Superfund RAG AI

Toscano, Adam; Howard, Ella; Ergezer, Memo; Larson, Josh

Department of Digital Technology Services, Schools of Sciences & Humanities and Computing & Data Science, Wentworth Institute of Technology

SuperfundGitHub

Abstract

- In an interdisciplinary, experimental research project, we are exploring the potential of retrieval-augmented generation for analysis of a large corpus of historical documents.
- Focusing on the historical records of the federal Superfund project, we seek to test and document analytical workflows using emerging technology.
- In this first phase of the project, we are establishing appropriate workflows to ensure accurate data collection and documentation practices.

Introduction

- Superfund is a program through the Environmental Protection Agency that seeks to remediate pollution sites in the United States.
- The relevant project documents are stored online in legacy government websites, creating a labyrinth of information that is difficult for users to navigate and aggregate.
- These barriers to data access have left the subject understudied.
- Prof. Howard set out to download and preserve relevant Superfund documents prior to any possible governmental deletion.
- As part of this research, she is partnering with the School of Computing and Data Science to use retrieval-augmented generation to experiment with data analysis.
- We are using RAG to do experimental research reading and analyzing historical documents.
- As part of the larger effort to download and analyze more than 250,000 EPA Superfund documents, this project will examine how different computational methods help users extract meaning from environmental policy texts.
- After constructing a clean, searchable corpus through AI-assisted OCR, the project will compare traditional sentiment analysis with newer interpretive approaches using large language models capable of contextual reasoning.
- Superfund texts are dense, technical, bureaucratic, and often emotionally muted, making them an ideal testing ground for exploring the limits of older sentiment tools (which often fail on jargon-heavy or neutral administrative prose) and the affordances of modern generative models, which can identify tone, stance, and rhetorical positioning even in highly formal federal documents.

We anticipate this RAG AI will save time that would have been spent reading the files and will surface new insights.

Methods

- **Data Collection**: Superfund site reports were downloaded from the EPA National Priorities List database as PDF Files
- **Text Extraction**: With the help of Virtual Studio Code and Python, we were able to download to our local computers the information from the site as each to read PDFs for Python.
- **Document Preparation**: Extracted text was cleaned and divided into smaller chunks for efficient retrieval throughout the program code.
- **Vector Indexing**: Document chunks were embedded and stored in a FAISS vector database to enable similarity searching to make the Ai have an easier time with problem solving.
- **RAG Question Answering**: User questions have their most relevant report sections retrieved, and an Ollama-based language model generates source-grounded answers based on the codes stored read in data.
- **Performance Testing**: The system was tested locally and on Unity Cluster to improve response speed and model efficiency to find which model works best for our RAG AI code.

Results

Key Document Data Fields:

- Contaminants: These are the pollutants that Superfund aims to clean up
- State/City/County: This is where the site is located
- Clean Up Cost/Process: These fields pertain to how it works and much it takes to clean up after one of the sites.
- Priority: This determines when in the Superfund queue the site will be attended to.

- We downloaded 2863 Superfund PDFs
- Out of the 2863 PDFs, 2862 were easily read by the RAG AI
- The RAG AI is very good at answering questions that have a concrete answer straight from the PDFs, such as contaminant name.

```
print(ask_superfund("Pennsylvania Ryeland Road Arsenic was responsible for what type of contamination?"))

To determine the type of contamination caused by Pennsylvania Ryeland Road Arsenic, we need to analyze the information from the Ryeland_Road_Arsenic.json
file.

**Primary Contaminant Information:**
Unfortunately, the primary contaminant categories and contaminants are not specified in the provided context.

**Similar Sites:**
However, since the Ryeland Road Arsenic site is an arsenic-related site, we can look at other sites with similar names. For example:

* The Perham Arsenic Site (MND980609572) also involves arsenic contamination.
* The Arsenic Mine (NYD982531469) site is another example of a site associated with arsenic.

**Conclusion:**
Based on the context provided and similarities with other sites, it can be inferred that Pennsylvania Ryeland Road Arsenic was responsible for arsenic-re
lated contamination.
```

Discussion

- The challenges faced during this project were primarily related to the data types and how the software can read the files
- When the code reads in the pdfs, it works best when the file is mostly standard text.
- Standard text is easy for OCR to handle, unlike images.
- Any document that is heavily reliant on images for displaying data is hindered due to the limitations of what OCR could work with.
- Despite the drawback of OCR and image recognition, most of the documents were primarily text making the RAG AI very efficient for consuming the data.
- Once the first working variant of the RAG AI was finished, the next goal was to improve the quality of responses and lowering load times.

Conclusions

This project gives the student practical experience building a scalable document-processing pipeline using AI-assisted OCR, Python automation, NLP techniques, and metadata engineering. These skills mirror real industry workflows in data engineering, information extraction, and AI-driven document analysis. By working with a large volume of raw federal PDFs, the student gains hands-on exposure to the kinds of high-volume, unstructured data challenges common in tech, government, and environmental consulting sectors.

Acknowledgements

We are grateful for the assistance of the staff at the High Performance Computing Center, including Berent.

References

1. Jacques, E. T., Roberts-Semple, D., & Blackman-Lees, S. (2025). Readability of Superfund information: an assessment of public health literacy in a tri-state area of the United States. Journal of Public Health: From Theory to Practice, 1–6. https://doi.org/10.1007/s10389-025-02546-6
2. Sabhanayakam, K., Kamat, A., & Zaidi, S. (2024). Comparative Study of Machine Learning Techniques in Prediction of Superfund Sites. 2024 International Conference on Machine Learning and Applications (ICMLA), Machine Learning and Applications (ICMLA), 2024 International Conference on, ICMLA, 1249–1252. https://doi.org/10.1109/ICMLA61862.2024.00194