# Enhanced ExpertSearch System

CS 410 - Text Information System Course Project Team Hybrid

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

**Team members**

| Name | NetIDs |
|---|---|
| Rasinee Pongchairerks *(Captain)* | rasinee2 |
| Aksh Gupta | ag26 |
| Pakhi Gupta | pakhig2 |
| Woo Jeon | woojeon2 |
| Xingsi Zhang | xingsiz2 |

**Datasets and Algorithms**

ExportSearch system is powered by the faculty dataset which is crawled from the faculty directory page from the university URLs. The university URLs are the starting point for the web crawler to gather faculty information and store it. The faculty dataset is unstructured data as it requires data parsing and transforming into more structured data.

We will utilize multiple algorithms and techniques for this project. ExpertSearch System can be split into different stages to get the indented information based on a user's query:

1. **Web Crawling**: First stage is web crawling through all the university pages and using document selection with Boolean queries to classify whether an individual page is a faculty directory page.
2. **Faculty Webpage Identification**: After gathering all the faculty director pages, we can use document selection with Boolean queries again to classify if a link is a faculty webpage or not.
3. **Data Extraction**: The next stage is to use algorithms such as regular expression for text parsing, Name Entity Recognition (NER) and topic mining for text extraction.
4. **Chatbot Implementation**: The last stage will include building the chatbot using pre-existing python packages for AI type chatbots and implementing functionality to input user queries to make output judgements using the data extracted from the second stage.

## **Enhancements**

Features:

- Improve ranking algorithms - We will test using the five main ranking algorithms that were utilized in 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 are planning 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 and the chatbot will output all relevant faculty according to it.

Future Improvements:

- Save data to database - The structured data could be stored in the database for systematic retrieval and efficiency of the system.

## **Demonstration**

Our enhancement focuses on the improved data extraction methods which will be then employed in the chatbot and added functionality in the UI (e.g., email professor / open social media), thus our implementation can be visually demonstrated through our webpage. We

In the backend, we will show how we have taken all the important data from the Illinois website and transformed it into structured data and have all key information needed for text retrieval based on user query. The data structure can be presented to illustrate how the data is extracted and structured.

## **Communication and Utilization of System**

Our code will be built upon the current implementation of the ExpertSearch System. Since our project is primarily focused on enhancing the system, we will use the existing system code as our base and starting point. Our enhancements will be integrated with the existing code base. We will also deliver any additional bug fixes and code optimizations required for the current code base to make it function accurately and efficiently.

## Programming Languages

For this project, we are planning to use Python as our main programming language. There are many available packages we can use to achieve our goal such as MetaPy, requests, and packages like FastAPI to host our server for our chatbot.

## Workload and Task Distribution

There are four main features: automatic faculty pages crawling, data extraction, ranking improvement, and live chatbot. We are planning to spend our time on tasks below:

- Web Crawler Development - 20 hours
- Faculty Webpage Identification - 20 hours
- Data Extraction - 20 hours
- Chatbot Development - 20 hours
- Testing and Debugging - 10 hours
- Documentation and Presentation - 10 hours

Total: 100 hours (20 Hours * 5 Members)