Assignment #2: RAG-based Question-Answering System Development

1. Team and Platform Details

Team Members: Musab – 29409, Hussain – 29410.

Platform: The system was developed and executed on Google Colab due to its accessibility and support for GPU acceleration. The entire implementation, including model training and inference, was conducted in the Colab environment.

2. Data Details

Dataset Source: The corpus used for this task is the Student Handbook 2023-24 PDF document, retrieved from IBA's official website.
Dataset Size: The corpus contains approximately 57 pages and 210 number of documents, totaling about 2.11 MB in size.
Preprocessing: The PDF was processed using the pdfplumber library to extract raw text. The text was then split into chunks of approximately 500 characters.

3. Algorithms, Models, and Retrieval Methods

3.1 Algorithms and Large Language Models (LLMs)

We utilized the following models for the question-answering pipeline:

- **deepset/roberta-base-squad2**: A robust model for answering questions based on SQuAD data.
- **bert-large-uncased-whole-word-masking-finetuned-squad**: A large, fine-tuned BERT model optimized for SQuAD tasks.
- **t5-small**: A lightweight T5 model chosen for its efficiency in text generation and its ability to answer questions quickly, despite occasional issues with large inputs.

3.2 Retrieval Methods

Semantic Search (FAISS): We employed FAISS (Facebook AI Similarity Search) for semantic search, which allowed us to index text chunks and search them based on the similarity of embeddings. The **Sentence-BERT model** (paraphrase-MiniLM-L6-v2) was used to generate dense vector embeddings for each chunk.

Justification:
Semantic search was chosen over keyword-based search to improve the quality of the retrieved text. This method enables the model to understand the context of the question better and retrieve relevant chunks, even if they don't contain the exact keywords.

3.3 Summarization and Preprocessing

Summarization: We considered applying extractive summarization (using models like BART or T5) on long documents before passing them to the LLM to reduce unnecessary information and improve answer precision. However, we decided to use the entire context to retain more information for the answers.

3.4 Other Techniques for Long Context Handling

Context Chunking: We split the document into smaller text chunks of approximately 500 characters each to handle the document's large size. This ensured that the input to the model remained within token limits, while still capturing relevant information.

4. Performance

4.1 Accuracy

Since the dataset does not contain labeled answers, we manually evaluated the relevance of the retrieved answers based on a set of test questions.

4.2 Efficiency

Inference Time: The average inference time for generating an answer using the models was:

- **deepset/roberta-base-squad2**: 2.83 seconds per question
- **bert-large-uncased-whole-word-masking-finetuned-squad**: 8.42 seconds per question
- **t5-small**: 1.09 seconds per question

Retrieval Time: The time taken for FAISS search was about **0.5 seconds** for retrieving the top 3 relevant chunks.

5. Reproducibility

5.1 Preprocessing and Setup

Preprocessing: The PDF file was processed using the pdfplumber library to extract raw text. The text was then split into chunks of approximately 500 characters.
Model Configuration: We used the SentenceTransformer model paraphrase-MiniLM-L6-v2 for generating embeddings and FAISS for indexing and searching the text chunks.
Environment: The system was developed on Google Colab with a GPU runtime enabled.

5.3 Code and Configuration Files

Code Repository: https://github.com/musabjamil/textanalytics.git

Dependencies: The system relies on the following libraries:
transformers, sentence-transformers, faiss-cpu, langchain, spacy, os, time, pdfplumber

Environment: Python 3.x, Colab/Jupyter notebook environment.

Colab link:
https://colab.research.google.com/drive/194ilsmJ_1ogXRkBJwquysIUVukHYp_3Y?usp=sharing

6. Table

Table 1: Performance comparison of different models used.

Model	Q1 Time (seconds)	Q2 Time (seconds)	Q3 Time (seconds)	Q4 Time (seconds)	Q5 Time (seconds)	Total Time (seconds)
deepset/roberta-base-squad2	2.83	0.88	0.82	0.77	0.68	**6.68**
bert-large-uncased-whole-word-masking-finetuned-squad	8.42	3.35	2.28	2.2	2.5	**19.7**
t5-small	1.09	1	1.03	1.45	0.6	**5.17**

7. References

- HuggingFace documentation: https://huggingface.co/docs

- FAISS library: https://faiss.ai/index.html

- Sentence-BERT: https://sbert.net