Legal Document Similarity Analysis

Corresponding Author / Report Owner: Aayushi Jeeban Tripathy

Aayushi Jeeban Tripathy,
Department of Computer Science and Engineering,
Amrita School of Computing,
Bengaluru, India
BL.EN.U4CSE21003@bl.students.amrita.edu

S Navin Sunder,
Department of Computer Science and Engineering,
Amrita School of Computing,
Bengaluru, India
BL.EN.U4CSE21176@bl.students.amrita.edu

Siddhant Ashwani
Department Computer Science and Engineering,
Amrita School of Computing,
Bengaluru, India
BL.EN.U4CSE21189@bl.students.amrita.edu

***Abstract*: The similarity analysis of legal documents is done with the intention of improving document retrieval, grouping, and legal understanding. It uses ideas from text mining, machine learning, natural language processing, and information retrieval techniques. The main goal of the comparison is to compare the legal document's specifications, which will be vectorized, with the rest of the vectorized documents using text analysis. From there, a similarity score will be determined. The document that has the highest document score is assumed to be closely linked to the current document and will thus be shown. The article compares the accuracy of the models with those of the others and covers the various techniques currently employed for document similarity analysis.**

Keywords: legal document similarity, document matching, text similarity, common law system, machine learning, Natural language processing, Information retrieval

II. INTRODUCTION

A fundamental component of the legal profession, legal papers are the main method used to record and uphold laws, contracts, and agreements. These records come in a variety of formats, such as contracts, laws, rules, judgments, and legal opinions. The administration and analysis of these papers are of utmost importance since the legal profession strongly depends on their accurate interpretation and use. The research necessary for each legal matter is a crucial part of the preparatory process. These records are frequently voluminous and conceal crucial details of the case.

The legal documents are classified into two main categories: (I) previous court decisions and (ii) statutes, which are the written versions of a jurisdiction's laws that specify and punish certain crimes. The Common Law System places a great deal of weight on precedents, or earlier decisions that are comparable to the one at hand. Therefore, in order to comprehend and defend various legal features of a specific case, legal professionals must find numerous case papers that are comparable to it. Law professionals and academics need automated systems for searching and suggesting similar instances for a specific case since there are so many previous cases. The paralegals and solicitors find it challenging to draw out the aforementioned features as a result. For it, a deep analysis of the various legal documents is required, and for doing so, there are a set of methods and algorithms that have been in current use. The methods currently in use broadly classify themselves into two categories: (1) network-based methods depend on the citations from the old documents, and (2) text-based methods make use of the textual information. Among these two categories, we have some methods, namely Text similarity, which compares the set of documents using techniques like word overlap, stemming, etc. Document structure, which compares structure of the documents like the order of the paragraphs, structure of matter, etc., and legal concept similarity, which compares the legal concepts that are discussed in the documents

This report finds that its goal is to compare the new cases with the ones that have already been resolved, focusing on the ones that have the greatest relative similarity in terms of case domain and issue for effective document clustering and semantic understanding in the legal domain. This is done in order for the paralegals to relate to the old cases and be prepared for court proceedings, taking into account the laws and sections of IPC used in the previous cases to be solved.

III. LITERATURE SURVEY

Definition: Legal document similarity analysis is a technique used to measure the similarity or relatedness between two or more legal documents, such as contracts, statutes, case opinions, or legal texts. The primary objective of this analysis is to assess how closely related or similar these documents are in terms of their content, structure, or legal concept

Key concepts and aspects of legal document similarity analysis:

1. **Textual Similarity:**

At its core, legal document similarity analysis involves comparing the textual content of legal documents. This comparison can be based on various NLP techniques, including but not limited to:

· **Cosine Similarity**: This measures the cosine of the angle between the vector representations of documents. Documents with a smaller angle (a cosine value close to 1) are considered more similar.

· **Jaccard Similarity**: This calculates the size of the intersection of terms in two documents divided by the size of their union. It is often used for comparing sets of words or phrases.

· **Word Embeddings**: Techniques like Word2Vec or GloVe can be used to represent words and phrases in high-dimensional vector space, and document similarity is calculated based on the similarity between these vector representations.

2. **Legal Ontologies:**

To enhance the analysis, legal ontologies or knowledge graphs may be used. These ontologies capture legal concepts, relationships, and terminologies. By incorporating legal ontologies, the analysis can consider the semantic relationships between legal terms and concepts.

3. **Document Matching:**

It is the process of identifying and pairing similar or related documents from a corpus of legal texts. This task is crucial in the legal field for various purposes,such as finding precedents, identifying potential plagiarism, or locating relevant documents for legal cases.

