Project Title

AI-Powered Discovery System for Transforming Human-Centric Structured Data into Actionable Knowledge

Project Description:

Tabular data and charts, which are commonly embedded in pdf files, excel sheets, web pages, and other document types stand out through their eye-catching layouts as objects storing and representing critical information, e.g., extracted insights, statistics, summaries, and aggregate data. These tables and charts typically have sophisticated structures to deliver a wealth of information in a concise and dense way primarily targeted for human understanding and consumption. We refer to these as "***human-centric structured data***" (**HCSD**) (refer to Figure 1). With many organizations generating thousands of such documents, each containing hundreds of HCSD objects, the traditional manual processing and visual inspection over individual HCSD objects by an analyst is no longer a viable option. Therefore, the pressing challenge we address in this proposal is to devise scalable and advanced techniques for ***data discovery*** (e.g., discovering relationships, novel transformations, and relatedness to a given query), ***querying*** (e.g., SQL-like queries expressed in natural language), and ***advanced analytics*** (e.g., leveraging visual signals in HCSD understanding and transformation, and lineage tracking) over the rich and ever-increasing HCSD objects. Such techniques are especially critical because for most organizations, HCSD represent the real information catalyst for making knowledge-based decisions.

Team's Targets. Our team will investigate, research, and develop solutions to overcome the following challenges:

(1) Unlike relational tables, the sophisticated and non-standard structure of HCSD makes it very hard for systems to understand and process.
(2) The highly mature AI- and LLM-based technologies for processing unstructured text are far behind when applied over structured data in general, and HCSD in specific, due to their complexities. And thus, no out-of-the-box solutions using these technologies are readily available.
(3) HCSD (and their container files) are usually generated from different sources, using different ETL processes, and by different people, over possibly long periods of time. As a result, performing integration, discoveries, and queries across HCSD objects is very cumbersome.
(4) Users' queries over HCSD are typically expressed in natural language (NL) expressions. As a result, ambiguous semantics is inherent in both queries and data, which raises the problem complexity.

Our proposed solution relies on three pillars, namely ***Data-Centric HCSD Transformations***, ***Fused Models for HCSD Data Discovery and Querying***, and **HCSD *Advanced Analytics***, all powered by the state-of-the-art innovations in AI and data management. Unlike existing techniques, the creativity of the project lies in its holistic approach for managing the complex processing lifecycle of HCSD.

Project Type: Research or Engineering

Internship Batch: May 12 to July 11, 2024

Duties/Activities:

1- Help building and collecting datasets related to the project
2- Building an interface for annotating the collecting datasets
3- Building a web crawler for specific domains

Required Skills: Programming skills in Java, Python, and Panda, database knowledge, LLM knowledge

Preferred Intern Academic Level: An undergraduate student in Computer Science or similar degrees (IT or Data Science)

Learning Opportunities:

1- Learning research skills
2- Learning state-of-the-art techniques related to the project scope
3- Polishing her programming skills
4- Hands-on experience in practical research projects
5- Learning how to work with and handle large-scale data
6- Leaning teamwork skills

Expected Team Size: *He/She will join a team of 5*

Mentors

Name: Mohamed Eltabakh email: meltabakh@hbku.edu.qa

Mourad Ouzzani email: mouzzani@hbku.edu.qa