Similar legal documents retrieval

Aditya Kishore, Nimisha Mittal, Kruthika Manjunath, Omkar Kulkarni, Yufei Pan
University of Southern California
{adityaki, nimisham, km80974, oskulkar, yufeipan}@usc.edu

Abstract

We propose that legal document similarity can be improved by combining elements from text summarization and conventional methods.

1 Project Domain and Goal

Studying and researching over vast volumes of written data in the domain of law and judiciary cases require accurate and reliable analysis of similar cases and their judgements . Reliable retrieval system for similar cases drastically reduces the human effort and time to figure out what cases to cite and summarize. Similarity in legal cases vary with minor evidences and contexts and needs to be analysed carefully. Thus the goal of this project is to retrieve similar legal cases and summarize them to reduce human effort and research time. Parsing over text documents and comparing the similarity among them can be optimized and be made reliable with the use of Natural Language Processing algorithms and techniques. This will also drastically reduce the time for case hearings and outcomes with high level citations backed analysis of cases and presentations.

2 Related Work

There has been a lot of work done in the past to measure the similarity of legal documents and summarize them. Some of these works are examined in the section that follows.

2.1 Measure Similarity of Legal Documents

2.1.1 Citation based similarity measure R. Wagh and D. Anand (Wagh and Anand, 2017) investigated cosine similarity and citation network analysis. During this, they proposed that the concept of citation connections surpasses cosine similarity and aids in comprehending the interrelationships between various legal concepts.
Similarly, in (Bhattacharya et al., 2020) many precedent citation similarity metrics are available, including bibliographic coupling, co-citation and Node2Vec. However, S. Kumar et al. (Kumar et al., 2011) concluded that citation graphs modeled for this purpose are often quite sparse, resulting in the use of text-based similarity measures.

2.1.2 Text-based or hybrid similarity measure Researchers have developed a few methods for text-based similarity, such as paragraph linkages (Kumar et al., 2013), which evaluate two paragraphs in a document to measure similarity using TF-IDF. Full text similarity (Mandal et al., 2017) is another technique in which Doc2Vec can be used to embed the entire document and then utilize cosine similarity measurements.
Similarly, (Chavan et al., 2020) employed hypergraphs to represent citation networks and text-based similarity using Doc2Vec and cosine similarity. In this case, the citation network aids in data organization while also minimizing pair-wise comparisons.
However, in (Mandal et al., 2017) the authors evaluated that advanced semantic techniques such as neural networks and topic modeling outperform baseline techniques such as TF-IDF.

2.2 Summarizing Legal Documents

A. Kanapala et al. (Kanapala et al., 2017) and P. Bhattacharya et al. (Bhattacharya et al., 2019) have provided numerous strategies that might be utilized for either legal or non-domain documents. There are supervised, unsupervised, and citation-based techniques for legal text summarization.

The contributions of our model can be summarized as follows:

1. We propose using different vectorization techniques to generate vector representations and training Siamese LSTM to find documents that are related.
2. Summarize related documents acquired via the Transformer model.

3 Datasets

We use the US Supreme Court Database that contains information of all cases and judgements from 1946 and 2020. The dataset has 122,088 rows, each row representing a dispute with case information and a set of justice votes. It has 247 attributes for each case which can be summarized in 6 categories:

1. Identification variables (e.g., citations and docket numbers)
2. Background variables (e.g., how the Court took jurisdiction, origin and source of the case)
3. Chronological variables (e.g., the date of decision, term of Court)
4. Substantive variables (e.g., legal provisions, issues, direction of decision)
5. Outcome variables (e.g., disposition of the case, winning party)
6. Voting and opinion variables (e.g., how the individual justices voted)

The link to database that we will be using is: http://scdb.wustl.edu/data.php

4 Technical Challenges

4.1 Technical problem

We aim provide a solution for document similarity in the context of legal documents. We aim to improve upon the current baseline performances which directly employ vectorization and a similarity measure.

4.2 Difficulty of the Problem

Challenges that need to be tackled when checking for similarity of legal documents:

1. Identification of appropriate vector representation of a document.
2. Identification of correct similarity measure to be used between two documents.

4.2.1 Vector Representation Identification Documents and texts from legal domain can be converted to corresponding vectorized representation using techniques such as *TF-IDF*, *Word2Vec*, *Doc2Vec*, *BERT* model etc.
A legal document is likely to integrate multiple passages of different legal issues which may not be directly related to the current case. This can be resolved by selecting appropriate paragraphs from the document which talk only about the case being discussed.

4.2.2 Document Similarity Measure Similarity between two documents is calculated using cosine similarity between the vectorized representation of the two documents. The technical challenge will be to come up with an architecture which can encapsulate context of a document. We propose using the TextRank algorithm to check for semantically similar documents. We also aim to check for the performance of a BERT model for semantic text similarity for the legal document.

4.3 Research Avenues Beyond Coursework and Existing Research Work

We aim to try a different architecture, which leverages information from different documents having similar context. We propose using inputs from TextRank algorithm while building vectors.
Recent works (Mandal et al., 2017) try to leverage document based similarity, we aim to leverage the same and incorporate a Siamses LSTM architecture. Work cited here[1] provides a basis for comparison, with legal experts providing a similarity score for the documents being compared. We aim to use the same scores as a testing benchmark for evaluation discussed further in **section 4.4**.

4.4 Evaluation

Baseline evaluation has been generated in the cited works[1], we plan to use the same evaluation metric. We will use a *Pearson correlation coefficient*. Pearson correlation coefficient is defined as

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Where X is the similarity score provided by an expert and Y is the similarity measure we will obtain from our model.

5 Division of Labour

Task	Contributor
Idea Brainstorming	Everyone
Text Summarization	Nimisha Mittal
Dataset Exploration	Kruthika Manjunath
Feature Engineering	Aditya Kishore
Model Integration	Omkar Kulkarni
Hyper parameter tuning	Yufei Pan
Alternate Architecture Reserach	Everyone

References

Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2020. Methods for computing legal document similarity: A comparative study. *ArXiv*, abs/2004.12307.

Paheli Bhattacharya, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. 2019. A comparative study of summarization algorithms applied to legal case judgments. In *ECIR*.

Suyash Chavan, Janani Balasubramanian, Jai Puro, Meghana Naik, and Anant V. Nimkar. 2020. Similarity analysis of legal documents using content and network based approach. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2017. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Aditya Singh. 2011. Similarity analysis of legal judgments. In *Bangalore Compute Conf.*

Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Malti Suri. 2013. Finding similar legal judgements under common law system. In *DNIS*.

A. Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2017. Measuring similarity among legal court case documents. In *Compute '17*.

Rupali Sunil Wagh and Deepa Anand. 2017. Application of citation network analysis for improved similarity index estimation of legal case documents : A study. *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pages 1–5.