DCPAPERS

Leverage Natural Language Processing (NLP) to improve the discoverability of academic resources

Charlene Chou[†1] Shravan Khunti[2] Harshit Bhargava[3]

New York University, Division of Libraries, New York, NY, United States†1
New York University, Center for Data Science, New York, NY, United States 2,3

Abstract
This interdisciplinary project is a collaboration among library metadata librarians, data scientists, digital library technologists, university IT, and the university press. Its goal is to improve the discoverability of academic resources by enhancing metadata through Natural Language Processing (NLP) and embedding-based semantic search, addressing the limitations of traditional keyword-based retrieval. To support this pilot, a library NLP system architecture has been designed, including the development of a vector database to enable semantic search within discovery platforms.

Keywords
metadata management, NLP (Natural Language Processing), generative AI, semantic enrichment, controlled vocabularies, LLM (large language model), information retrieval

Introduction

The authors express a major concern that, despite the abundance of rich metadata, users are often disappointed with search results. This issue highlights inherent limitations, e.g. lack of semantic understanding, in current discovery platforms, such as PrimoVE (library services platform), eBook collection websites, and institutional repositories. A pilot study conducted in 2021 demonstrated the potential of using BERT for subject indexing; however, the methods employed could be improved through more advanced technologies. [1] This project is part of an ongoing initiative focused on AI and metadata within the library metadata department, [2] and specifically aims to develop a deep learning pipeline for text classification using retrieval-augmented generation (RAG), embeddings, and neural networks. The goal is to enhance semantic search for library resources, using an open access eBook collection as a test case, with particular emphasis on metadata enrichment, such as generating enriched descriptions and facilitating crosswalks between controlled vocabularies.

Goals & research questions

This research pilot seeks to address a central question: How can we improve the discoverability of academic and library resources by leveraging state-of-the-art

*Corresponding author. † 0000-0003-4736-7662 (C. Chou) Charlene.chou@nyu.edu (C. Chou); ssk10036@nyu.edu (S. Khunti); hb2976@nyu.edu (H. Bhargava)

https://doi.org/10.23106/dcmi.952586098

technologies and high-quality metadata, rather than relying solely on keyword searching, which lacks semantic search capabilities? Our approach shifts the data processing workflow from traditional methods to a collaborative, data-driven, generative, and agentic AI pipeline designed to deliver enhanced metadata and enable semantic search. We aim to leverage machine learning and natural language processing to develop high-quality metadata and support scalable metadata enrichment, classification, and semantic discovery. To foster interdisciplinary collaboration, the metadata department provided metadata training to data science students. This initiative encouraged brainstorming and enhanced mutual communication between metadata librarians and students, helping to bridge disciplinary knowledge gaps. We are committed to conducting an ethical and responsible AI project and have selected an open access eBook collection published by New York University Press for testing, with permissions granted.

Library NLP system architecture

After two metadata training sessions, data science students began designing a library NLP system architecture. [3] This pipeline transforms full-text library content into a semantic search and recommendation service that helps patrons discover the most relevant resources. The high-level workflow follows these stages: data ingestion, preprocessing, embedding, vector database creation, semantic retrieval, large language model (LLM) reranking, and finally, top-k output. Below, we highlight a few key steps for illustration. The system architecture begins with data ingestion, incorporating the 272 e-book titles of Open Square collection along with title metadata, full text, and API endpoints. Data preprocessing involves tasks such as correcting encoding issues and applying Unicode normalization. The vector database construction includes creating a metadata store that contains pointers to the original text, enriched fields such as subject tags, and structured entity metadata retrieved via APIs such as VIAF (Virtual International Authority File). A key part of the approach involves using the DLTS API to retrieve metadata for each book and parsing the API responses into a standardized JSON format.

Conclusion

The project is currently underway and aims to complete the pilot by the end of summer or early fall. Upon reflection, this interdisciplinary initiative serves as a potential model for the library metadata department to initiate and foster collaboration with stakeholders across multiple domains, such as faculty, students from the Center for Data Science, and metadata librarians, in alignment with NYU's strategic pathways [4]. Additionally, the project leverages emerging technologies to develop innovative solutions, supports global education, and provides students with applied experience in metadata enrichment and AI workflows within a real-world digital library context. Through this initiative, we have further recognized that metadata plays a pivotal role in enhancing the discoverability of academic resources through the use of state-of-the-art technology. The NYU Center for Responsible AI emphasizes its goal of making "responsible AI" synonymous with "AI" [5]. Most importantly, this research project has adopted an approach that manages data and metadata in a systematic and responsible manner.

https://doi.org/10.23106/dcmi.952586098

References

[1] Chou, C and Chu, T. (2022). An Analysis of BERT (NLP) for Assisted Subject Indexing for Project Gutenberg. *Cataloging & Classification Quarterly 60.8* (2022), 807–835. Doi:10.1080/01639374.2022.2138666

[2] Chou, C. (2024). The Impact of AI on Metadata: AI Study Group for Learning and Interdisciplinary Collaboration. *International Conference on Dublin Core and Metadata Applications, 2024.* https://doi.org/10.23106/dcmi.952409098

[3] Bhargava, Harshit & Khunt, Shravan. *Library NLP system architecture.* URL: https://drive.google.com/file/d/1UK4beA1YmDjX1-pBnllyzAQw8s1K1UVo/view?usp=sharing

[4] New York University, Office of the President, Strategic Pathways. URL: https://www.nyu.edu/about/leadership-university-administration/office-of-the-president/strategic-pathways.html

[5] NYU Center for Responsible AI. URL: https://engineering.nyu.edu/research-innovation/centers/center-responsible-ai

https://doi.org/10.23106/dcmi.952586098