Legal Knowledge and Information Systems
M. Araszkiewicz and V. Rodríguez-Doncel (Eds.)

doi:10.3233/FAIA190307

Similarity and Relevance of Court Decisions: A Computational Study on CJEU Cases

Kody MOODLEY[a,b,1], Pedro V. HERNANDEZ SERRANO[a], Gijs VAN DIJCK[b] and Michel DUMONTIER[a]
[a] *Institute of Data Science, Maastricht University*
[b] *Faculty of Law, Maastricht University*

Abstract. Identification of relevant or similar court decisions is a core activity in legal decision making for case law researchers and practitioners. With an ever increasing body of case law, a manual analysis of court decisions can become practically impossible. As a result, some decisions are inevitably overlooked. Alternatively, network analysis may be applied to detect relevant precedents and landmark cases. Previous research suggests that citation networks of court decisions frequently provide relevant precedents and landmark cases. The advent of text similarity measures (both syntactic and semantic) has meant that potentially relevant cases can be identified without the need to manually read them. However, how close do these measures come to approximating the notion of relevance captured in the citation network? In this contribution, we explore this question by measuring the level of agreement of state-of-the-art text similarity algorithms with the citation behavior in the case citation network. For this paper, we focus on judgements by the Court of Justice of the European Union (CJEU) as published in the EUR-Lex database. Our results show that similarity of the full texts of CJEU court decisions does not closely mirror citation behaviour, there is a substantial overlap. In particular, we found syntactic measures surprisingly outperform semantic ones in approximating the citation network.

Keywords. Text Similarity, Word Embeddings, Network Analysis, CJEU

1. Introduction

Within the setting of case law, the identification and citation of relevant court decisions to support judicial decision making is a central activity. Network Analysis methodology [1,2,3,4] has proven to be useful for *a posteriori* analysis of court decision citation behavior, for example, in identifying legal precedents and measuring the influence of decisions. However, an *a priori* understanding of what constitutes a relevant case (w.r.t. to a given case) remains a complex and multifaceted question. In law generally, the concept of relevance has been previously studied and there have been attempts to define it for Legal Information Retrieval (LIR) tasks [5]. However, to date, there has been no measurable specialisation of this definition for case law.

The publishing of court decisions online as full texts in databases such as EUR-Lex (`https://eur-lex.europa.eu`) and HUDOC (`https://hudoc.echr.coe.int`), and the advancement of text similarity algorithms [6,7,8], has enabled the automatic search

[1] Corresponding Author: Institute of Data Science at Maastricht University, Universiteitssingel 60 (1st floor), 6229 ER, Maastricht, The Netherlands; E-mail: kody.moodley@maastrichtuniversity.nl

and retrieval of similar (and potentially relevant) cases. Many such measures are implemented in proprietary software such as ROSS (`https://rossintelligence.com`) and Lex Machina (`https://lexmachina.com`). The commercial success of these platforms suggest that the algorithms have promising accuracy and, therefore, text similarity may prove to be a useful tool for computationally characterising case relevance. However, there are caveats to these technologies. One is that many of these platforms do not explain *why* they found particular cases relevant, and therefore, it is difficult to measure and benchmark their legal merit. In particular, we are interested in measuring *recall* or *completeness* of these algorithms (and to a lesser extent, their *precision* or *accuracy*).

In order to establish a benchmark for completeness, we need to capture an understanding of relevance in a legal context, case law in particular. One possible strategy to achieve this is to solicit legal experts to annotate court decision texts with information (e.g. legal principles, topics and arguments) that they use to evaluate case relevance [9]. While we advocate such an approach for the longer term, there are alternatives to explore in the interim that would yield equally interesting insights with lower demand on time and resources. One of these, which we adopt in this work, is to select our base understanding for relevance to be equivalent to *citation* as captured in the *court decision citation network* (CDCN for short). Accepting this notion of relevance, we compare it to several state-of-the-art text similarity measures applied to cases from the Court of Justice of the European Union (CJEU). We use these algorithms to generate what we call a *court decision similarity network* (CDSN) - an analogue of the CDCN in which links between decisions imply high textual similarity. The graphical difference between a CDSN and a CDCN is that the edges of a CDSN are undirected, whereas those in a CDCN are directed. The goal of our study is to evaluate the size of overlap between the CDSNs generated by selected text similarity algorithms and the CDCN. Our results contribute towards an answer to the question: *to what extent can state-of-the-art text similarity measures capture the citations in the CJEU CDCN?*

