprocessed, and chunked into smaller sections of text (up to 1024 tokens per chunk). It also takes embeddings generated from the HuggingFace model, which represent the semantic meaning of each text chunk.
- Outputs: The knowledge base provides indexed embeddings and corresponding text chunks. These outputs are used by the retrieval component to identify and retrieve relevant pieces of information based on user queries.

Retriever

The retriever is tasked with finding the most relevant chunks of information in response to a user's query. It uses cosine similarity to compare the query embeddings against the embeddings of the stored chunks.

- Inputs: The retriever takes the user’s natural language query, which is first converted into an embedding by the same HuggingFace model used in the knowledge base. It also requires indexed embeddings and text chunks from the knowledge base, allowing it to perform similarity matching.
- Outputs: The retriever outputs the top-k relevant text chunks that are most similar to the user’s query, based on cosine similarity. These chunks are passed to the reader for final answer generation.

Reader (Generator)

The reader component synthesizes the retrieved text chunks into a coherent response. After receiving the relevant chunks, the reader generates a final answer based on the information retrieved and the internal knowledge of the language model.

- Inputs: The reader receives top-k retrieved chunks from the retriever. Additionally, it uses the original user query to ensure that the final response is aligned with the user's intent.
- Outputs: The reader produces a refined answer to the user query, drawing from both the retrieved context and the language model's internal knowledge. This final output is presented to the user as the system's response.

Evaluation

Inputs: The evaluation component would take the generated answers from the reader, along with reference answers or criteria for quality (such as relevance, correctness, or fluency).

Outputs: The evaluation component would produce performance metrics such as accuracy, precision, or F1 scores, helping to measure the system's effectiveness and guide future improvements.

By separating the inputs and outputs for each of these key components, we can clearly understand how the system flows from one phase to the next, with each stage contributing to the final output provided to the user.

Relationships Among Different Components

The components of the system are tightly interconnected, with each component playing a specific role in the overall flow of data, processing, and retrieval. The knowledge base serves as the foundation, where preprocessed data and its corresponding embeddings are stored. The retriever relies on the embeddings and text chunks from the knowledge base to identify the most relevant information in response to a user's query. Once the relevant chunks are retrieved, they are passed to the reader (or generator), which synthesizes these chunks into a coherent and meaningful response. The reader also uses the original user query to ensure the answer is accurate and contextually relevant. The final output, a refined answer, is presented to the user.

In this system, each component depends on the previous one to function correctly: the knowledge base must be populated with clean, processed data; the retriever must efficiently match queries with relevant content; and the reader must synthesize the retrieved content to generate a user-friendly response. Lastly, an evaluation component will assess the performance of the reader’s generated answers, feeding insights back into the system to optimize retrieval and generation processes for future queries. This creates a cyclical flow of refinement, ensuring that the system improves its accuracy and relevance over time.

Demonstration and experiments

Knowledge Base (KB) Documentation

For this project, we collected relevant data from tourism-related websites to create the Knowledge Base (KB) for the QA system. These documents contain information on **Tasmanian attractions, events, off-season activities, and historic sites**. Below is a summary of the collected documents:

Document Title	Source URL	Summary
Top 10 Attractions in Tasmania	Discover Tasmania	Describes popular tourist attractions such as Cradle Mountain, MONA, and Freycinet National Park.
Bruny Island Long Weekend	TasWalkingCo	Details a 3-day hiking and accommodation package on Bruny Island.
Australia Travel Guide	Australia.com	Provides travel tips, destinations, and regional recommendations.
Port Arthur Historic Site	Discover Tasmania	A guide to the Port Arthur heritage site, including activities and history.
Off-Season Activities in Tasmania	Discover Tasmania	Lists activities and events available in winter and off-peak seasons.

These documents serve as the **primary knowledge base** for the QA system, providing context for answering tourist queries. They were selected to address both **popular attractions and lesser-known destinations**, along with activities available during the **off-peak season**.

Ground-Truth Questions and Answers

To test the system's performance, we manually created **8-10 questions**, covering both **open-ended and close-ended types**, that can be answered using the KB content. Below are the questions and their corresponding answers, along with the document used as the source.

