Data Science Clinic: Argonne Knowledge Graph Final Report

       Argonne Operations is responsible for ensuring that Argonne National Laboratory activities adhere to a myriad of recorded guidelines. Our team of UChicago students aims to reduce the amount of human effort and error involved in reviewing the fidelity of the lab's policy documents by leveraging AI models and knowledge graph systems to map out document relations.

       Previously, an AI model was built that divided these documents into segments and calculated a score to quantify similarities between segments. This project's contributions involve visualizing the model's similarity scores using Neo4j, with document segments as nodes and the similarity scores between them as edges.

       Upon building an initial visualization (Figure 1), two problems arose. (1) The document segments had an imbalanced number of words and contained irrelevant data, such as "Intentionally Left Blank" pages. (2) The current model of the graph revealed clusters of closely related document segments, but the nature of these clusters remained unclear.

       For the first problem, the team developed an enhanced segmentation algorithm to re-segment the policy documents consistently at the paragraph level. The text post-processing algorithm has also been improved to remove symbols and sentences irrelevant to the content. Following this, the AI model was reapplied to recalculate the similarity between each segment.

       Furthermore, to enhance the informativeness of the graph, our team constructed an intuitive, keyword-based version of the knowledge graph using the latest GPT-4 model. First, the GPT-4 model was used to extract five unique, non-synonymous keywords from each segment. These keywords were then matched across segments to establish clearer connections. As demonstrated in Figure 2, the visual representation was further refined: the size of each keyword node in the graph corresponds to the number of segments related to that keyword, offering a more intuitive understanding of the document's content and the relationships between different segments.
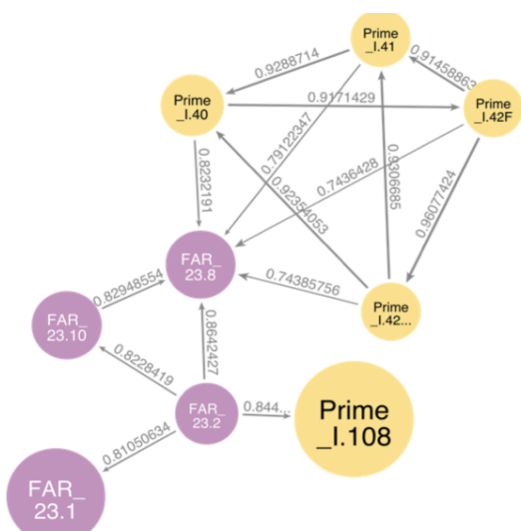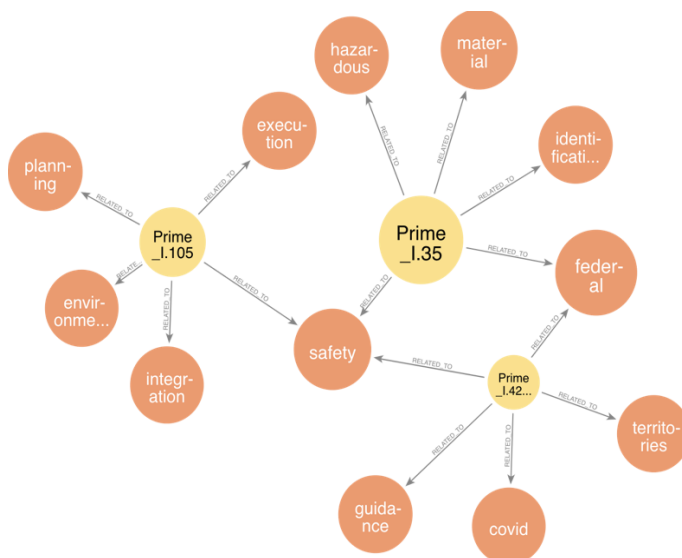

Figure 1. Neo4j Graph with Cosine Similarity   Figure 2. Neo4j Graph with Keywords