Automating Textbook Identification From Course Syllabi and Availability Tracking For Libraries

Parsa Mallik
pmallik22@earlham.edu | Computer Science Senior Project

Introduction

This project aims to automate the identification and tracking of required and recommended readings from course syllabi at Earlham College to ensure textbooks are accessible in the Lilly Library. Currently, the process relies on professors notifying the library, which can lead to inconsistencies. This tool uses a large language model (LLM) to accurately extract textbook information, even from varied formats, and stores this data in a database. Through web scraping, the tool checks book availability in the library catalog. This system will help reduce students' financial burden and streamline library resource management, with potential applications for broader academic use in other institutions.

Design

This project utilizes an LLM to perform zero-shot named entity recognition (NER), web scraping, and a structured database to streamline the tracking of course readings.

1. **Zero-shot NER** LLMs are a type of AI capable of processing and generating human-like text. They are trained on vast datasets and can be adapted to specific tasks. In this case, the tool leverages zero-shot NER on a pre-trained LLM. The model is prompted to directly identify and extract textbook references from syllabi text files across varied formats without the need for additional training.
2. **Web scraping** The program then checks the availability of these readings in the library's catalog through web scraping, using advanced parsing tools to identify matches based on title.
3. **Database interface** The extracted readings are stored in a database, and missing details are filled using an external API. This structured database is used to generate a human-readable report, showing textbook availability at the library.

Methods

This data architecture diagram outlines a pipeline for processing course syllabi to automatically extract and track required and recommended readings. The workflow begins with PDF syllabi files, from which prose is extracted and fed into a pre-trained model that performs NER to list the reading information. The list then formats each reading and stores it in a central database. To verify availability, a web scraping module queries the library catalog, updating the metadata with availability information. The final output is a report detailing course readings and their availability status, facilitating resource tracking for the library.

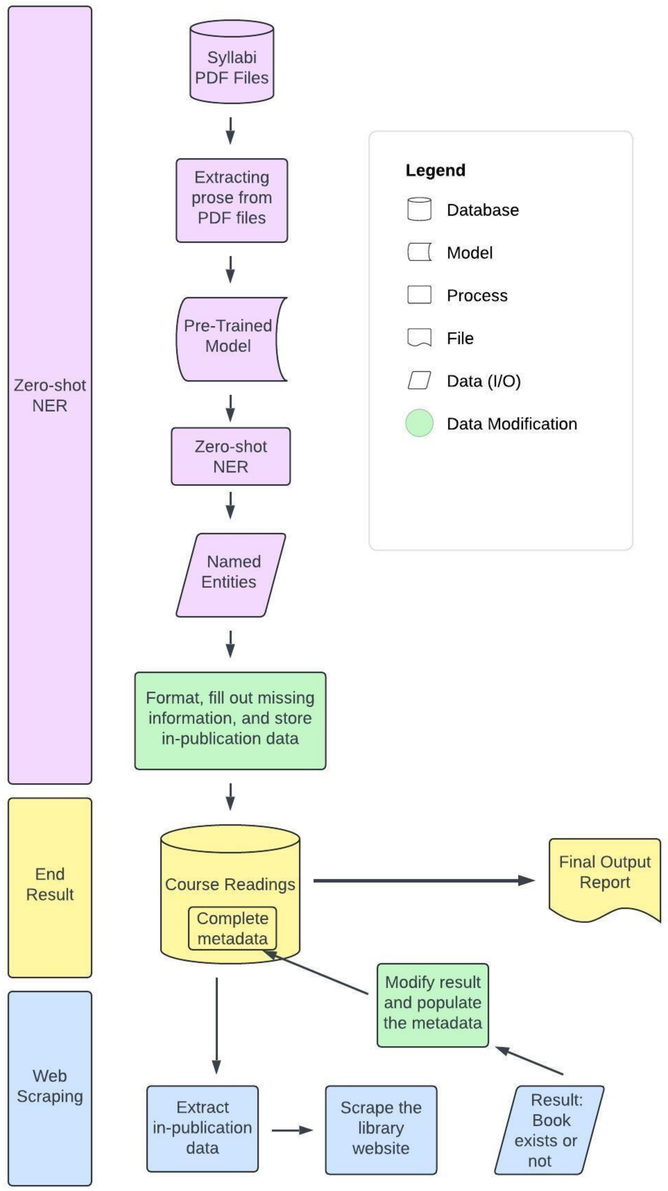

Figure 1: Data architecture Diagram

Result

The efficacy of the tool was verified by running the program with one syllabus and manually checking the textbook requirements in the syllabus and their availability on the library website. The result produced by the tool was compared to the manually checked one.

- The tool identified 50% of the books listed in the syllabus and produced 31% false positive results.
- The web scraping part correctly labeled books as available or not available.

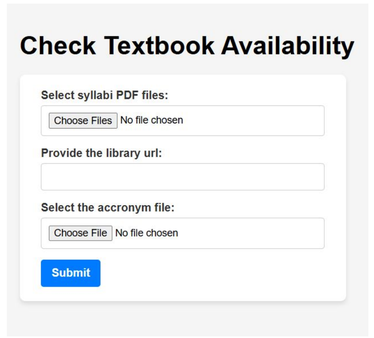

Figure 2: UI home page.

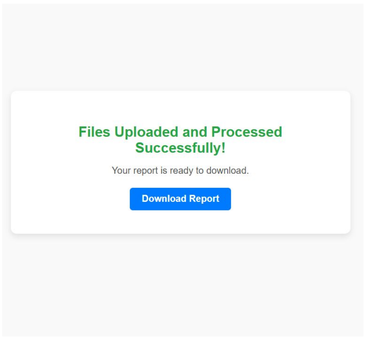

Figure 3: UI result page.

	A	B	C	D	E
1	id	isbn	book_title	availability	
2	1	9780313378980	contemporary sociological studies	no	
3	2	9780812976717	The Satanic Verses	no	
4	3	9780807031452	The Ebonics debate	no	
5	4	9781582976082	academic monograph	no	
6	5	9783161484100	Chicago	no	
7	6	9780415908085	Teaching to Transgress	no	
8	7	9780367222796	Routledge history of queer America	no	
9	8	9780813317298	rism and AIDS Prevention/Education	no	
10	9	9780415922135	tional, Ethnic and Religious Identities	no	
11	10	9780801846328	Genealogies of Religion	no	
12	11	9780684856575	Black reconstruction	no	
13	12	9781517912253	Virtue hoarders	no	
14	13	9780199747252	ant Ethic and the Spirit of Capitalism	yes	

Figure 4: Excel report.

Future Work

- Improve the syllabi data extraction part to reduce false positive results.
- The tool can be improved to correctly identify and fetch missing in-publication data.
- The web scraping part can be improved to identify the type of the books available on the website: eBook or physical copy.
- The web scraping part of the tool strictly works for the library at Earlham College. The tool can be adapted to work for any college and their library website.