Assignment Overview

Design and implement a multi-modal Retrieval-Augmented Generation (RAG) system for describing images, with multilingual support. The system should integrate textual and visual data to retrieve or generate relevant descriptions for a set of images in multiple languages.

Project Objectives

1. Develop a multi-modal RAG model that retrieves accurate textual descriptions from image queries.
2. Implement an end-to-end pipeline for data processing, embedding creation, and retrieval.
3. Create a system that matches images with the most relevant textual descriptions in multiple languages.

Dataset

You will work with the **Clip Images Data**, containing images and their descriptions.

Dataset Link: Clip Images Data on Kaggle

Dataset Details

- Image files: 10 images (JPEG/PNG format).
- Descriptions: Textual descriptions corresponding to each image.

Core Components

1. Data Preprocessing (25%)

- **Images**: Resize and normalise the images.
- **Text**: Clean and preprocess the textual descriptions in the base language (English).

2. Embedding Creation (20%)

- **Image Embeddings**: Use a pre-trained model like CLIP to extract embeddings.
- **Text Embeddings**: Generate textual embeddings for the descriptions in multiple languages.

3. RAG Model Implementation (25%)

- Design a system that allows both **image-to-text** and **text-to-image** retrieval using embeddings.
- Implement efficient search techniques for fast retrieval.

4. Multilingual Support (15%)

- Integrate multilingual capabilities for the descriptions.
- Translate text descriptions to languages like **Spanish**, **French**, or any other language of choice using models like **MarianMT** or **M2M100**.

5. Evaluation and Fine-Tuning (15%)

- Use metrics like **BLEU** or **Cosine Similarity** to evaluate the system.
- Propose fine-tuning of retrieval results to improve accuracy across multiple languages (*optional*)

Evaluation Criteria

1. **Model Performance (35%)**: Accuracy in retrieving or generating multilingual descriptions.
2. **System Design (25%)**: Complexity and efficiency of the multi-modal and multilingual system.
3. **Multilingual Support (20%)**: Robustness of multilingual retrieval and generation.
4. **Code Quality (10%)**: Clarity, organisation, and documentation of code.
5. **Innovation (10%)**: Novelty in approach or implementation.

Submission Guidelines

- **Source Code**: Well-documented code for the RAG system.
- **Technical Report** (Max 3 pages):
  - Detailed explanation of approach.
  - Evaluation of the system’s performance.
  - Approach to multilingual support.
- **Sample Outputs**: Provide examples of image-description retrieval results in multiple languages.

Technical Recommendations *(better alternatives are welcome)*

- Use **PyTorch** or **TensorFlow** for model building.
- Use **CLIP** or Hugging Face's **Transformers** library for embedding extraction.
- Use multilingual models like **MarianMT** or **M2M100** for translation.
- Consider using **FAISS** for efficient similarity search.

Important Notes

- Focus on creating a working end-to-end system rather than optimising every component.
- Be prepared to discuss the trade-offs in your approach, especially for multilingual support.
- Consider aspects like computational efficiency and scalability in your design.