Measuring Similarity among Legal Court Case Documents

Abhyudaya Agrawal[a], Dr. P.J.A. Alphonse[a], Avinash Trivedi[a], Divakar Keshri[a]

[a] *Computer Applications, NIT Trichy, Tiruchirappalli, Tamil Nadu,*

Abstract

In the legal area, it is critical to determine the degree of similarity between two legal documents. It is useful for tasks such as determining which examples from the past are relevant to a current situation. A variety of methods have been proposed for doing this, such as examining the text itself or leveraging networks. Although prior techniques have suggested both text-based and network-based strategies for this task, each has drawbacks of its own. Because legal citation networks are scarce, network-based approaches frequently fail, whereas text-based approaches—which mostly use TF-IDF—have some promise but are not very sophisticated. We investigate how to improve text-based approaches for comparing legal documents' similarity. We investigate cutting edge methods such as topic modeling and neural network models like word and document embeddings, going beyond conventional TF-IDF metrics. We compare various approaches for textual similarity evaluation after conducting a thorough experimentation with a large dataset of Indian Supreme Court decisions.

I obtained 33,545 documents from the Indian Supreme Court in order to do my investigation. From 1950 to 2016, a span of 67 years, all the cases the court has heard are covered by these materials. I obtained these from the Legal Information Institute of India's website, LLLofIndia, which offers a large number of court cases in a public way. Here, I concentrated on the Word2Vec and Doc2Vec models in this work.

Keywords: Legal Information Retrieval, Legal Document Similarity, Court Cases, Topic Modeling, Word Embeddings, Word2Vec, Doc2Vec

1. Introduction

Many nations, including Australia, India, the United Kingdom, and the United States of America, use the Common Law System. This system's emphasis on prior cases is one of its main features. Stare decisis is a legal doctrine that states that if two cases are similar, then too should their outcomes be. Thus, judges consider prior decisions made in similar circumstances and make similar decisions.

Many legal documents are now accessible online thanks to technology. Due to this, it is difficult for attorneys to manually locate past cases that may be relevant to their current ones. They hence require automated systems to perform this.

Similar to internet searches, however instead of typing a few words, attorneys tell the system about their present case, and it identifies precedents that are similar to it.

The tough thing is this: legal documents frequently contain a wide range of legal topics and are lengthy and complex. This makes determining if two texts are comparable difficult for anyone lacking specialized understanding. Thus, having technologies that can perform this automatically is crucial.

Some of the techniques have been applied in the past to compare the similarity of legal documents. These techniques can be broadly divided into two categories: those that concentrate on the text itself and those that examine the connections between documents (such as those found in citations). However, these approaches have certain drawbacks. For instance, because legal networks might be sparse, those that rely on links between papers don't always function. Moreover, the text-based techniques currently in use tend to be primitive.

I so set out to devise more accurate methods of measuring the degree of resemblance across legal papers. I experimented with displaying the documents in various ways, such as viewing the entire manuscript, simply the paragraphs, summaries, or the text around citations. I then compared these representations using a variety of contemporary methods to see how comparable they were.

Converting the documents into vectors and comparing the similarities between the vectors is one popular method for doing this.

I discovered from the publications that Word2vec and Doc2vec, two neural network-based techniques, performed the best. This is the second time that similarity between legal texts has been measured using these types of models.

2. Related Work

Text-based methods and network-based methods are the two primary categories of approaches used to determine the degree of similarity between two legal documents. Some even combine the two. First, let's discuss network-based techniques. These techniques examine an interconnected system of citations within legal documents. Every document can be thought of as a point on a large map, with a line connecting them if they are cited in one another. A citation occurs, for instance, when a later case cites an earlier one as an illustration. Certain approaches tally the number of citations that two documents share. For instance, two texts may be similar if they have a large number of references in common. An additional approach examines the degree of connectivity between the common references. It's like seeing if they're all part of the same group. These strategies are effective, but there's something to consider. Citations are not always abundant in legal papers. Actually, they're usually really infrequent, with only a handful of links. This is due to the fact that attorneys and judges tend to focus on the most significant cases when discussing relevant cases. Therefore, systems that rely on text or combine text and network data are frequently more beneficial. Apart from the previously discussed network-based techniques, Kumar and colleagues developed two text-based metrics: ”all-term cosine similarity” and ”legal-term cosine similarity.” These metrics make use of TF-IDF scores, which essentially indicate the relative importance of words in a document. This is how it operates: Every document is converted into a series of numbers, where each number denotes the weight of a distinct word. The similarity between the papers is then determined by comparing these numbers. They have developed a hybrid approach that considers the citation network in addition to the text. Something known as ”paragraph-links,” or PLs, were introduced. They begin by segmenting each document into paragraphs. Subsequently, they combine every paragraph from one document with every paragraph from another. When two paragraphs are very similar, they connect those two documents. Like previously, they use TF-IDF to convert two paragraphs into numbers, which are then compared to determine how similar they are. We're basing our work on a study that used 3,866 Court Case Judgments from the Indian Supreme Court, and they discovered that this approach performed better than simply looking at citations. 1. Deciding how to portray each document: We can either utilize the full text or just the sections that address the various legal topics it discusses. 2. Examining how comparable the representations are between the documents: After deciding on each document's representation, we need a means for evaluating how similar one is to the others. Let's now discuss how we choose a legal document's representation:

• Use the Entire Document:

This is the most straightforward method. However, legal experts have taught us that a single document can address a wide range of legal issues. Thus, two texts may be mentioned together even when they only cover the same subject. We investigate alternative representations of a document in order to capture these partial similarities.

