CS4221 Mid-term Project: Data Curation and RAGs

Instructions

- **Collaboration policy:** The project should be done in 2-3 people groups.
- **Submission:** Submit your solutions online on the course platform. Follow the submission instructions in the appendix.
- **Generative AI Policy:** Use of generative AI is allowed. However, you should document your usage, including the prompts used. Verify the correctness of all outputs.
- **Late Policy:** Late submissions are not allowed.
- **Presentation:** You will present your solution during the relevant tutorial session.

1 Database and Dataset

In this project, you are supposed to use RAGFlow to help you build your own RAG workflow. Here is a quick tutorial. API reference can be found here. Download the dataset AAPL.zip in which **news.json** lists the metadata of each news article; the **news** folder contains the raw HTML files.

2 Task 1 [40 points]

Build a QA bot based on RAG. The raw HTMLs are noisy and need to be cleaned. You can use BeautifulSoup, Unstructured, or any package you like to help extract useful text from raw HTMLs and generate cleaned TXTs. Next, use the default chunking method to construct a knowledge base. Pick an embedding model that fits your hardware. You are required to create two knowledge bases: one using raw HTMLs and the other using parsed TXTs. For generation, pick an LLM model that fits your hardware. Example queries can be found in query.txt. Record the results and comparison of RAGs with and without parsing.

3 Task 2 [60 points]

Tuning to improve the query efficiency and accuracy in Task 1. The tuning options include:

- Try and compare different chunking methods from RAGFlow, or implement your own. One option for the custom parser and chunker could be to deploy a local LLM. You may also try different solutions mentioned in the lecture.
- Try and compare advanced RAG features such as re-ranking and multi-way retrieval.

4 Report Instruction

In your report, describe your solution and how to run it. For each group, submit a PDF report with code.