4. **Text Similarity:**

It is a fundamental concept that helps measure the degree of resemblance between pieces of text, such as legal documents, contacts, case law, and regulations. Legal professionals often use text similarity techniques to identify similar documents, precedents, or relevant passages for various purposes.

5. **Natural Language Processing (NLP):**

It is a subfield of Artificial Intelligence that focuses on the interaction between computers and human languages. It aims to enable computers to understand, interpret, and generate human language in a valuable way.

6. **Information Retrieval:**

It is the process of assigning labels or categories to documents based on their content. Finding relevant documents or passages from a large corpus of text in response to user queries

7. **Machine Learning:**

It is a technique to implement Artificial Intelligence that can learn from data without being explicitly programmed.

8. **Named Entity Recognition (NER):**

Identifying and comparing entities such as names of individuals, organizations, dates, and monetary amounts within legal documents can contribute to similarity analysis.

9. **Application Areas:**

· **Legal Research:** Researchers and legal professionals can use similarity analysis to find relevant cases, statutes, or legal documents related to a specific topic.

· **Contract Analysis:** Businesses can analyze contracts to identify similarities or divergences between different versions or templates.

· **Plagiarism Detection**: In academia or legal practice, similarity analysis helps detect cases of plagiarism or unauthorized copying of legal documents.

Comprehensive Research on the papers:

1. Legal case document similarity:

This research paper focuses on the similarities between two legal case documents. The two methods used for the task are citation network-based and text-based.

Citation network- based approaches consider only prior cases, which are also called as precedents (PCNet). Researchers have proposed Hier-SPCNet, which augments PCNet with a heterogeneous network of statutes. The dataset used in the research is from the Indian Supreme Court, where the similarities between the document pairs are annotated by Law experts from two different reputed Law institutes in India: (1) the Rajiv Gandhi School of Intellectual Property Law (RGSOIPL) and (2) the West Bengal National University of Juridical Sciences (WBNUJS). The first dataset is used as the validation set for tuning various hyper-parameters in the similarity estimation methods, and the second dataset is used as the test set to evaluate the performance of the methods [3]

2. Legal document similarity: a multi-criteria decision-making perspective

Legal information retrieval (LIR) aims at retrieving legal information objects relevant to a user's query. Legal information objects are various documents like court transcripts, verdicts, legislation documents, and judgments that are generated during the course of a legal process. These documents are primary resources for the interpretations of the law of any judiciary and hence are required by a legal professional for decision-making as well as argumentation. Methods and techniques used in LIR originate from confluence of four major technologies: artificial intelligence (AI), Network Analysis, Machine learning, and NLP. [4]

3. Methods for Computing Legal Document Similarity: A Comparative Study

This paper talks about the existing automatic techniques for finding similarity in documents that are network-based and text-based. It focuses on comparing all the existing methods of finding legal document similarity in order to ensure a fair comparison. The dataset contains 47 pairs of Indian Supreme Court case documents, where the similarity between each pair of documents is annotated on a scale from 0 to 10 by law experts. [5]

4. Analysis of Law Document based on Word2vec:

As the title suggests, this research paper talks about using the Word2Vec model for conducting document analysis of legal documents. The Word2vec model has the advantage that it can improve the accuracy of analysis by '0.20' when compared with the 'BOW model'. To summarize the index terms, they are: Learning, intelligence, Word2Vec, and similarity analysis. [6]

5. Similarity Analysis of Legal Documents Using Content and a Network-Based Approach:

This paper has proposed 'CaseRex,' which is a case recommendation system. It uses both content-based and network-based similarity measures for document similarity. The CaseRex algorithm is divided into two parts: Document Modeling and Relevant Document Identification. This algorithm structures the database of documents in a way that preserves their inherent relationships via the citations. Keywords used: natural language processing (NLP), Document Handling, and text analysis. [7]

6. An Intelligent approach towards legal text document retrieval:

The main focus of this paper is to retrieve relevant documents for particular Supreme Court cases in India from a set of prior case documents. For each distinct document, the current system lists a set of documents with their ranking scores using N_similarity, Fastt-Tex, and 'BERT'. The system returns the relevant document based on their ranking score. Database used: Document from Supreme Court of India Index terms used: information Retrieval, Legal Document, Natural Language Processing (NLP), Document Similarity [8]

7. JPReg, a novel method to identify paragraph regularities in legal case judgments:

Legal experts have to prepare a lot of legal documents, which is time- consuming. Paragraphs of existing documents that can be used, known as paragraph irregularities, need to be identified. JPReg is a method to help with this aspect that can be adopted by legal experts. JPReg adopts a two-step approach: in the first step, all similar documents are clustered according to semantic content, and in the second step, regularities are identified in the paragraphs for each cluster. Text embedding methods are adopted to represent a numerical nearest neighbor search Method is used to retrieve most similar paragraphs with respect to target document. This method has been found to be very effective in several experiments performed on real-world datasets.