• Use Document Summary:

Summaries provide us with a document's essential information without including all of the supporting details. Therefore, we can get a better understanding of the key concepts in a document by using summaries rather than the entire text. To create summaries, we follow a method from previous research. Each sentence in the document gets a score based on how important it is, then the top 10% of sentences with the highest scores are picked to make up the summary.

3. Proposed Method

After selecting the most crucial elements from the documents under comparison, we must determine how similar they are. There are several ways to accomplish this, but one popular technique is to convert each document to a numerical list. These numbers may stand for words found in the papers, subjects they address, or even more general concepts. After we obtain these lists of numbers, we compare them using a technique known as cosine similarity. In essence, cosine similarity indicates the degree to which two lists of numbers are similar. Cosine similarity will be near to 1 if the lists are substantially similar. It will be closer to 0 if they are significantly different. So, the higher the cosine similarity, the more alike the documents are.

Preprocessing:

Before we turn text into vectors, we do some basic clean-up steps: 1. We make all letters lowercase. 2. We split the text into individual words. 3. We get rid of any words that aren't letters, except for those with hyphens, dots, or commas. 4. We remove common words like "the" or "and." 5. We shorten words to their root form, like changing "running" to "run." 6. We remove words that don't appear in at least three documents in the whole collection.

Word2Vec:

Word2Vec and Doc2Vec are tools that use a large amount of text to learn words. Every word is assigned a unique number that indicates its significance in relation to the words surrounding it. I discovered from the study that the ideal size for these numerals is 200. However, I have to move from single words to entire paragraphs of text. I so create a single large number by combining the word numbers in each part. I accomplish this by averaging the word counts, emphasizing the importance of each word. After I get these large values for two text passages, I use cosine similarity to compare them and determine how similar they are.

Doc2vec:

It's like Word2vec, but instead of just looking at words, it looks at whole pieces of text. I found that using a size of 200 works best. Just like with Word2vec, we compare the big numbers it gives us for two pieces of text using cosine similarity.

4. Experimental Setup

We undertake experiments in this part to investigate two key questions: (1) What is the best approach to represent legal documents, and (2) Which document similarity technique produces the most accurate results? We need a dataset of legal papers and a gold standard—provided by domain experts—for comparing documents in order to accomplish this.

33,545 court case records from the Indian Supreme Court covering a 67-year period from 1950 to 2016 make up the dataset we assembled. The text version of these documents was downloaded from the LIIofIndia website, which has a number of legal resources. Following headnote removal and HTML file parsing, each document was saved as a list of phrases.

We cited a previous study by Kumar et al., which offered similarity ratings for 50 pairs of Indian Supreme Court case documents, as evaluated by legal professionals, in order to provide a gold standard for document similarity. Nevertheless, because three couples' datasets had missing documents, we were only able to analyze 47 pairs' similarity ratings. These scores were on a scale from 0 to 10, with 10 denoting documents that were extremely similar and 0 representing the least similar.

We used several strategies to generate similarity scores for each of the 47 test pairs in order to assess our methodologies. We next evaluated how closely our techniques matched expert opinions by comparing these scores with the similarity scores supplied by experts. This alignment was measured using the Pearson correlation coefficient, which expresses how much of a linear link there is between two variables. Perfect positive correlation is denoted by a coefficient of 1, and perfect negative correlation is denoted by a coefficient of -1. In essence, the Pearson coefficient indicates the degree to which the similarity scores determined by our techniques correspond with the expert's scores.

5. Results

Below in the table are results:

Table 1: Similarity measures for some pairs of document, as inferred by (i) legal experts, and (ii) the best methodology discussed in this work **(Doc2vec over the whole document)**. The scores inferred by the Doc2vec method has much higher correlation with the expert scores, as compared to the baseline methodology

