DSI Argonne Project

Argonne National Laboratory has a large amount of policy documents. Searching through these documents for specific information is labor intensive. To support employees in making decisions, they want to make the policy information more accessible by creating a searchable network. The network would serve as the basis for a chatbot-style user interface as well as a digital twin simulation demonstrating downstream effects of policy changes.

The Data Science Clinic students were tasked to build an early version of this network.

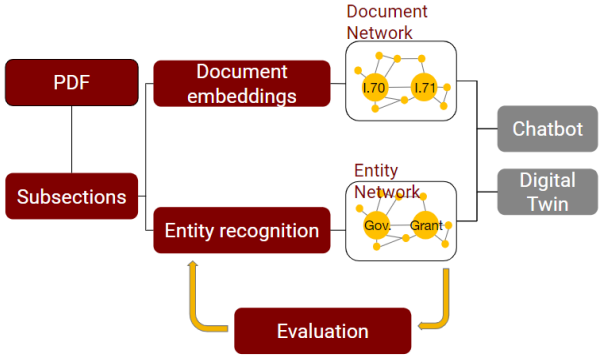

Fig 1. Workflow of the Argonne & DSI project. Red represents processes the students in the clinic completed, and grey represents future use-cases.

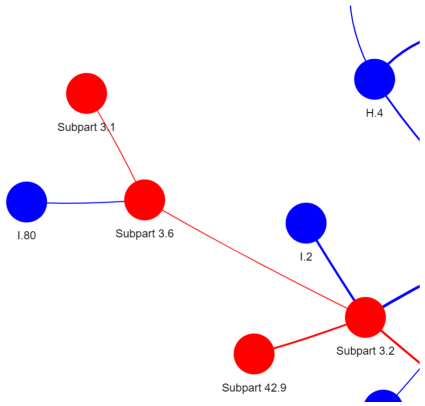

Fig 2. Final network connecting subsections. Blue and red represent different documents.

Following the workflow in Figure 1, students extracted cleaned text and metadata (e.g. bold, italics, bullet points) from PDF files of internal policy documents, consisting of thousands of pages. Students then built tools to utilize state-of-the-art machine learning models to transform text data into numerical representations (embeddings). These numerical representations were used to create two types of networks: one where documents are connected to each other through similarity (e.g. I.70 has a 0.9 similarity to I.71), and another where entities are connected to each other through relationships (e.g. the government pays contractors). Out of five state of the art models, one was chosen for its simplicity and accuracy. A portion of the resulting network connecting subsections to each other is exemplified in Figure 2. The network was then manually evaluated by comparing similarity scores given by the model to scores given by people in order to determine a meaningful threshold for when connections should be drawn. The evaluation also provides a baseline comparison for future improvements.

The network developed by the students served as an initial step towards a full knowledge graph, building the foundation for a chatbot and digital twin interface. Searching for relevant documents in the document network allows a chatbot to answer questions (e.g. “Who pays contractors?”) or identify documents affected by policy changes (e.g. “A grant will now pay for contractors”). Thus, students have created a tool utilizing state of the art machine learning models that bridges the way to a searchable policy database.