The remainder of the paper is organised as follows: in Section 2, we provide an overview of related work in relevance and textual similarity of court decisions. In Section 3, we introduce the methodology of our study which includes descriptions of the selected dataset, sampling strategy and text similarity algorithms. Section 4 discusses our main findings, Section 5 outlines the caveats, limitations and challenges of the evaluation, and Section 6 summarises what we learned in the study, our plans for extending the work, and the licensing and availability of the data and software used.

2. Related Work°

In terms of efforts to define relevance for legal information retrieval, van Opijnen & Santos [5] provide a conceptual framework to categorise and define dimensions of relevance. There are six types listed: *algorithmic, topical, bibliographic, cognitive, situational* and *domain.* While this work provides a foundation for defining legal relevance, to date there has not been any mechanism proposed for *measuring* these relevance dimensions for specific legal topics.

In a separate endeavour, van Opijnen [10] has also established a model for ranking importance of case law. In this work, the author arrives at predictors for whether a case will play a marked role in future legal debate (based on its discussion in the legal community from the point of inception). Malmgren [11] also studies the notion of relevance in LIR and also within the context of CJEU decisions. However, it appears that both these efforts presume a futility in developing reliable computational algorithms for finding relevant cases, that only take the decision texts or content into account. One major reason being intrinsic subjectivity in the notion of what legal experts might consider relevant. Therefore, there are many studies that try to measure the importance and relevance of case law by means of studying the CDCN through the use of Network Analysis metrics [12,13,14]. Network Analysis was also validated as a useful way to measure relevance and importance for Dutch cases [15]. There are also many efforts to apply text similarity measures to find relevant cases in the literature. Sugathadasa et al. [16] apply deep learning to train a similarity classifier for cases from FindLaw (https://www.findlaw.com). In order to measure performance, ground truth is based on validation by legal experts. Raghav, K. [17] also provide a method to augment similarity analyses of cases based on Network Analysis with text similarity on the paragraph level. The authors found a very high agreement between citation metrics and paragraph similarity on their dataset of Indian Supreme Court judgements. Panagis et al. [18] performed an interesting study on CJEU decisions to identify what they call "implicit" citations. These are references between cases that are not explicitly stated in the cited instruments of the decision but those identifiable from the text. They use the *Tversky index* measure [19] to compare similarity of paragraphs between cases. This approach proved that the CDCN does not provide the full picture of relevant cases and provides motivation for further research into increasing recall of case retrieval.

3. Methodology

In this section we detail our methodology for constructing the CDCN and CDSNs in the study and how we calculated the size of their overlap.

Corpus selection and extraction: we selected to first study decisions by the CJEU as published in the EUR-Lex database. While we would like to extend our investigation to other case law corpora in the future, we focus on EUR-Lex initially because: 1) EUR-Lex judgements are translated into English (unlike many national case law databases), which provided our analysis team with a *lingua franca* through which to interpret and communicate the results of the text similarity algorithms, 2) While databases such as HUDOC also provide English translations of cases, EUR-Lex cases can be downloaded directly from their webpage in both XML and HTML formats which are more readily processable with software tools (as opposed to HUDOC cases available in PDF and Microsoft Word format). We extracted the full texts of all judgements and *orders* (abridged judgements) from EUR-Lex / CELLAR (the central data store of the EU publications office). We did this for all decisions until December 2018 (according to their document dates). We excluded decisions from the General Court, Civil Service Tribunal and Court of First Instance. This gave us a corpus of 13,828 decision texts in total across various topics. In addition to the full texts, we also extracted the citations (exclusively to other CJEU judgements and orders) and *subject matters* for each case, as reported in the metadata published on the EUR-Lex webpage for the case. Subject matters are keywords denoting legal topics that a case deals with (the topics are part of a classification system for EUR-Lex documents aligned with the evolution of EU policies). Details about how the extracted information is stored, published and licensed (for further research) is found in Section 6.

