Optimizing Historical Knowledge Retrieval: Leveraging an LLM for Content Cleanup

The Challenge

Enterprise Knowledge (EK) recently worked with a Federally Funded Research and Development Center (FFRDC) that was having difficulty retrieving relevant content in a large volume of archival scientific papers. Researchers were burdened with excessive search times and the potential for knowledge loss when target documents could not be found at all. To learn more about the client's use case and EK's initial strategy, please see the first blog in the **Optimizing Historical Knowledge Retrieval** series: *Standardizing Metadata for Enhanced Research Access*.

To make these research papers more discoverable, part of EK's solution was to add "about-ness" tags to the document metadata through a classification process. Many of the files in this document management system (DMS) were lower quality PDF scans of older documents, such as typewritten papers and pre-digital technical reports that often included handwritten annotations. To begin classifying the content, the team first needed to transform the scanned PDFs into machine-readable text. EK utilized an Optical Character Recognition (OCR) tool, which can "read" non-text file formats for recognizable language and convert it into digital text. When processing the archival documents, even the most advanced OCR tools still introduced a significant amount of noise in the extracted text. This frequently manifested as:

- A table, figure, or handwriting in the document being read in as random symbols and white space.
- Inserting random punctuation where a spot or pen mark may have been on the file, breaking up words and sentences.
- Excessive or misplaced line breaks separating related content.
- Other miscellaneous irregularities in the text that make the document less comprehensible.

The first round of text extraction using out-of-the-box OCR capabilities resulted in many of the above issues across the output text files. This starter batch of text extracts was sent to the classification model to be tagged. The results were assessed by examining the classifier's evidence within the document for tagging (or failing to tag) a concept. Through this inspection, the team found that there was enough clutter or inconsistency within the text extracts that some irrelevant concepts were misapplied and other, applicable concepts were being missed entirely. It was clear from the negative impact on classification performance that document comprehension needed to be enhanced.

Optimizing Historical Knowledge Retrieval: Leveraging an LLM for Content Cleanup

Auto-Classification

Auto-Classification (also referred to as auto-tagging) is an advanced process that automatically applies relevant terms or labels (tags) from a defined information model (such as a taxonomy) to your data. Read more about Enterprise Knowledge's auto-tagging solutions here:

- 4 Steps to Content Auto-Classification with High Accuracy
- Expert Analysis: When should my organization use auto-tagging? Part One
- Knowledge AI: Content Recommender and Chatbot Powered by Auto-Tagging and an Enterprise Knowledge Graph
- A Guide to Selecting the Right Auto-Tagging Approach

The Solution

To address this challenge, the team explored several potential solutions for cleaning up the text extracts. However, there was concern that direct text manipulation might lead to the loss of critical information if blanket applied to the entire corpus. Rather than modifying the raw text directly, the team decided to leverage a client-side Large Language Model (LLM) to generate additional text based on the extracts. The idea was that the LLM could potentially better interpret the noise from OCR processing as irrelevant and produce a refined summary of the text that could be used to improve classification.

The team tested various summarization strategies via careful prompt engineering to generate different kinds of summaries (such as abstractive vs. extractive) of varying lengths and levels of detail. The team performed a human-in-the-loop grading process to manually assess the effectiveness of these different approaches. To determine the prompt to be used in the application, graders evaluated the quality of summaries generated per trial prompt over a sample set of documents with particularly low-quality source PDFs. Evaluation metrics included the complexity of the prompt, summary generation time, human readability, errors, hallucinations, and of course - precision of auto-classification results.

Optimizing Historical Knowledge Retrieval: Leveraging an LLM for Content Cleanup

The EK Difference

Through this iterative process, the team determined that the most effective summaries for this use case were abstractive summaries (summaries that paraphrase content) of around four complete sentences in length. The selected prompt generated summaries with a sufficient level of detail (for both human readers and the classifier) while maintaining brevity. To improve classification, the LLM-generated summaries are meant to supplement the full text extract, not to replace it. The team incorporated the new summaries into the classification pipeline by creating a new metadata field for the source document. The new 'summary' metadata field was added to the auto-classification submission along with the full text extracts to provide additional clarity and context. This required adjusting classification model configurations, such as the weights (or priority) for the new and existing fields.

Large Language Models (LLMs)

A Large Language Model is an advanced AI model designed to perform Natural Language Processing (NLP) tasks, including interpreting, translating, predicting, and generating coherent, contextually relevant text. Read more about how Enterprise Knowledge is leveraging LLMs in client solutions here:

- What is a Large Language Model (LLM)?
- Choosing the Right Approach: LLMs vs. Traditional Machine Learning for Text Summarization
- The Role of Semantic Layers with LLMs

Optimizing Historical Knowledge Retrieval: Leveraging an LLM for Content Cleanup

The Results

By including the LLM-generated summaries in the classification request, the team was able to provide more context and structure to the existing text. This additional information filled in previous gaps and allowed the classifier to better interpret the content, leading to more precise subject tags compared to using the original OCR text alone. As a bonus, the LLM-generated summaries were also added to the document metadata in the DMS, further improving the discoverability of the archived documents.

By leveraging the power of LLMs, the team was able to clean up noisy OCR output to improve auto-tagging capabilities as well as further enriching document metadata with content descriptions. If your organization is facing similar challenges managing and archiving older or difficult to parse documents, consider how Enterprise Knowledge can assist in optimizing your content findability with advanced AI techniques.

Enterprise Knowledge (EK) is a services firm that integrates Knowledge Management, Information Management, Information Technology, and Agile Approaches to deliver comprehensive solutions. Our mission is to form true partnerships with our clients, listening and collaborating to create tailored, practical, and results-oriented solutions that enable them to thrive and adapt to changing needs.

Our core services include strategy, design, and development of Knowledge and Information Management systems, with proven approaches for Data and Information Management, Knowledge Graph Implementation in support of NLP, ML, and AI initiatives, Taxonomy Design, Project Strategy and Road Mapping, Brand and Content Strategy, Change Management and Communication, and Agile Transformation and Facilitation. At the heart of these services, we always focus on working alongside our clients to understand their needs, ensuring we can provide practical and achievable solutions on an iterative, ongoing basis.

info@enterprise-knowledge.com | 571-403-1109 | @EKConsulting