Group Member* & Contact

Jinghan HUANG	jinghanhuang@g.harvard.edu
Zhiyu LI	zhiyuli@fas.harvard.edu
Haozhuo YANG	haozhuoyang@gsd.harvard.edu
Luozhong ZHOU	lzhou1@g.harvard.edu

*Members are listed in alphabetical order by last name.

Who Used the Dataset for Exact What?

— Developing an AI-Powered Research-centered Dataset Search System

Background

How do researchers efficiently discover relevant literature? Traditional academic search platforms are often limited by keyword-based categories, which can restrict the depth and precision of searches. While datasets are a critical component of many research papers, they are not consistently or comprehensively categorized in current search portals. As a result, researchers may spend hours locating articles that reference specific datasets or relationships between data, only to struggle in determining whether the content is relevant to their work.

The rapid advancement of large language models (LLMs) presents a new opportunity for enhancing academic searches. By offering fine-grained categorization of datasets and their relationships, researchers could gain more control over search filters and results. We propose developing a platform which allows for more customized navigation, empowering users to cluster and refine search results based on specific dataset criteria, ultimately streamlining the research discovery process.

Problem Statement

Current popular academic search portals fall short in helping users efficiently locate papers based on users' specific research needs. This project seeks to develop an alternative search mechanism for academic papers, specifically for researchers whose work depends heavily on datasets.

Several key areas for improvement are there in current search portals:

- First, users are often limited to searching by dataset-related keywords. However, these keywords may correspond to variables that play vastly different roles within a paper. In social science research, for example, a keyword might represent a primary variable in one paper, driving the core analysis, while in another it might be a secondary or contextual variable, offering less relevance to the user's research goals. Current search platforms do not provide sufficient clarity on how datasets are used within a paper, making it difficult to discern the relevance of search results.
- Second, when users search for multiple datasets, they have no way of filtering results based on how those datasets are related within the research. As a result, papers that mention both datasets, regardless of their actual interaction or importance, are included in the search results. This makes it difficult to isolate papers where the datasets have meaningful relationships critical to the user's work.

To address these limitations, we aim to create a dataset-centered search portal. This portal will enable researchers to use fine-grained dataset categories and filter papers not just based on the presence of datasets, but also on the nature of their relationships. By offering more precise search parameters, researchers can more efficiently navigate the literature, focusing on papers that are directly relevant to their specific dataset-driven research needs.

Data Sources

We have two categories of data: academic journal data and data of datasets. For the former, we would first construct an index of journals and obtain the information of all articles on these journals within a timespan, say 10 years or 20 years, and utilize the DOIs as handles to further download full articles. For the latter, we can call existing APIs, if any, provided by data repositories, (e.g., ICPSR). The academic journal dataset would contain two parts: metadata of research articles and the full text of research articles. The dataset of datasets would contain metadata of popular datasets in social science, including the publishing year, theme, purpose, principal investigator, etc. Metadata of both research articles and datasets are structural, while the dataset of journal articles is semi-structured and textual. All the data we are to obtain are typically well maintained by publishers or third-party service providers. Hence, the data quality is unlikely to be a concern in our project.

Tech Stack

Data Collection (Web Scraping)

- **Tools:** Scrapy, BeautifulSoup, PyPDF, etc.
- **Procedure:**
    1. Manually curate a list of target journals;
    2. Use Scrapy to automatically download meta information of articles from these journals within a given time period;
    3. Download the full text of articles with DOIs.

AI/ML

- **Tools:** PyTorch, Tensorflow, Transformer, etc.
- **Procedure 1 (Info Extraction):**
    1. Create a training set manually with the assistance of 3rd party LLMs;
    2. Fine-tune an LLM (e.g., LLAMA) to make it fit our task the best;
    3. Deploy our LLM to process journal articles at scale.
- **Procedure 2 (Info Match):**
    1. Curate a training set consisting of similar semantic units;
    2. Fine-tune a Bi-Encoder and generate off-shelf embeddings for future fuzzy matches.

Database

- **Tools:** SQL (MySQL, PostgreSQL)

Front-End

- **Tools:** HTML, CSS, JavaScript

DevOps

- **Tools:** Google Cloud, Docker, GitHub

Milestones

Data collection and preprocessing ~ 1 week;

Model fine-tuning ~ 2 weeks;

Data processing ~ 1 week;

Front-end and Deployment ~ 3 weeks.