8. Clustering legal documents using Topic Modeling

The amount of information available for the justice system is voluminous, and it is extremely difficult for legal experts to sieve through so much of legal document information that is available. Legal documents, if managed properly as nuggets of knowledge, would help legal experts. If the legal documents are in a knowledge repository that is properly organized and if the legal experts can use a search engine on this knowledge repository for retrieval,analysis, and presentation, it is a very useful service. The clustering of similar documents can be done using topic modeling.

9. Exploring and Inferring Precedent Citations in Legal Documents Using a Web Visual System:

The number of legal cases is ever increasing, and if binding precedents can be created for the lower courts by the Supreme Court, it will help in faster resolution of cases. A database of frequently cited binding precedents can be created, and this can be made available in a web-based visual analytics system to support analysis of legal documents that cite a binding precedent. Binding precedents can be created. The visual system can have three interactive components: the first showing an overview of data showing temporal patterns; the second grouping relevant documents by topic; and the third pointing to paragraphs that are likely to mention the binding citation.

This has been successfully implemented in the Brazilian judicial system with an accuracy of 96%.

10. Clustering and Summarization of Legal Documents

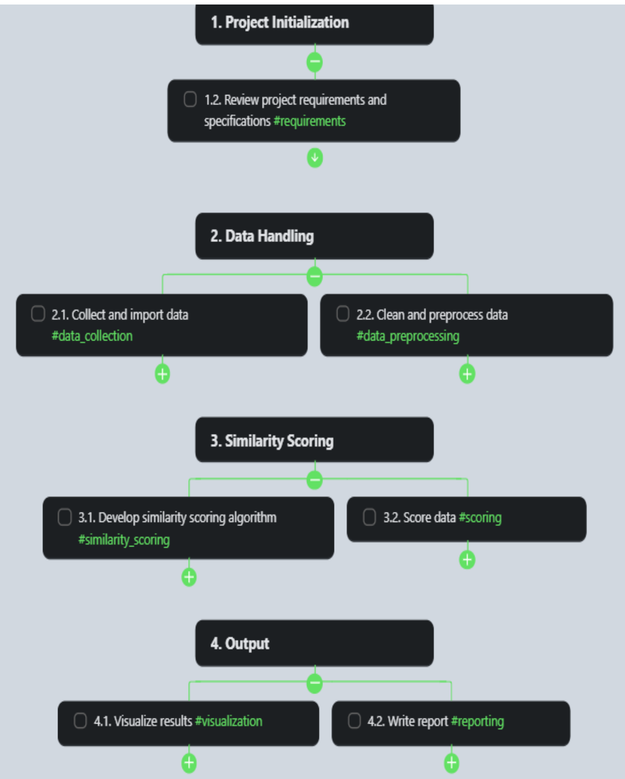
The increase in legal documents available led to digitization of the legal documents and the need for knowledge management of all these digitized documents. There is a need for an automated tool for organizing, analyzing, retrieving, and displaying relevant content. A search engine can then be used to access the system of digitized documents. This can be very beneficial if automated summary of legal documents is also available, as that can be easily understood by common man. As an added feature, the system can display recommendations from well-known lawyers.

IV. DATASET DESCRIPTION AND METHODOLOGY

The dataset that is made use of in this report is provided in the form of an excel workbook, "19CSE305_LabData_Set3.xlsx," comprising 2 sheets, namely: "thyroid0387_UCI," which is a thyroid-related dataset from the UCI Machine Learning Repository, and the other one called "marketing_campaign," which includes marketing-specific customer behavior data. These datasets are utilized for study purposes and for understanding various data- related applications and operations, and thus form the basis of our report methodology for deriving the facts and analyzing them.

There are numerous critical phases involved in designing a system for customer and patient segmentation utilizing the "thyroid0387_UCI" dataset. A high-level summary of the system that includes a flow diagram, an architectural diagram, parameter definitions, and explanations for each one is provided below.

I) *FLOW DIAGRAM*



1. Project Initialization

□ 1.2. Review project requirements and specifications #requirements

2. Data Handling

□ 2.1. Collect and import data #data_collection

□ 2.2. Clean and preprocess data #data_preprocessing

3. Similarity Scoring

□ 3.1. Develop similarity scoring algorithm #similarity_scoring

□ 3.2. Score data #scoring

4. Output

□ 4.1. Visualize results #visualization

□ 4.2. Write report #reporting

II) *ARCHITECTURE DETAIL ANALYSIS*

1. PROBLEM STATEMENT:

i) Acquire the dataset "thyroid0387_UCI".

ii) Load the dataset into memory so that it may be processed later.

