TeamStation Technical Challenge

Juan Camilo Parra Diaz

I. Exploratory Data Analysis (EDA)

For detailed information on the Exploratory Data Analysis, please refer to the following README file: `EDA/README.md`.

I-A. RAG Proposal

For the provided questions and answers, it is proposed that this information be presented as a cohesive text. Specifically, the index should integrate both the question and answer into a single semantic unit.

I-B. Missing Data Analysis

The pattern of missing values is illustrated in the following image:

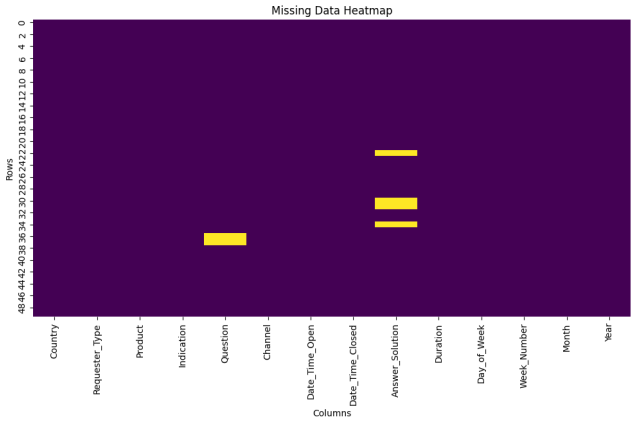

Figura 1: Missing Value Pattern

The percentage of missing values is depicted in this image:

```
Missing Data Summary:
                 Missing Count  Missing Percentage
Question                     2                 4.0
Answer_Solution              4                 8.0
```

Figura 2: Missing Value Percentages

The missing values will be removed, as the percentage (12 %) is deemed insignificant.

I-C. Clustering and Embedding

Three sources of text were analyzed:

1. **Text:** Question + Answer
2. **Questions**
3. **Answers**

For each source, embedding generation and clustering analysis were conducted. This analysis allows for the extraction of potential labels, facilitating the understanding of relationships with categorical variables. The following image displays the T-SNE plot of the clusters for the text, indicating three clusters:

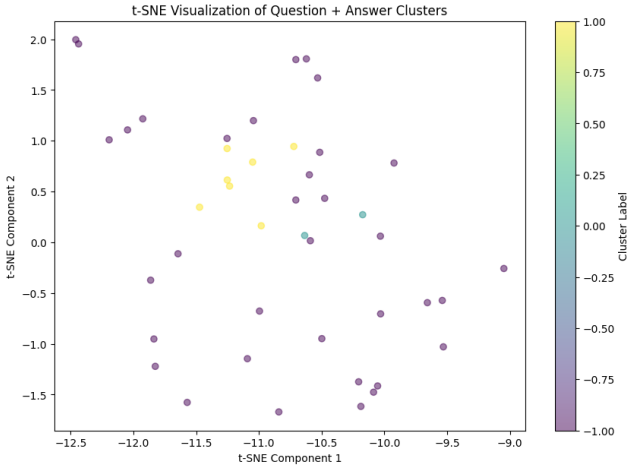

Figura 3: T-SNE Plot of Clusters

From the combination of questions, answers, and text, up to seven potential clusters can be formed, as shown in the following graph:

	question_cluster	answer_cluster	text_cluster	count
0	-1	0	-1	16
1	-1	1	-1	2
2	0	0	-1	12
3	0	1	-1	3
4	1	0	-1	2
5	0	0	0	2
6	0	0	1	7

Figura 4: Potential Clusters

I-C1. Insights from Clusters:

I-C1a. First Cluster: The first cluster is more closely related to the effects of the drug:

1. **What are the common side effects of Keytruda?** Common side effects include fatigue, nausea, and skin rash.
2. **Can Keytruda cause changes in blood pressure?** Yes, fluctuations in blood pressure can occur, and patients should monitor their blood pressure regularly.

I-C1b. Second Cluster: The second cluster is also related to effects but focuses more specifically on NSCLC:

1. **How do the results of KEYNOTE-006 influence future research directions for NSCLC treatment?** The results encourage further research into immunotherapy combinations, personalized medicine based on genetic markers, and strategies to overcome resistance to PD-1 inhibitors.
2. **Can Keytruda be used in NSCLC patients with autoimmune diseases?** Keytruda should be used with caution in patients with pre-existing autoimmune diseases due to the risk of exacerbating their condition.