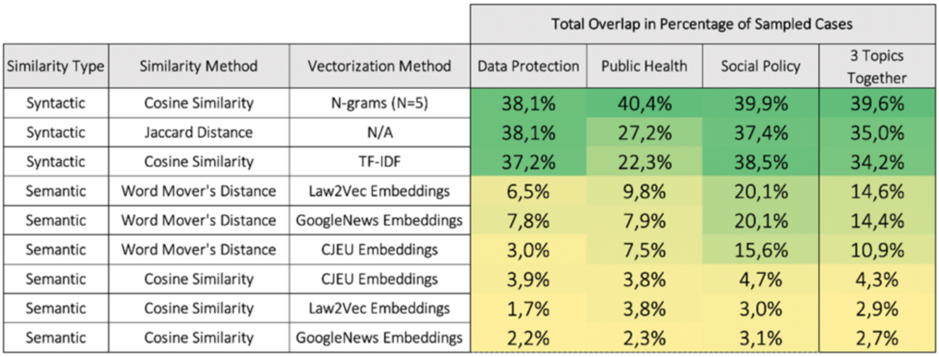
Case sampling strategy: analysing all 13,828 CJEU cases would require in the region of 95 million similarity checks for each algorithm that we evaluate (*n choose k* where *n*=13,828 and *k*=2). We therefore elected to focus on a sample subset of the CJEU CDCN. To be representative of the CJEU cases, we chose to sample variance in the citation frequency of a case (to avoid bias). For the selection of topics, there is an option to perform a similar sampling across the case topic distribution in the CJEU corpus. However, while the advantage of this approach gives us a sample that contains a broad variety of topics, it also presents a challenge. This is because we would like to generate human interpretable visualisations of the CDSNs. If we have many topics within a particular visualisation, it is more challenging to represent all of them in the CDSN while still retaining a graphical representation of the CDCN in which patterns are self-evident. Therefore, we selected three topics of cases for our evaluation based on their currently heightened societal relevance: 1) Data protection, 2) Social policy and 3) Public health. Extracting all cases concerning these topics, we had 42, 707 and 181 for data protection, social policy and public health, respectively. We calculated sampling size based on population size and margin of error. Selecting a sampling error of 10% and confidence of 95%, resulted in a sample size of 63, 85 and 29 cases for each topic, respectively. To ensure that we sample cases uniformly across citation frequency, we sorted them by number of citations. We then partitioned them into N quantiles equidistant from each other, where N is the sample size for the case topic. The cases located at each quantile then serve as the sample cases for our analysis.