Case_1	Case_2	Doc2Vec Cos-Sim	Legal Score	Cos Similarity Class	LSE Class	Word2Vec Cos-Sim
1992_47.txt	1992_76.txt	0.277961	0	0	0	0.731328702
1992_76.txt	1992_182.txt	0.279482	0	0	0	0.739862618
1972_11.txt	1984_115.txt	0.121031	0	0	0	0.588155489
1969_57.txt	1980_91.txt	0.366385	0	0	0	0.870145311
1959_151.txt	1982_28.txt	0.386851	0	0	0	0.880942982
1976_200.txt	1959_151.txt	0.2381	0	0	0	0.846950946
1985_114.txt	1959_151.txt	0.33172	0	0	0	0.913255411
1966_236.txt	1967_267.txt	0.380725	0	0	0	0.92695288
1961_34.txt	1979_110.txt	0.299225	0	0	0	0.867322991
1961_34.txt	1987_37.txt	0.335277	0	0	0	0.808825368
1992_47.txt	1987_315.txt	0.384198	0	0	0	0.927543692
1971_138.txt	1992_47.txt	0.487341	0	0	0	0.926846399
1992_47.txt	1992_76.txt	0.258591	0	0	0	0.731328702
1984_115.txt	1987_315.txt	0.418575	0	0	0	0.778734271
1983_129.txt	1983_27.txt	0.642085	1	1	0	0.917416809
1979_110.txt	1953_28.txt	0.305152	2	0	0	0.882468655
1963_170.txt	1979_158.txt	0.469809	2	0	0	0.950989716
1983_27.txt	1983_37.txt	0.606873	2	1	0	0.921440929
1983_27.txt	1979_33.txt	0.538635	2	1	0	0.907486228
1984_115.txt	1981_49.txt	0.572566	2	1	0	0.895251556
1979_110.txt	1989_233.txt	0.330643	3	0	0	0.844658008
1983_129.txt	1976_176.txt	0.466026	5	0	0	0.951813215
1971_111.txt	1972_291.txt	0.41302	5	0	0	0.881878767
1990_171.txt	1988_88.txt	0.318247	5	0	0	0.914149733
1972_31.txt	1984_115.txt	0.614136	5	1	0	0.914646093
1984_118.txt	1971_336.txt	0.442697	5	0	0	0.954967592
1961_232.txt	1987_380.txt	0.470643	5	0	0	0.942422488
1964_25.txt	1955_79.txt	0.522034	5	1	0	0.959058593
1976_43.txt	1985_257.txt	0.475821	5	0	0	0.954912182
1987_154.txt	1964_144.txt	0.368014	5	0	0	0.895262643
1973_186.txt	1986_218.txt	0.366749	5	0	0	0.911014335
1990_96.txt	1990_171.txt	0.444356	5	0	0	0.906012942
1958_3.txt	1992_144.txt	0.386149	5	0	0	0.873915808
1979_158.txt	1965_111.txt	0.419282	7	0	1	0.945581841
1962_303.txt	1972_291.txt	0.537909	7	1	1	0.864502544
1987_37.txt	1989_233.txt	0.418592	7	0	1	0.85569245
1953_40.txt	1953_24.txt	0.745867	7	1	1	0.982557823
1966_154.txt	1976_43.txt	0.34037	7	0	1	0.95480329
1953_24.txt	1957_52.txt	0.271108	7	0	1	0.808632331
1984_115.txt	1971_49.txt	0.530668	7	1	1	0.909634669

	Precision	Recall	F1-score	Support
0 (Not Similar)	0.84375	0.818182	0.830769	33
1 (Similar)	0.666667	0.705882	0.685714	17
Accuracy	-	-	0.78	50
Macro Avg.	0.755208	0.762032	0.758242	50
Weighted Avg.	0.783542	0.78	0.781451	50

Table 2: Performance metrics for the classification task.

6. Conclusion

We tested a number of approaches in this research to determine how similar two legal papers are to one another. Before then, only rudimentary frequency-based methods like TF-IDF could be employed as text-based metrics for this purpose. The Doc2vec similarity throughout the entire document was shown to correlate most with the expert judgment among the similarity measures covered in this paper. This finding defies common sense in that we expected paragraph-based approaches to perform better because, according to legal experts, various paragraphs in a legal document typically address distinct legal concerns. Nonetheless, the gold standard taken into consideration in this work was produced by legal professionals who evaluated the resemblance of two complete documents. This is possibly why measuring semantic similarity between the two full documents correlates best with the similarity scores assigned by the legal experts.

References

[1] Lars Backstrom and Jon Kleinberg. 2014. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. In Proc. ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW). 831–841.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3 (March 2003), 993–1022.

[3] Stefanie Brüninghaus and Kevin D. Ashley. 2001. Improving the Representation of Legal Case Texts with Information Extraction Methods. In Proceedings of the 8th International Conference on Artificial Intelligence and Law (ICAIL '01). ACM, New York, NY, USA, 42–51. https://doi.org/10.1145/383535.383540

[4] Filippo Galgani, Paul Compton, and Achim Hoffmann. 2012. Towards Automatic Generation of Catchphrases for Legal Case Reports. Springer Berlin Heidelberg, 414–425.

[5] Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Aditya Singh. 2011. Similarity Analysis of Legal Judgments. In Proc. ACM Compute Conference. 17:1–17:4.

[6] Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, and Malti Suri. 2013. Finding Similar Legal Judgements under Common Law System. 103–116.

[7] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In Proc. International Conference on Machine Learning (ICML), Tony Jebara and Eric P. Xing (Eds.). JMLR Workshop and Conference Proceedings, 1188–1196.

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781

[9] Akshay Minocha, Navjyoti Singh, and Arjit Srivastava. 2015. Finding Relevant Indian Judgments Using Dispersion of Citation Network. In Proc. International Conference on World Wide Web (WWW) Companion. 1085–1088.

[10] Paul Zhang and Lavanya Koppaka. 2007. Semantics-based Legal Citation Network. In Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL). 123–130.