Andover ArchiveBot: Unlocking History in the OWHL Archives Using Fine-Tuned Retrieval Augmented Generation

Tianyi Gu for CSC600
April 7, 2025

This project aims to develop an AI-powered retrieval-augmented generation (RAG) system that enables intuitive, conversational access to archived materials from the Oliver Wendell Holmes library. This would be available to students, librarians, faculty, and any others who are curious about engaging with the school's history in an accessible manner. This project will enable users to leverage natural language search, in order for users to be able to retrieve summaries and context-rich information from historical documents.

Research Question and Project Description.

1. What fine-tuning and retrieval strategies are best for aligning the system's responses with the context specific to the OWHL archives?
2. What does an ideal user interface for exploring school archives through natural language interaction look like?
3. In what ways can we measure the quality, relevance, and accuracy of responses generated by a RAG system, specifically as built upon the OWHL archives?
4. How could a system like this be generalized to be able to serve as a foundation for systems working with other archival or historical datasets?

If successful, this project will be deployed successfully at the Andover OWHL, being actively used by students, faculty, and alumni. It could be a physical booth located somewhere in the entrance that is usable, as well as deployed on a website, aiming to provide a simplification in research conducted that involves the archives. This project will also contain a well-documented and reusable codebase that could be used for future projects with similar intent and a high-quality, well-documented, reusable codebase, serving as a reference for future student projects and potentially adopted or expanded upon in subsequent terms. Hopefully also include promotion/support of use by the OWHL staff, offering it as a resource in History/other classes that engage with archival materials.

Audience

I hope that my audience for this project can be two-fold. Firstly, one that can talk about and give feedback on the applied nature of the project, that is how well it functions and serves its purpose. For this, I would like to present and demo my project to the OWHL archival team, specifically Dr. Roberts, the Director of Archives and Special Collections. Here I hope to receive feedback on refinements, how well it works, and thinking about how such revisions can best support work done with the archives. I also hope to be able to present some more of the technical aspects behind how the system works, through disseminating it either to CS Club or perhaps through a video presentation or newsletter.

Outline of Work to be Done.

Week 1

- Identify and locate all the archive materials that I am thinking of working with.
- Set up Zotero for source management.
- Set up GitHub repository with for code versioning plus documentation system.

In this week, I plan to show Nick:
My Zotero organization, initial collected archival documents, and repository structure.

Week 2

- Build OCR/text extraction pipeline from PDFs or scanned documents into workable data.
- Establish chunking methodology for text.
- Store processed text for future embedding process.

In this week, I plan to show Nick:
A working pipeline demonstrating text extraction, cleaning, and chunking results.

Week 3

- Select an embedding model (e.g., SentenceTransformers).
- Set up vector database (FAISS or Chroma) and populate it with embeddings.
- Use semantic search tests to verify the retrieval quality.

In this week, I plan to show Nick:
A functional vector database capable of accurate semantic searches on archival data.

Week 4

- Deploy a local open-source LLM (e.g., Mistral or Llama-3 quantized).
- Try some initial prompt engineering and see how the basic generation is.

In this week, I plan to show Nick:
A running local LLM able to respond heuristically well to basic test prompts.

Week 5

- Integrate vector retrieval with LLM inference to form a RAG pipeline.
- Test pipeline by generating answers using retrieved archival content.
- Evaluate initial accuracy and make adjustments to the setup.

In this week, I plan to show Nick:
A preliminary working RAG system that answers questions using archival context.

Week 6

- Create a fine-tuning dataset (QA pairs) from archival material.
- Conduct LoRA-based fine-tuning on the local LLM using Hugging Face PEFT.
- Evaluate improvements and tune hyperparameters.

In this week, I plan to show Nick:
Successful fine-tuning setup and clear improvements over the initial LLM performance.

Week 7

- Develop a interactable web interface.
- Integrate the UI with the existing backend RAG system.
- Conduct user testing and get some initial feedback.

In this week, I plan to show Nick:
A working interactive UI with backend integration.

Week 8

- Optimize the entire system based on user feedback.
- Review the technical documentation, including all the source annotations.
- Complete the final video presentation.

In this week, I plan to show Nick:
My final project with thorough documentation!

Mindset/Skill Plan.
Mindset Name: Curator

"How I can make my work tell a compelling story?"

A Curator approaches projects and problems as a thoughtful collector, organizer, and presenter of complex ideas. The Curator carefully selects and integrates elements to design and develop systems that are both technically sound and also serve meaningful purposes to the intended users. They combine experimentation and iterative development with intentional design and communication. The Curator's strength lies in their ability and capability to shift between thinking about the technical depth of projects while also maintaining user-centered thinking.

Vital Skills:

- Experimental Design and Observation
- Intentional Design
- (Re)factoring Code

Supporting Skills:

- Questioning Impact
- Collaboration
- Testing Code

Underlying Skills:

- Writing Code
- Exploration and Research

Bibliographical Plan.

I plan on using Zotero for managing all the sources that I use, with each entry containing the full citation, link to the source, as well as a couple annotated notes with a main idea summary of 2-3 sentences, any key ideas or quotes relevant to my project, and why the source matters to my specific research. Each day, I will record any new resource (article, tutorial, documentation, dataset, or technical guide) that I used for my work, making sure that I add anything that was relevant to my work or influenced my thinking. Every week, at the end during Sunday evening, I will review sources recorded during the week and add any missing annotations.