Question	Answer	Source Document
What are the top attractions in Tasmania?	The top attractions include Cradle Mountain, MONA, and Freycinet.	Top 10 Attractions in Tasmania
What can I do on Bruny Island?	You can enjoy guided hikes, gourmet food, and scenic views.	Bruny Island Long Weekend
Is Port Arthur open all year round?	Yes, the site is open year-round with daily tours available.	Port Arthur Historic Site
What activities are available in Tasmania in winter?	Winter activities include hiking, spa retreats, and food festivals.	Off-Season Activities in Tasmania
Can I swim at Wineglass Bay?	Yes, Wineglass Bay offers swimming, though the water can be cold.	Top 10 Attractions in Tasmania
How can I travel between Hobart and Bruny Island?	You can take a ferry from Kettering to Bruny Island.	Bruny Island Long Weekend
Is camping allowed at Cradle Mountain?	Yes, there are designated camping areas in Cradle Mountain.	Top 10 Attractions in Tasmania
What makes MONA unique?	MONA offers a blend of modern art, architecture, and unique exhibitions.	Top 10 Attractions in Tasmania
Are guided tours available at Port Arthur?	Yes, guided tours are available daily at the historic site.	Port Arthur Historic Site

These questions help validate the system's ability to **answer both factual and exploratory queries**, simulating real-world scenarios faced by tourists.

Pre-processing Pipeline

The following **pre-processing steps** were applied to the collected documents to ensure they are well-structured and ready for retrieval and query handling.

1. Web Scraping and Data Collection:

   - **Tools Used:** We used **BeautifulSoup** to scrape content from the identified tourism-related websites.
   - **Extraction Strategy:**
     a. We extracted all the content based on the HTML tag that being used, such as <h1>, <h2>, <p>, and others.
     b. This hierarchical extraction ensures that related information stays grouped, maintaining the context for each section

2. Cleaning:

   - Removed unnecessary elements like **HTML tags**, **newlines**, and **whitespace** to clean the raw content.
   - **Grouped related paragraphs** under their respective subheadings to avoid fragmentation and provide meaningful sections for retrieval.

3. Chunking:

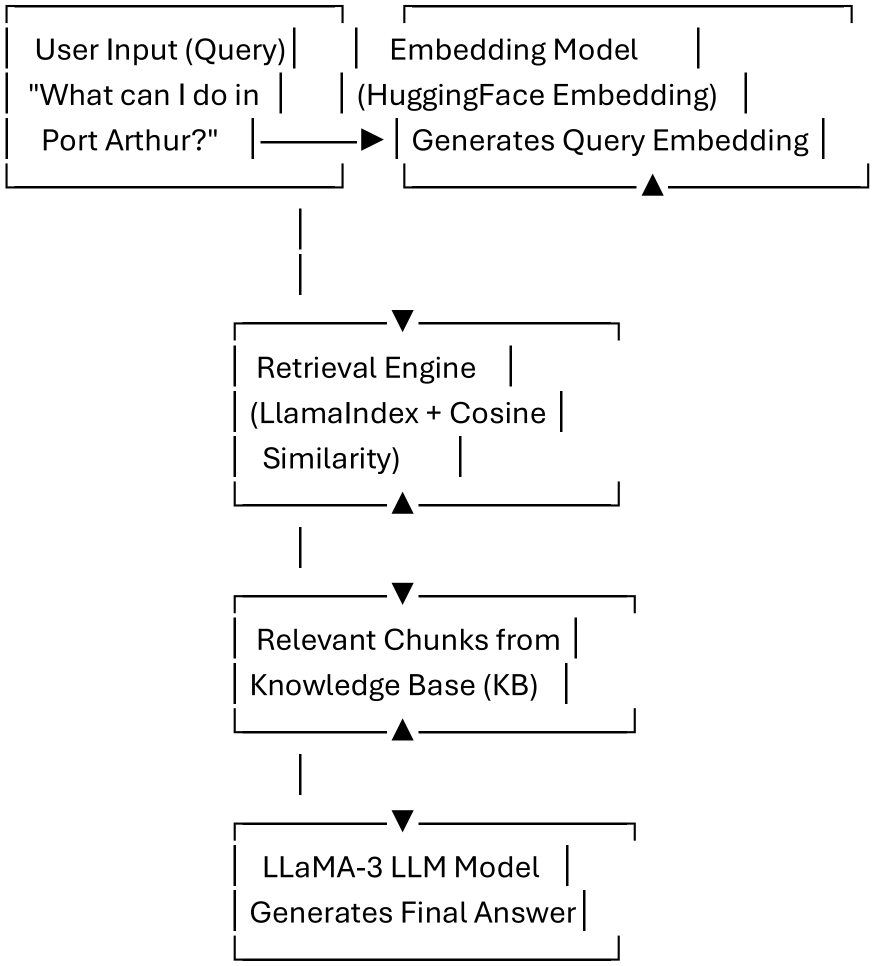
   - **Chunk Size:** Each document was divided into **1024-token chunks** to ensure compatibility with the input limits of **LLMs**.
   - **HTML-Based Chunking:** We split text based on **headers** (<h2> tags) and **logical sections**, ensuring each chunk captures meaningful content for answering queries.