Selection of text similarity measures: Text similarity algorithms generally fall into two broad categories: *syntactic* and *semantic* [8]. Syntactic measures are generally based on calculating and comparing the frequency of characters or words between texts. Semantic measures provide mechanisms to take into account context of words within the text - i.e., their neighbouring words. For this initial study, we chose to evaluate three methods in each category. For syntactic measures, we elected to evaluate *Term Frequency - Inverse Document Frequency* (TF-IDF) [20], *Jaccard distance*, and *N-grams* (N=5). For the N-grams method, we found that the overlap of similarity links and citation links in the CDCN continues to increase until N=5. Thereafter, the overlap starts to drop (hence we choose N=5). TF-IDF and N-grams provide a method for vectorising the CJEU case texts into *document vectors*. In order to measure similarity of documents, we need a vector distance measure. We elected to use the popular *cosine similarity* distance measure for these two methods. The only preprocessing applied to the texts was removal of *stop words*. The stop words removed were a combination of: 1) the set of all English language stop words available in the Natural Language Toolkit Python library (https://www.nltk.org), and 2) the set of words that occur most frequently in the case texts (those appearing in at least 90% of the documents), and 3) a selection of words and phrases which were identified by legal researchers as particular to the corpus (e.g. “Court of Justice”). For the semantic measures, we chose to implement *word embeddings* [21] as the primary means to vectorise the texts. In order to gain insight into the question of whether general or domain-specific word embeddings are more successful, we used three types: 1) a general model pre-trained on news articles - the *GoogleNews* vectors (https://code.google.com/archive/p/word2vec), 2) a more specialised model pre-trained on legal documents from the EU (including EUR-Lex) and the US, called *Law2Vec* [22], and 3) a model trained by us on all EUR-Lex judgements and orders until December 2018. We shall refer to these models in the sequel as the *GoogleNews, Law2Vec* and *CJEU* embeddings, respectively. Our CJEU embeddings were trained using the following steps: Firstly, we removed stopwords from each case in the corpus of 13,828 cases. We then used the Word2Vec model implementation offered by the Gensim (https://radimrehurek.com/gensim) Python library in order to train the word embeddings. We varied the following parameters: 1) the vector dimension size (2^n where n=[5,9]), the number of training epochs (increments of 5 from 5-50), the *window size* (increments of 5, from 5-20). Window size refers to the number of words to the left and right of a word in the text that the embedding model should consider as its "context". For vector size, we tried dimensions that are powers of 2 to speed up training time by making efficient use of memory. We found a vector size of 256, number of training epochs of 30, and window size of 5 for the CJEU embeddings provided highest overlap size with the CDCN. Hence, this is the model reported in the sequel. In terms of document distance measures, we considered two measures: cosine similarity and *word mover's distance* (WMD) [23], the latter has given state-of-the-art performance for various applications. WMD can only be calculated with word vectors and therefore cannot be used for TF-IDF and N-gram, which use *document* vectors.

Evaluation setup: in summary, we selected three syntactic text similarity measures for the evaluation: Jaccard distance, TF-IDF and N-grams (N=5), the latter two methods are applied with cosine similarity to calculate document similarity. For semantic measures, we selected three word embedding models: GoogleNews, Law2Vec and CJEU embeddings. With each of these models, we applied cosine similarity and WMD to calculate document similarity. This gives us nine methods in total for the evaluation. For each of our sample cases in each topic, we calculate the top 20 similar cases to it (according to the given method). The motivation for choosing 20 as an upper bound for the size of the similarity list is that we found 99% of CJEU judgements and orders in our corpus of 13,828 to have fewer than 21 citations (with a mean of 4.2). Computing the top 20 similar cases thus gives the algorithms the theoretical possibility to capture all the citations for 99% of the cases. While there are cases in the other 1% which have up to 55 citations, it would be computationally infeasible for us to compute the top 55 similar cases for all the sample cases, using all the algorithms. For each similarity link computed by the algorithms, we check in the CDCN (for the sample cases) if there is a citation link between these same cases. If there is a citation link, we count it as an overlap. We record the overlap counts per case, per case topic and per algorithm. The CDCN for the sample cases is defined as the subset of the full CDCN that contains only the sample cases and their *direct* citations (one link). We do not include links with a length of more than one in this initial study.

4. Results

The results of the overlap, which contribute towards the main research question of the study, are depicted in Figure 1:



