SOP
Machine Learning

Kriti Arora (12240880) Mallarapu Hema Varshini(12240950)

Kaki Venkata Vaneesha(12240740) Bhavya Jain(12240420)

September 2024

Introduction:

Estimating similarity between legal documents is a pivotal task with significant implications for legal research and practice. Legal documents are often interlinked through citations, making the challenge of determining document similarity both intricate and crucial. Our objective is to develop a robust methodology to address this challenge, leveraging advanced machine learning techniques to enhance legal information retrieval and citation recommendation systems.

Motivation:

The ability to accurately estimate the similarity between legal documents is of paramount importance. Legal professionals and researchers frequently need to retrieve prior cases and recommendations based on the relationships between documents. Traditional methods have struggled with the complexity of legal language and the structured citation networks inherent in legal texts. By improving similarity measures, we can facilitate more effective case retrieval and citation recommendations, ultimately streamlining legal research and decision-making processes.

Objective:

Our primary objective is to develop a novel approach that measures the similarity between legal documents by integrating text embeddings and citation network structures. By achieving accurate similarity assessments, we aim to enhance various information retrieval tasks, including prior-case retrieval and citation recommendation. This work will bridge the gap between citation network-

based and text-based approaches, offering a comprehensive solution to document similarity in the legal domain.

Relevant Study:

Previous research has explored various methodologies for document similarity, including Recurrent Neural Networks (RNNs), Doc2Vec, and combinations of citation networks and text embeddings (Paheli ewt al.(2022). Notably, TextGCN (Liang Yao et al. (2019)) has shown promise in text classification by generating embeddings without relying on external transformer models. Our approach seeks to build upon these findings by employing TextGCN to obtain document embeddings and leveraging Graph Convolutional Networks (GCNs) to analyze citation structures. Also we apply event extraction (Joshi et al.(2023)) to the documents before forming the graph to feed to TextGCN. This dual approach aims to combine the strengths of both text representation and citation analysis for improved document similarity assessment.

Proposed Solution:

We propose a novel approach that adapts the TextGCN model to the legal domain to measure document similarity. Our solution involves several key steps:
Dataset Collection: Compile a comprehensive dataset of legal documents interconnected through citations. This dataset will serve as the foundation for our similarity analysis.
Pseudo Similarity Generation: Utilize an existing model to generate initial similarity values for the documents, providing a baseline for comparison.
Event Extraction and Embedding Generation: Apply event extraction (Finding Subject-Verb-Object triplets) techniques to the dataset and employ the TextGCN model to generate embeddings for both documents and words. This step captures the semantic and contextual information embedded in the legal texts.
Graph and Embedding Integration: Integrate the citation graph of legal documents with the embeddings produced by the TextGCN model. This combined approach will enable a comprehensive analysis of both textual content and citation relationships. We can use Neural networks here or just simply concatenate or add the embedding vectors.

Ensemble Learning: Combine the insights from both TextGCN and GCN using ensemble learning techniques. This approach will enable the model to learn from both text representations and citation networks, enhancing the accuracy of similarity assessments.

Similarity Verification: Validate the similarity values obtained by comparing them with the pseudo similarity values. Refine the model to ensure high

accuracy and reliability in document similarity assessment.