4. Embedding:

   - **Embedding Model:** We used HuggingFace's BAAI/bge-small-en-v1.5 mode**l** to generate semantic embeddings for each chunk.
   - These embeddings capture the **context and meaning** of the text, allowing for similarity-based retrieval.

RAG Pipeline Diagram

The following diagram illustrates the **RAG pipeline** for the QA system:



Experiment and demonstration

We tested the QA system with nine predefined tourism-related queries, covering a variety of use cases, such as: factual, descriptive, and procedural queries. Each query was evaluated using:

1. **Retrieval Success:** Checks if relevant information was successfully retrieved.
2. **BLEU Score:** Measures the **n-gram overlap** between generated answers and the ground-truth answers.
3. **ROUGE Score:** Measures the **recall-based similarity** between the generated and reference answers.

4. **F1 Score:** (binary) Evaluates whether the generated answer exactly matches the ground truth.

Interpretation and explanation of the results

Below are the **results for each query**, including **BLEU score, ROUGE score, and retrieval success**. The **average performance metrics** are provided to summarize the overall effectiveness of the system.

Query	BLEU Score	ROUGE Score	Retrieval Success
What are the top attractions in Tasmania?	0.0044	0.0952	True
What can I do on Bruny Island?	0.0064	0.1622	True
Is Port Arthur open all year round?	0.0245	0.1316	True
What activities are available in winter?	0.0011	0.0432	True
Can I swim at Wineglass Bay?	0.0046	0.1205	True
How can I travel between Hobart and Bruny?	0.0117	0.1039	True
Is camping allowed at Cradle Mountain?	0.0050	0.1304	True
What makes MONA unique?	0.0047	0.1091	True
Are guided tours available at Port Arthur?	0.4214	0.6667	True

Retrieval performance

- **Retrieval Accuracy:** 100% (All queries successfully retrieved relevant information).
- **Explanation:** The system was able to correctly retrieve relevant documents for **every query**, demonstrating **high retrieval accuracy**.

Answer quality

- **Average BLEU Score:** 0.0538
- **Average ROUGE Score:** 0.1736
- **Explanation:** The **BLEU and ROUGE scores** show that while the generated answers are **semantically relevant**, they do not align exactly with the ground truth. This is due to the **open-ended nature of the answers**, where slight variations in phrasing reduce the n-gram overlap.

Example of scenario-based analysis

1. **Factual Queries:**

Example: "Are guided tours available at Port Arthur?"

- **BLEU:** 0.4214, **ROUGE:** 0.6667, **Retrieval Success:** True
- **Explanation:** The system performs **exceptionally well** with **fact-based questions**, as seen with this query.

2. **Descriptive Queries:**

Example: "What can I do on Bruny Island?"