			Total Overlap in Percentage of Sampled Cases			
Similarity Type	Similarity Method	Vectorization Method	Data Protection	Public Health	Social Policy	3 Topics Together
Syntactic	Cosine Similarity	N-grams (N=5)	38,1%	40,4%	39,9%	39,6%
Syntactic	Jaccard Distance	N/A	38,1%	27,2%	37,4%	35,0%
Syntactic	Cosine Similarity	TF-IDF	37,2%	22,3%	38,5%	34,2%
Semantic	Word Mover's Distance	Law2Vec Embeddings	6,5%	9,8%	20,1%	14,6%
Semantic	Word Mover's Distance	GoogleNews Embeddings	7,8%	7,9%	20,1%	14,4%
Semantic	Word Mover's Distance	CJEU Embeddings	3,0%	7,5%	15,6%	10,9%
Semantic	Cosine Similarity	CJEU Embeddings	3,9%	3,8%	4,7%	4,3%
Semantic	Cosine Similarity	Law2Vec Embeddings	1,7%	3,8%	3,0%	2,9%
Semantic	Cosine Similarity	GoogleNews Embeddings	2,2%	2,3%	3,1%	2,7%

Figure 1. Percentage overlap of the similarity links in the CDSNs with the citation links in the CDCN.

Figure 1 demonstrates that the overlap remains fairly consistent across the three topics and that we reach a 40% overlap in the best case. N-grams with N=5 proved to be the method with the largest overlap. It is a surprise that syntactic measures far outperform semantic measures. We also observed that though the semantic measures have far lower overlap with the CDCN, they do find overlaps which the syntactic measures miss. To be precise, 12% and 21% of the WMD and cosine similarity overlaps, respectively, are missed by the syntactic measures. There is also, interestingly, only an overlap of 13% between cosine similarity and WMD (see Figure 2).

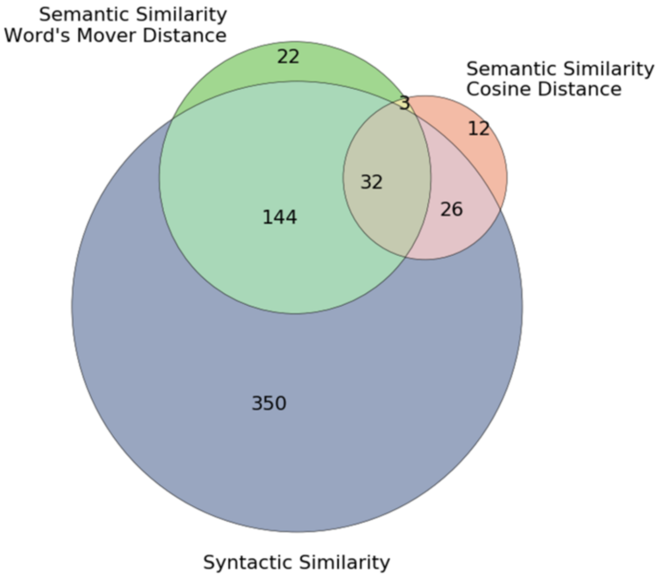


Figure 2. Venn diagram showing the degree of consensus among the algorithm categories concerning the CDSN and CDCN overlap.

It is a surprise that syntactic measures perform better because it was hypothesised that ambiguity in meaning would be an important factor in legal text. For example, the words ‘violation’ and ‘infringement’, although semantically related, are syntactically distinct.

Semantic measures would still recognise this relationship, while syntactic measures do not. The CDSNs for the three methods having highest overlap with the CDCN are plotted in Figure 3 below:

