Task Assignment Document: Synthetic Deep Research Question Generation

Update

Our current focus has shifted from the energy vertical domain to building a general-domain synthetic deep research question generation workflow.

Overall Goal / Direction

Focus on synthetic deep research question generation with reference. Start with general domain, and then move on to special domains (energy to start with).

Version 0 (V0)

doc → KNN docs → LLM summarization of reports → LLM generated deep research question

Use prompts to make the reports and questions better

Version 1 (V1)

Define criteria to make the questions "deep research" questions, not trivial ones

This Week's Goal We need to complete the total coding of document collection and implement the basic report and question generation process. All work must be completed on the cluster.

Step 1: Document Selection

First, log in to the boston cluster using 'ssh username@boston-cluster.lti.cs.cmu.edu'

Copy the file /bos/tmp1/ClueWeb22\_B/ClueWeb22-ID\_URL-Hash\_map.csv.gz to your own folder in the cluster and unzip it. This gives a list of all document IDs in ClueWeb22-B.

Randomly sample 10 document IDs from this list.

Use the cw22.get\_doc function provided in ClueWeb22.py from https://github.com/cxcscmu/ClueWeb-operator to retrieve full document content.

Once documents are retrieved, proceed to the next step.

Step 2: Call DeepResearchGym API

Treat each retrieved document as a separate domain. For each document, generate a corresponding report and then generate deep research questions based on that report.

Call the API using each retrieved document as the query and set k=100. Download the top 100 similar documents for each domain as its initial corpora. Refer to the official website https://www.deepresearchgym.ai/ for API documentation and usage examples.

Step 3: Generate Report and Research Questions with LLM

Use each domain's 100-document corpora as input.

Call the LLM API with appropriate prompts to summarize the domain corpus into a coherent expert-style report.

Then, use a second stage of prompts to generate deep research questions based on the report.