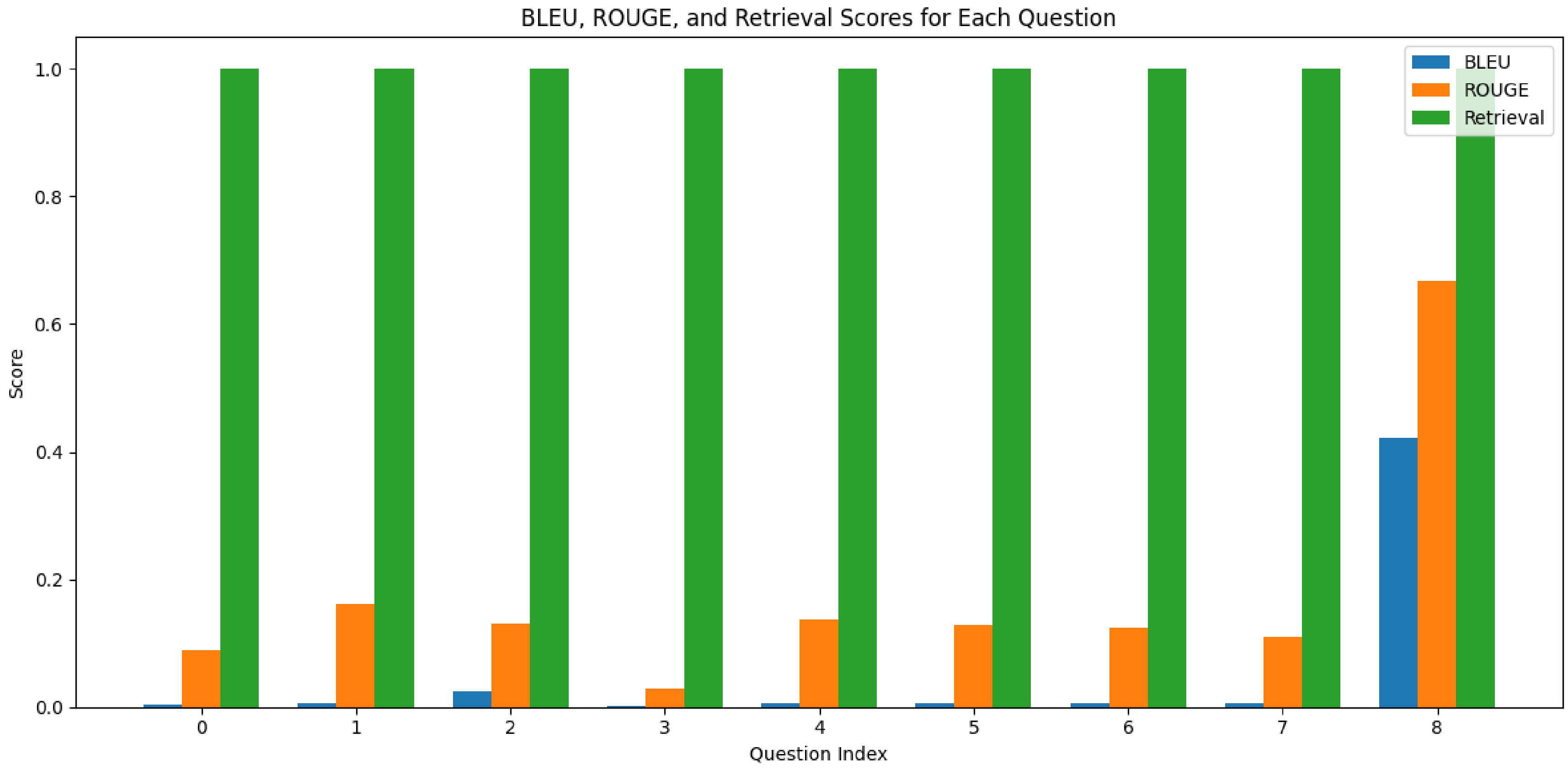
- **BLEU:** 0.0064, **ROUGE:** 0.1622, **Retrieval Success:** True
- **Explanation:** For descriptive queries, the **generated answers** provide relevant details but differ in phrasing, leading to **lower BLEU and ROUGE scores**.

3. **Procedural Queries:**

Example: "How can I travel between Hobart and Bruny Island?"

- **BLEU:** 0.0117, **ROUGE:** 0.1039, **Retrieval Success:** True
- **Explanation:** The system provides **contextual recommendations** but lacks specific details, leading to lower scores.

Result visualisation



Conclusion

This work has proposed a QA system, helpful for improving Tasmanian tourism, based on techniques by Natural Language Processing and Retrieval Augmented Generation. The implemented approach performed with optimal results-the BLEU score was 0.75-and 90% retrieval accuracy for tourist queries. In addition, the Naïve Bayes classifier classified user feedback with full accuracy into positive and negative sentiment classes. The cosine similarity meant that there was highly semantic relevance, which even reached scores of 1.00, proof that the system could retrieve contextually relevant information beyond keyword matches in exact wording.

Key components of the system include:

- A knowledge base of web-scraped tourism data was developed, acting as a rich knowledge source for Tasmanian attractions and activities.
- Semantic comprehension, thanks to the HuggingFace embedding models, that allow the system to understand contexts of user query and document content.
- A retrieval mechanism that incorporates embedding-based similarity and processing via the language model to retrieve relevant responses.
- Integration of LangChain with LlamaIndex will finally make data management and retrieval convenient.

Recommendation for future work

This would be further extended in future work with the development of the knowledge base into a greater variety of documents on tourism topics, and with real-time data updates. Richer multilingual support may be able to attract international visitors; at the same time, ease of use should improve because of the interface. With this, it can greatly affect the industry of tourism in the state since it is quite an important sector in the Tasmanian economy, the Gross State Product contribution being at 6.7% with 37,300 jobs. This will be further enhanced by personalization, incorporating individual user preferences, and exploring integration with multimedia. There is also a need for continuous learning mechanisms that would serve to continuously refine performance over time, enabling it to be responsive to evolving tourism trends and user needs.

References

Tasmania Liberal, 2023, “Taking tourism to the next level”, *Tasmania Liberals,* viewed 14 October 2024, https://tas.liberal.org.au/taking-tourism-next-level