2. DATA HANDLING:

i) Dealing with missing values Decide whether to add the missing data or remove it.

ii) Categorical characteristics should be converted to a numerical representation (for example, one-hot encoding).

iii) Calculate the mean and standard deviation of numerical characteristics on the same scale to Standardize and normalize data.

iv) Create training and testing sets from the data.

v) Select the segmentation attributes that are the most informative.

vi) Reduce the number of features while maintaining crucial data.

vii) Scale features to a comparable range.

3. SIMILARITY SCORING:

i) Calculate similarity scores between people using the proper metrics by making use of cosine similarity and Euclidean distance.

ii) Produce a matrix of similarity.

iii) Use clustering methods to put people in groups depending on how similar they are, using K-means or hierarchical clustering.

iv) Use techniques like the elbow method or silhouette score to determine the ideal number of clusters.

4. OUTPUT:

i) Create separate patient and customer groups.

ii) If you can, visualize the portions to aid in comprehension.

iii) Offer analysis or suggestions for each part.

III) *PARAMETERS AND JUSTIFICATION*

1. *Managing Missing Values*:

Parameter: Method to handle missing values, such as mean imputation, median imputation, or elimination.

Justification: The most effective technique will depend on the kind and quantity of missing data. The mean imputation approach is widely employed when data are randomly missing, and we may easily utilize the data when training the model rather than discarding it.

2. *Feature selection*:

Parameter: Lowest acceptable level of feature relevance, value at 0.5

Justification:The selection of characteristics reduces noise. The efficacy and readability of the model can be improved by setting a threshold to ensure that only essential features are taken into account.

3. Similarity metric:

Parameter: Measurement unit for similarity, such as the cosine similarity or the Euclidean distance.

Justification: A similarity metric is chosen based on the job and the type of data. Cosine similarity might be advantageous for text-based data.

4. Cluster quantity:

Parameter: The silhouette score should be close to 1, and the David-Bouldin score should be kept lower.

Justification: For segmentation to be effective, the appropriate number of clusters must be chosen. Techniques like the elbow approach and silhouette score may be helpful in making this decision.

V. RESULT ANALYSIS

The "thyroid0387_UCI" dataset was carefully studied during the completion of the report. For completion of A1 and A2 questions, work on both types of data was done, namely nominal and numeric data. Label encoding was performed for ordinal variables and one-hot encoding for nominal data values; categorical properties were carefully encoded into numeric values for the analysis. Missing values for several categories were filled in, and a note of outliers was made in the data.

After that, we used a suitable central tendency for filling in missing values during the data imputation phase.

The features that needed to be normalized were selected and the appropriate procedures applied to establish a standardized dataset during the data normalization/scaling phase, through which the importance of data normalization was understood.

For the A4, A5, and A6, the similarity measure analysis was done, where the cosine similarity, Jaccard method, and SMC method were applied and the results were compared.

VI. CONCLUSION

In this report, the need for analyzing legal documents for similarity is analyzed, as legal documents come in a variety of formats, and going through a lot of information in these documents can be very challenging. There is a need for knowledge management that can allow for a search engine for easy retrieval, analysis, and presentation of similar documents and paragraphs that can be easily applied to the current legal document that is being prepared. In this report, we are discussing the two popular methods used: network-based and text-based. We are touching upon a variety of methods and technologies that can be applied for similarity analysis, such as Machine Learning, Natural Language Processing, Named Entity recognition, etc. This is a field where a lot more research is required, and the tools and technology can help ease the lives of legal experts and deliver better outcomes.

VI. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the university, Amrita Vishwa Vidyapeetham, our lecturer, Dr. Peeta Basa Pati, and Ms. Roshni for assigning us the assignment and project and helping us with the insights through the project. Their knowledge of the subject assisted in the successful completion of this project report.

References

[1]https://aws.amazon.com/what-is/overfitting/#:~:text=Overfitting%20is%20an%20undesirable%20machine,on%20a%20known%20data%20set

[2]https://www.javatpoint.com/regression-vs-classification-in-machine-learning

[3]Legal case document similarity: You need both network and text - ScienceDirect

[4]Legal document similarity: a multi-criteria decision-making perspective, PMC (nih.gov)

[5]2004.12307.pdf (arxiv.org)

[6] https://ieeexplore.ieee.org/document/8859429

[7] https://ieeexplore.ieee.org/document/9225586

[8] https://ieeexplore.ieee.org/document/10054858

[9]https://link.springer.com/chapter/10.1007/978-3-031-16564-1_8

[10]https://link.springer.com/chapter/10.1007/978-981-19-0095-2_17

[11] https://ieeexplore.ieee.org/document/9716779

[12] https://ieeexplore.ieee.org/document/10010585