AI-Augmented Content Analysis to Remediate Duplicate Content

The Challenge

In alignment with the UN Paris Agreement, a global energy company with operations in over 60 countries began actively working to reduce their carbon emissions to achieve Net Zero by 2050 within three key domains:

- Direct greenhouse gas emissions produced from sources that the company can operationally control;
- Indirect greenhouse gas emissions from the generation of purchased energy produced during the effort of producing energy (e.g., by the company's energy-producing assets);
- Other indirect greenhouse gasses, including those emitted from the process of using energy products sold by the company.

The company's information management (IM) team engaged Enterprise Knowledge (EK) to evaluate and refine their strategic roadmaps, which yielded a plethora of new initiatives that built on their existing efforts. One of these initiatives focused on developing a "green" information management sustainability strategy. To achieve this, their IM team sought a pilot to address some of their most pressing challenges, as identified and synthesized by EK:

- Proliferation of duplicative content, based on the organization's internal formula for carbon emissions per GB of content storage;
- High barrier to entry to reduce duplication proactively (e.g., linking existing content requires more effort than making copies); and
- Collaboration software unintentionally built silos and promoted content duplication due to a lack of visibility and awareness.

The Solution

To address the challenge of carbon emissions created due to content storage, EK designed a web application that uses AI technologies to identify duplicate content and evaluate options for handling it. The application dashboard displays aggregate statistics on the presence and type of duplicate content, allowing users to make a decision as to whether it should be retained, updated, archived, or deleted. This provides a clear view into duplication and its connection to CO2 emissions, promoting a cultural shift among employees to increase awareness about carbon footprint and the role they play in contributing to a wider sustainability strategy.

One of the striking aspects of this effort was the sheer magnitude of the IM team's content collection, estimated at 500 million documents and many petabytes of information. The EK team worked with our stakeholders to gather and prioritize requirements and build the application with this scale in mind, including a number of key components:

- Identification of content storage locations;

- Indexer to crawl the content collection;
- Data pipeline to convert content to a vector database that allows for closer examination;
- AI model to continuously identify duplicate content within designated content locations; and
- Method of showing scope and impact of duplicate content to the end user.

The technical team leveraged Azure Open Source AI and Power BI to design a prototype dashboard to quantify duplicate unstructured and semi-structured assets based on scans, indexes, and queries. EK also used reusable code wherever possible to further minimize computing power and carbon emissions. Leveraging the metadata and textual content, AI-based analysis can rate the likeness of other information previously indexed.

The team created a value statement and strategic roadmap that will continue to provide guidance to the company on continued expansion of their Green IM tool. It included consideration of the complexities of their environment for topics like scaling, rollout, and customer footprint growth, underscored with the importance of change management as a critical component. As part of the roadmap, the team was also able to identify opportunities to selectively introduce proactive and reactive automation in order to help users reduce their content duplication throughout the course of their normal day, such as warning notifications when uploading content that meet a certain similarity threshold with an existing content item in a given repository (proactive) or enabling system triggers to remove duplicative content through the dashboard interface itself (reactive). The ultimate future state goal, as identified in the roadmap, is to enhance the web application and make it actionable in supporting push-of-a-button content deletion through the application itself, promoting a pure “don’t make me think” content experience and further behavioral change.

The EK Difference

EK leveraged our history with the company, our understanding of their strengths and challenges, and a balanced team of technical and strategy subject matter experts (SMEs) to proactively propose the idea of a “green” application, based on their enterprise effort to reduce greenhouse gas emissions.

The EK team also developed a customized scorecard to evaluate the pilot’s success and ensure alignment to the company’s strategic objectives, including measurable factors like:

- The ability to calculate carbon footprint with at least 80% accuracy;
- 75% of non-technical users report being able to use the application on their first try with minimal training; and
- A solution that abides by all architectural and security requirements established by the technical team.

AI-Augmented Content Analysis to Remediate Duplicate Content

This comprehensive understanding of the challenge, paired with the delivery of a tailored mechanism for assessing the proof of concept's success, enabled EK to perfectly position the organization to take on similar efforts in the future.

The Results

EK identified potential pathways to quickly address carbon emissions within the IM team, with an initial focus on reducing the amount of duplicate content within their repositories.

In partnership with the client, EK identified a pilot set of 226 million of the company's approximately 500 million total documents to prove out the concept.

With a 15% target deduplication rate this company has the potential to remove over 34,000 kilograms of CO2 from the environment through a reduction in physical server usage, directly supporting their objective to remove greenhouse gas emissions from operations that they are capable of controlling.

Just over 1 PB (or, 1.07 million GB) **adds nearly 228,000 kg of CO_2** to the environment, just from existing

226,000,000

Documents Stored in their Repositories

Beyond providing tangible, quantifiable statistics upon which to build a business case for larger ESG initiatives, this initiative also provided the IM team with a repeatable framework for running similar "green" efforts in the future, as well as faster and more accurate decision making through less clutter and quicker access to content.

33,900,000

Documents Removed through 15% Reduction in Duplication

Removing 160,500 GB **removes over 34,000 kg of CO_2** from the environment

In doing so, EK was able to demonstrate the technical and business viability of an AI-driven content deduplication tool. The pilot use of this tool demonstrated the accurate identification of duplicates based on conversion to a vector database and AI modeling to identify – and fine tune – duplicate content.