The directory contains the following folders. A brief description of the folders have been provided below.

Before-Chunking

The directory is used to download unstructured text from online sources, such as API endpoints and PDFs hosted on various web applications. If additional sources are used in the future, this same directory can be utilized to save those resources.

Embedding_Downloads_CSV

The application provides the capability to download text embeddings once they are created or updated in the vector database. This directory is designated for saving these embeddings for future reference or troubleshooting purposes.

Gephi_Images

After importing the .gdf files into Gephi and plotting network diagrams, you may need to take screenshots or download images of these diagrams for cataloging, future reference, and demonstrations. This directory is designated for saving such images. Please note that there is no automated code to capture or save Gephi images in this directory; this task must be performed manually.

Gephi_Imports_GDF

A Python program has been implemented to process the text embeddings and convert them into a .gdf file, which is essential for conducting network analysis in Gephi. This designated directory is used to store the .gdf files produced by the Python program.

Gephi_Projects

Throughout this project, the team created network diagrams for the application domains studied—USSP, GAO, and Farm Bill. Each domain was analyzed separately, and corresponding network diagrams were generated. After the diagrams are produced, Gephi requires the resources to be saved as project files, which are stored in the designated folder. Please note that there is no code to automate the download and saving of these files; this task must be done manually.

Logos

Certain logos were used on the application's user interface, and these logos are stored in the designated directory. In the future, if new application domains are added to the tool, logos for these domains should also be stored in this directory.

Notebooks

The project team implemented four Jupyter notebooks to host the application code. Two notebooks were developed to fetch raw text from external sources, one to build the core logic of the application, and the last one to process embeddings and create .gdf files. A brief description of each notebook will be provided later in this document.

Transposed_Embeddings_CSV

Before creating the .gdf file, an intermediate task involves building a data frame with document chunks as column headers and listing the embeddings vertically beneath them. The program that generates the .gdf file requires a .csv file constructed from this data frame. These .csv files are eventually downloaded and stored in the designated directory.

Vector_DB_Embeddings

The core program for this application is designed to implement persistence when embeddings are created and written to the vector database. This persistent logic enhances efficiency by enabling direct retrieval of embeddings whenever needed to answer user prompts. The persistent directories for each application domain are stored in the designated directory.

Report - Technical Implementation

This document is not a formal project report. To access the official project report, please visit: Project Report.

The purpose of this technical implementation report is to detail the various features built into the prototype application and their development process. Additionally, this report serves as a guide for custodians and the development team to enhance the system and implement future use cases effectively.