Problem Statement by Overlayy
Gen AI PS

Objective:

Your task is to develop a system that automates the process of generating relevant questions from website content. The system should follow these steps:

1. **Website Scraping:**
    - Given a website URL, scrape or collect all the links in a json/csv file present on that website.(some url crawler api)
    - For each link, retrieve and save the response(content can be generated through open source tools like jina ai, but look for edge cases which cover all information) (content of the webpage) in a JSON file.
2. **Question Generation:**
    - Use the saved responses to generate 10 questions related to the content of each webpage (each url of the website).
    - Each question should be concise and contain fewer than 80 characters.
    - You may use any language model (LLM) or API to assist in generating these questions.
3. **Relevant Links and Topics:**
    - For each webpage, identify and select 5 relevant links from the scraped URLs that are pertinent to the content and questions generated.
      Sample format
    - Save the generated questions, relevant links, and topics in a structured JSON file.
4. **Automation and Validation:**
    - Implement automation to verify that each webpage has exactly 10 questions, each under 80 characters, and that each entry includes 5 relevant links and topics.
    - Develop a metric to evaluate the performance of the question generation and relevance detection process. This can be done using an additional LLM or custom evaluation method.

Evaluation Criteria:

- Accuracy and relevance of generated questions.
- Correctness and appropriateness of the selected links.
- Performance and reliability of the automation and validation process.

Tools and Technologies:

- Web scraping tools or libraries (e.g., BeautifulSoup, Scrapy) or any online crawler.
- Language models or APIs for question generation (e.g., OpenAI GPT,gemini,etc).

- JSON or CSV handling libraries for data storage.
- Automation frameworks for validating outputs.

Sample output file: https://pastebin.com/bN0LdNXc