Progress Report - Enhanced ExpertSearch System

CS 410 - Text Information System Course Project

This is the project progress report for CS410 Fall 2023 semester. The team chose the **System Extension** topic to improve the existing **ExpertSearch System,** which aims to find faculty based on their specializations and research topics.

Enhancements

Features:

- Improve ranking algorithms - We will test using the five main ranking algorithms that were utilized in MP2.1 (BM25, pivoted length normalization, absolute discount smoothing, Jelinek-mercer smoothing, Dirichlet prior smoothing), all with multiple different tuning parameters, and choose the one with the highest accuracy as the primary ranker in the project.
- Automatic Faculty Page Crawling - We will add the feature to identify the faculty directory pages and crawl the data for data extraction and query.
- Data extraction – We plan to extract additional information from the faculty biography page. This requires text parsing and processing since the information in these pages are likely unstructured data. Once we extract more data points and transform the data into structured information, we can add more filtering and/or enhance our search query ability to capture more than just faculty research topics. Data extraction will require enhanced regex-based and NER-based methods, along with topic mining and keyword extraction.
- Live chat – Provide users with the ability to look up relevant faculties based on a live chat system where the user can provide a query relating to their search parameters. The chatbot will output all relevant faculty according to it.

Progress

- AI Chatbot - We tested many pre-trained Python libraries. We decided to use the Google-powered chatbotAI as our primary pre-trained language model since this package can adapt to a custom knowledge-based database, which suits our use cases. The next step is to integrate the chatbot into the existing User Interface of the ExpertSearch system and utilize the enhanced data extraction features as our knowledge-based database.

- WebCrawling - Built base platform for web crawler, and when given the starting point as the UIUC website, it is able to crawl to the directory pages and each faculty's page and scrape the information of the respective faculty from their page. We are using Python libraries for the web crawler, We are currently passing string checks to identify the path to the individual faculty member's pages and now working on utilizing regex implementation instead.

Remaining Tasks

- Integrate chatbotAI with the existing UI and the backend. Writing rules for how the chatbot will respond.
- Connect the WebCrawler with the existing ExpertSearch platform and store the data.
- Extract data from the webpage and store it in the data model for the search and chatbot.
- Testing, debugging, and documentation.

Challenges

- Determining the most effective ranking algorithm that suits the varied nature of faculty information.
- Determining the best way to analyze and store large amounts of data.
- Integrating different components of the project due to varying object models.
- The starter code does not function properly, therefore it needs debugging.