The final cluster percentages are illustrated in the following image:

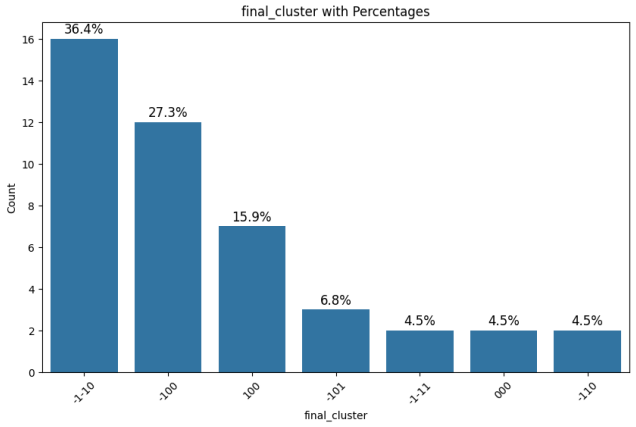

Figura 5: Final Cluster Percentage

The word clouds for the clusters of questions and answers are available in the following file: `EDA.ipynb`.

I-D. Analysis of Categorical Variables

After conducting an independence test between the clusters and categorical variables, it was determined that they are independent. This finding is logical, as the semantic meaning of the text does not relate to factors such as data or distribution channels.

I-E. Distribution of Tokens in the Text

The distribution of tokens is represented in the following image:

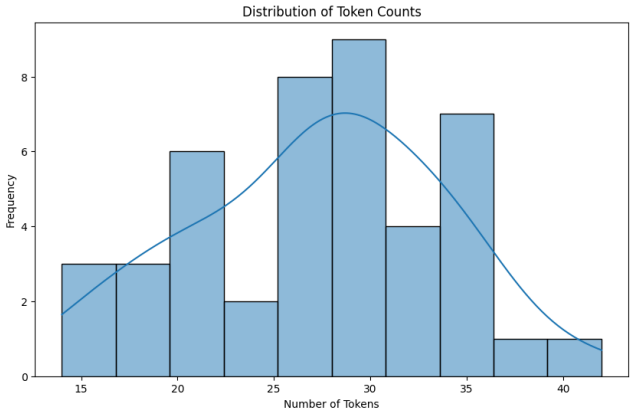

Figura 6: Token Distribution

The maximum number of tokens is around 50, indicating that the text is relatively short. Consequently, no chunking will be necessary when indexing the data.

I-F. Evaluation Set Generation

The evaluation of the RAG system will utilize a set of evaluation questions. This evaluation will be conducted using the RAGAS framework. For each question in the dataset, an augmentation process generates five new questions. The prompt used for this generation is as follows:

```
self.prompt = """
You are an AI assistant designed to help generate
     insightful questions based on a given
    medical topic,
specifically focusing on Keytruda, a medication
    used in cancer treatment.
Your task is to generate new questions that
    maintain the same core meaning, context,
and intent as the original question while
    providing unique perspectives or phrasings.

Please follow these guidelines strictly:
1. Retain the main topic and medical context of
    the original question.
2. Do not introduce new information or change the
     clinical meaning.
3. Ensure each question addresses the same
    subject matter in different words, focusing
    on Keytruda.
4. Use precise medical language consistent with
    the original question.

Here is an example based on these guidelines:

Given the question: "What are the common side
    effects of Keytruda?"

Generate 5 unique augmented questions in this
    JSON format:
```

```
{
    "question_1": "What are the most frequently
        reported side effects of Keytruda?",
    "question_2": "Which side effects are
        commonly associated with Keytruda
        treatment?",
    "question_3": "What adverse effects should
        patients be aware of when taking Keytruda
        ?",
    "question_4": "How often do patients
        experience side effects while on Keytruda
        ?",
    "question_5": "What are the typical side
        effects observed in patients using
        Keytruda?"
}

Notice how each augmented question maintains the
    focus on the side effects of Keytruda
without altering the original medical context or
    introducing unrelated topics.

Now, based on the following question:
"{question}"

Please provide {num_questions} augmented
    questions in the same format, ensuring they
    preserve the original meaning and medical
    context.
"""
```

II. Setting Up the Retrieval System

A modular retrieval system has been implemented, consisting of the following components:

1. **Search Module**
2. **Reranking Module**

II-A. Indexing

The indexing process is documented in the notebook `RAG/1_rag_indexing.ipynb`, utilizing **LlamaIndex**.

