Project Title: Developing an Efficient Vector Database System for Legal Document Retrieval

Faculty Member: Keith Perkins, CNU PCSE

External Institution: Width.Ai (Collaborator: Patrick Hennis from Width.Ai)

Project Summary:

This research project aims to develop an efficient vector database system for retrieving relevant legal documents based on user queries. The primary objectives are to explore and evaluate various embedding models, compression strategies, and vector database indexing techniques to optimize the retrieval performance for legal documents.

The project will leverage pretrained language models to generate embeddings for a dataset of legal documents provided by Width.Ai. These embeddings will be stored in a vectorized database, enabling efficient retrieval of documents similar to user queries encoded using the same embedding model.

The research methodologies will involve the following steps:

1. Conduct a literature review on vector databases, distance metric algorithms, embeddings and embedding compression strategies to identify best practices for legal datasets.

2. Preprocess the legal document dataset, addressing structured data challenges and identifying optimal preprocessing techniques to ensure high-quality data for embedding generation.

3. Evaluate various embedding models (considering context length, compute requirements, and cost) to find models that generates accurate embeddings for legal datasets.

4. Generate embeddings from the preprocessed dataset using the selected embedding models and evaluate different embedding compression strategies, such as quantization and binary/scalar techniques, to optimize storage and retrieval performance.

5. Evaluate various vector database similarity search indexes (e.g., flat, IVF) and distance metrics (e.g., cosine, dot product, Euclidean) to determine the most accurate and efficient combination for legal document retrieval.

6. Research methods to identify dissimilar yet domain-relevant contexts from a dataset, distinct from the user query, to enhance the retrieval capabilities of the system.

7. Collaborate with Width.Ai researchers, to share findings, address challenges, and explore potential avenues for further improvement.

The expected outcomes of this project include a comprehensive understanding of the trade-offs between different embedding generation models, compression strategies, and vector database indexing techniques in the context of legal document retrieval. Additionally,

the project aims to deliver recommendations for creating robust and efficient vector database systems tailored for the legal domain, enabling fast and accurate document retrieval based on user queries.