Institute of Business Administration
Introduction to Text Analytics
Assignment 04
Due Date: 13th April 2025 (11:55 PM)

RAG-based Question-Answering System Development

This is a group assignment where the group size can be of 2-3 members. In this assignment, your task is to:

- Develop a Retrieval-Augmented Generation (RAG) based question-answering system.
- The system should retrieve information from a specific domain corpus (you may choose any domain such as medical, legal, human resources, etc.) and answer questions posed in natural language.

You must perform and report extensive experimentation, including

- different retrieval strategies,
- parameter tuning, and
- model variations.

While the core concept of RAG programming may seem straightforward, many challenges arise during data preparation, and it's a common area where students encounter difficulties. The goal is to not only build a functioning system but to clearly document your approach in a reproducible manner.

Deliverables

1. Python code notebooks that you used for experimentation.
2. A well-organized report either in the docx or pdf format.

Report Guidelines

Organize your report into clearly defined sections. Your report should include the following components to ensure that someone else can replicate your process.

1. **Platform Details**: Specify the platform used for experimentation (e.g., local machine, Kaggle, Colab). If multiple platforms were used, clarify where each stage was executed.
2. **Data Details**: Clearly state the source of the dataset, including the size and number of documents used in the corpus.
3. **Algorithms, Models, and Retrieval Methods**: Clearly document the experimental setup and results, highlighting insights gained from multiple trials.
    - Describe the retrieval methods employed in your system. Did you use semantic search, keyword-based search, or another method? Justify your approach.
    - Specify the algorithms and large language models (LLMs) you used, and explain your choices.
    - Explain your chunking strategy, including how you segmented the documents and whether different chunking approaches were tested. Discuss how chunk size and overlap affected retrieval and answer quality.
    - If applicable, explain whether you applied summarization techniques before passing retrieved results to the LLM.
    - Detail any other techniques used to improve the quality of the input to the LLM, such as long-context reordering or similar schemes. Discuss how these methods impacted the system's performance.
4. **Performance Metrics**: Compare results across different models, retrieval strategies, and parameter settings, providing insights into how various choices impact performance.

- Implement and report evaluation metrics for generated answers, specifically faithfulness and relevance. If automated libraries like Ragas do not work reliably, design your own prompting method to evaluate these metrics.
- These evaluation metrics should be analyzed across different retrieval strategies, LLM choices, chunking methods, and post-retrieval processing techniques (e.g., summarization, long-context reordering). Your report should compare these variations and discuss their impact on answer quality.
- Moreover, analyze the latency and computational efficiency by measuring inference time, retrieval time, and overall system response time.

5. **Best Model Selection:** Justify your best model selection by assessing the effectiveness of using the above-mentioned performance metrics.
6. **Reproducibility**: Your report must provide enough detail to enable others to replicate your work. Include any information that is critical for reproduction, such as preprocessing steps, system configuration, or model fine-tuning techniques.

Additional Instructions:

- **Application Development (Optional)**: If you have developed a working application based on your system, include screenshots and relevant details in an appendix at the end of your report.
- **Figures and Tables**: Ensure that all figures and tables are properly numbered and cited in the text. Avoid vague references like "the figure below"; instead, use precise citations such as "Table 1 shows…" or "As shown in Figure 7…".
- **References**: If you have used external resources, such as blogs or GitHub repositories, ensure they are appropriately cited. Include a reference section before the appendix to acknowledge all sources and avoid any potential issues of plagiarism. Proper citation is a key part of your academic and professional training.
- **Submission File Name**: The file name should be as per the group members name and don't name it Assignment1 or Project1. So if there are two group members, Aamna and Zaid, then name it Aamna_Zaid.docx.

NOTE: Do not submit Google Drive or shared document links, as they can be modified after submission. The report must be submitted as a **Word or PDF** file on the **LMS**. If other files exceed the size limit, they may be uploaded via Dropbox. However, any submission via shared links will be considered incomplete and will not be graded.