II-B. Retrieval

The notebook for configuring and experimenting with the retrieval module can be found at `RAG/2_rag_retrieving_with_reranker.ipynb`. This module employs a top-k similarity search followed by the **RankGPTRerank** module for reranking.

II-B1. Query Example: When querying the retrieval system with the question: **"What are potential CONSEQUENCES of Keytruda?"**, the following results were obtained:

1. **Without Reranking:** The retrieved nodes are displayed below:

	Score	Text
0	0.694642	Can Keytruda cause immune-related adverse effects? Yes, Keytruda can cause immune-related adverse effects such as colitis, hepatitis, and pneumonitis.
1	0.655332	Are there specific side effects of Keytruda that NSCLC patients should monitor? NSCLC patients should monitor for cough, shortness of breath, and chest pain, as these could indicate immune-related pneumonitis.
2	0.643183	What should patients report immediately while on Keytruda treatment? Patients should report any new or worsening symptoms such as cough, chest pain, or changes in vision immediately.
3	0.632885	Can Keytruda be used in NSCLC patients with autoimmune diseases? Keytruda should be used with caution in patients with pre-existing autoimmune diseases due to the risk of exacerbating the condition.
4	0.631243	Can Keytruda cause changes in blood pressure? Yes, fluctuations in blood pressure can occur, and patients should monitor their blood pressure regularly.
5	0.630975	What are the common side effects of Keytruda? Common side effects include fatigue, nausea, and skin rash.

Figura 7: Without Reranker

2. **With Reranking:** The nodes retrieved based on their index positions [0, 6, 3, 7, 1] from the original list are shown below:

	Score	Text
0	0.694664	Can Keytruda cause immune-related adverse effects? Yes, Keytruda can cause immune-related adverse effects such as colitis, hepatitis, and pneumonitis.
1	0.630779	How does Keytruda affect immune system function in NSCLC patients? Keytruda enhances the immune system's ability to detect and destroy cancer cells but can also lead to immune-related adverse effects that need to be managed.
2	0.632893	Can Keytruda be used in NSCLC patients with autoimmune diseases? Keytruda should be used with caution in patients with pre-existing autoimmune diseases due to the risk of exacerbating the condition.

Figura 8: With Reranker

II-B2. Observations: A noticeable difference is observed in the nodes retrieved between the two methods. The reranked results appear to emphasize the immune effects associated with Keytruda, indicating that the reranking process effectively enhances the relevance of the retrieved information.

II-C. Integration of the Generation Component and RAGAS Evaluation

The generation component has been implemented using OpenAI's **gpt-3.5-turbo** model. The evaluation set is generated from the notebook: `RAG/4_evaluation_set.ipynb` using a few-shot prompting approach.

II-C1. Response Modes: When integrating the generation module, the option `response_mode=çompact"` was utilized, which specifies:

1. **Refine:** Create and refine an answer by sequentially processing each retrieved text chunk, resulting in a separate LLM call for each node/retrieved chunk.
2. **Compact:** Similar to refine but concatenates (compacts) the chunks beforehand, resulting in fewer LLM calls.

II-C2. Evaluation Metrics: The evaluation metrics derived from RAGAS include:

1. **Faithfulness**
2. **Answer Relevancy**

Note: These metrics are used as they do not require ground truth data for nodes—only the query, answer, and context (retrieved nodes).

II-C3. Evaluation Results: The evaluation results for the RAG system are presented as follows:

1. **With Reranking Module:**

	faithfulness	answer_relevancy
count	219.000000	220.000000
mean	0.789695	0.891236
std	0.321104	0.259486
min	0.000000	0.000000
25%	0.550000	0.949394
50%	1.000000	0.967308
75%	1.000000	0.979370
max	1.000000	1.000000

Figura 9: RAGAS Evaluation With Reranker

2. **Without Reranking Module:**

	faithfulness	answer_relevancy
count	218.000000	220.000000
mean	0.827316	0.859463
std	0.285554	0.301945
min	0.000000	0.000000
25%	0.600000	0.949778
50%	1.000000	0.964436
75%	1.000000	0.975003
max	1.000000	1.000000

Figura 10: RAGAS Evaluation Without Reranker

III. Conclusion

The evaluation results indicate that the RAG system performs better with the reranking module, demonstrating improved relevance and faithfulness in the generated responses. Future work will focus on optimizing the retrieval process and exploring additional features to enhance overall performance.