COMP 4560 - Industrial Project

University of Manitoba

Course Instructors: Robert Guderian & Aden Knelsen Dobson

Project Proposal

Project Title:

Enhancing Alumni Deceased Data Matching with Machine Learning for Donor Relations

2024

Prepared by :
Francis O. & Rishamdeep S.

ochiaghi@myumanitoba.ca
singhr50@myumanitoba.ca

ABSTRACT

Our project seeks to modernize and automate the process of identifying deceased alumni and updating their next-of-kin information. Currently, this involves paying a third party to web scrape online obituaries, generating approximate matches with alumni records through a weighted scoring system, and relying on human reviewers to confirm matches and extract next-of-kin details. This method faces challenges such as scraper fragility due to website updates and significant reliance on manual intervention for data extraction and validation.

To address these issues, we propose integrating a Large Language Model (LLM) to enhance the web scraping and data processing workflow. The LLM would dynamically adapt to website structure changes, extract obituary content more reliably, and automate next-of-kin identification directly from the text. Additionally, the LLM would perform fuzzy matching between scraped data and alumni records, replacing the current weighted sum approach.

BACKGROUND

Francis: I have relevant experience for this project from the courses I have taken. In COMP 3350 (Software Engineering 1) and COMP 4350 (Software Engineering 2), I learned about software development processes, project management, and designing and implementing complex systems. These courses taught me how to analyze problems, collaborate on solutions, and deliver effective results. Additionally, in COMP 3190 (Introduction to AI), I learned about artificial intelligence concepts, including machine learning techniques, which are directly relevant to this project. These experiences have provided me with the skills needed to take on this work.

Rishamdeep: I have relevant experience for this project from the courses I have taken. In COMP 3350 (Software Engineering 1), I gained experience in designing and implementing software systems, working collaboratively on projects, and applying structured development processes. Before the kickoff, I plan to deepen my understanding of deploying LLMs locally to ensure secure handling of sensitive data. I will focus on using AI for web scraping, exploring AI-driven tools like ScrapGraphAI, which offer dynamic, intelligent scraping capabilities, and Ollama, which provides a secure, local deployment solution for LLMs

PROBLEM STATEMENT

In University of Manitoba donor relations, there is a need to keep track of the key life events of an alumni in order to stop solicitations to the deceased and possibly send condolences, as well as options to create legacy gifts in their memory. Donor relations department of Advancement Services already have an automation that web scrapes online obituaries to extract data of the deceased, clean the data and then compare it to the alumni database using rules-based deterministic techniques. The issue is the web scraped data often does not match 1-to-1 with the data in the alumni database, especially when it comes to relationship data, such as, children, brothers, sisters, etc.

Goal: The goal is to now come up with a machine learning solution to be better able to extract data from web sources. This could involve using an open AI model over the current rules-based deterministic techniques in order to fully extract any relevant relationship information, as the current automation can already handle extracting the alumni personal information such as first and last name, also possibly the university studied at. Another goal is also to increase the accuracy of the matches between the “fuzzy” record matching against the alumni database to a level that is close to or even exceeds human accuracy, through making use of machine learning based solutions.

Industry Relevance: This project shows how machine learning-based solutions (including AI) are becoming more important in the nonprofit and education sectors for handling sensitive data carefully and effectively. Using these solutions can help donor relations address challenges like matching data more accurately, ensuring respectful and clear communication. Better tools for managing information about deceased alumni will strengthen donor relationships, improve efficiency, and demonstrate empathy.

Methodology and Timeline

The project will follow the Scrum methodology, which emphasizes iterative progress, collaboration, and regular feedback to ensure that the project is completed within the term. The work will be divided into sprints, each lasting about two weeks, and focused on specific aspects of the project.

Sprint	Sub-Tasks	Timeline	Check-In Presentation
1	Conduct initial research on obituary scraping and fuzzy data matching techniques.	Weeks 1-2	Research findings and initial approaches to obituary scraping and data matching.
	Familiarize the team with relevant tools and frameworks, including LLMs and machine learning libraries.		

2	Develop the Web Scraping Module using LLM-based techniques for obituary content extraction.	Weeks 3–4	Initial prototype of the Web Scraping Module capable of extracting obituary data dynamically.
	Test module on a fixed set of websites to ensure accuracy and reliability.		
3	Build the Data Matching System using machine learning to perform fuzzy matching against alumni records.	Weeks 5–7	Prototype of Data matching system with improved accuracy.
	Conduct tests to validate matching performance and refine algorithms		
4	Integrate the Web Scraping Module and Data Matching System into a unified proof-of-concept system.	Weeks 7–8	Functional proof-of-concept system combining scraping and matching capabilities.
	Perform end-to-end testing to identify and resolve integration issues.		
5	Prepare detailed technical documentation covering system design,	Week 9	Comprehensive technical documentation.

	implementation, and usage instructions		
6	Present the final proof-of-concept system to industry contact.	Weeks 10-11	Final proof-of-concept system delivered and reviewed by project supervisors and industry contact.

Infrastructure and Facilities

1. Computational power: When deploying LLMs locally, the computational requirements vary based on the model's size and complexity. Smaller models can be managed with moderate computing resources, including standard GPUs, less memory, and limited storage capacity. In contrast, more advanced and sophisticated models demand significantly greater computing power, including high-performance GPUs, increased memory, and expanded storage. If the required computational power is not available, we will adapt by focusing on smaller, more efficient models that can run on moderate hardware setups. These smaller models, while less resource-intensive, can still provide effective performance for our use case when fine-tuned appropriately. By optimizing these models for the specific tasks of data extraction and fuzzy matching, we can achieve the desired outcomes without the need for extensive computational infrastructure.
2. Alumni Database: Access to the alumni database is essential for validating and testing the system. If direct access is unavailable, we would be willing to work with securely anonymized, sample, or even dummy data sets that align with privacy and compliance standards. Using dummy data for development and testing would allow us to simulate real-world scenarios while ensuring data security and ethical considerations are met. This flexibility will ensure that the necessary inputs are available for building and validating the system, enabling us to deliver the project on time.
3. Additional Collaborations: To collaborate with graduate students or an industrial partner who are not currently in the Department of Computer Science, I would begin by identifying the relevant individuals through my supervisor or departmental resources and reaching out to them via email or professional platforms like LinkedIn. In my communication, I would provide a clear overview of the project and explain how their input or guidance could contribute to its success. Virtual meetings using tools like Zoom or Microsoft Teams would be proposed to discuss project details and seek their advice.

Outcome and Deliverables

At the completion of this project, we expect to deliver a functional proof-of-concept system that integrates a locally deployed Large Language Model (LLM) to automate the extraction and matching of obituary data with alumni records. Specifically, the deliverables will include:

1. **Web Scraping Module:** An LLM-powered system capable of dynamically extracting obituary content, including next-of-kin information, even when website structures change.
2. **Data Matching System:** A machine learning-based fuzzy matching solution to improve the accuracy of linking scraped data with alumni records.
3. **Documentation:** Detailed technical documentation covering system design, implementation, and usage instructions.

Stretch Goal: Integrate the proof-of-concept system with Raiser’s Edge to streamline the workflow further. This integration would enable the automated updating of alumni records directly within Raiser’s Edge, ensuring that deceased records and next-of-kin information are accurately reflected without requiring manual intervention.

Shrink Goal: The shrink goal would focus on delivering a simplified AI-powered proof-of-concept using LLM for extracting next-of-kin information from a fixed set of obituary websites and performing matching using AI, with outputs provided in a structured report format for manual review and updates.