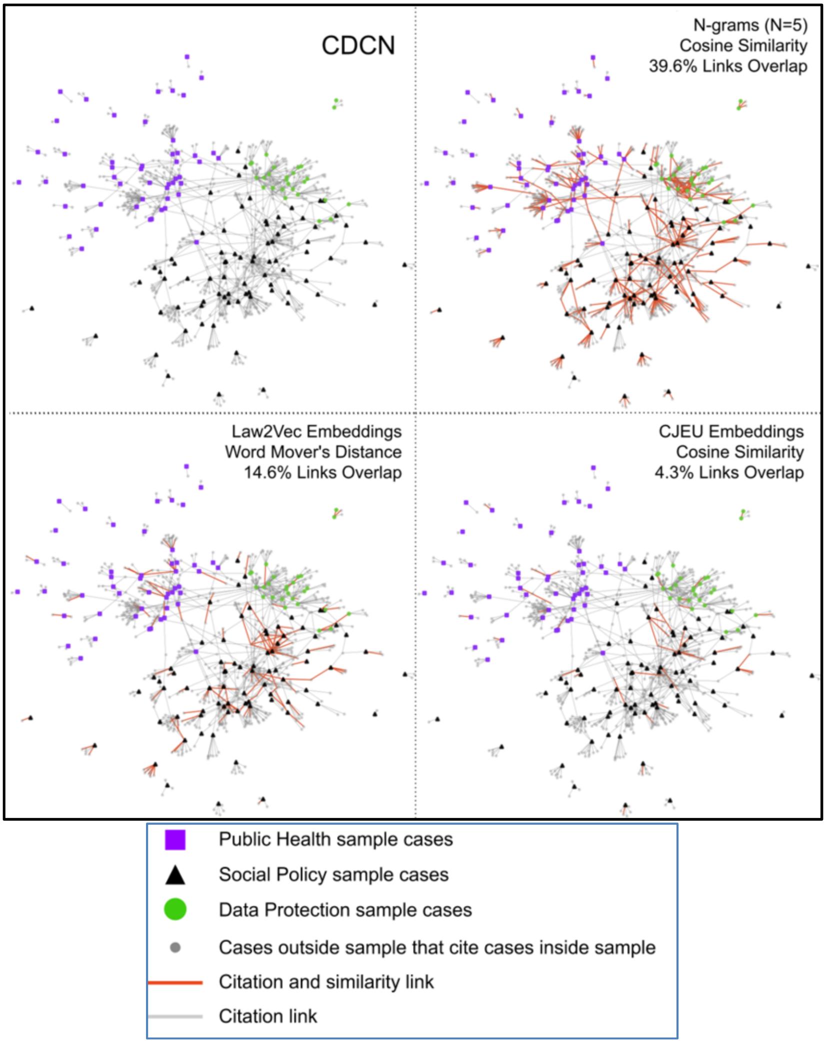


Figure 3. Visualisation of the CDSNs having the highest size of overlap with the sample cases CDCN (the best syntactic, cosine similarity and WMD methods are included).

Observing the difference between the networks for the CJEU and Law2Vec embeddings in Figure 3, we notice that there is very little improvement in overlap for the data protection cases. We also notice that the CDCN has a substantial number of cross-topic citations. It is confirmed for all methods that text similarity does not perform well at capturing these citations (most likely because the cases would be textually dissimilar, reflecting their different legal topics). Another surprising finding is that there was no significant difference in performance between general word embeddings (trained on news articles) and those trained on legal text (Law2Vec). Cosine similarity was found to be a poor measure of case similarity (in the sense that agrees with the CDCN). WMD is a substantial improvement on cosine similarity but still far behind the performance of the syntactic measures. We also performed an analysis to verify the hypothesis that, given a case A, and two cases B and C which are similar to A with almost the same degree, the one which A will cite will generally be the more cited one. We found that similar cases that are also cited have on average 2 to 8 more citations than those that are not - regardless of the similarity score. We also found that, within the top 20 similarity list for each sample case, the probability of overlap with the CDCN is highest for the 8th similar case (on average for all algorithms). If we examine the individual methods, we find one outlier - Jaccard distance - which has the highest probability of overlap with the 13th similar case. Jaccard is also the outlier in terms of variance in where the overlap lies on the similarity list. 75% of the overlaps are found in the top 18 similar cases and 50% within the top 15. The results are slightly better for cosine similarity with top 13 and top 8 respectively. However, the most reliable method was WMD with 75% of overlaps coming in the top 10 and 50% within the top 5 respectively.

5. Challenges & Limitations

One of the main limitations of the study is that we only consider three legal topics. It remains an open question about whether these results would generalise to other topics. Another caveat is that we only compare similarity links with *direct* citations from the CDCN. In general, there may be multiple indirect paths between two nodes in the CDCN, and these paths could still capture relevance between cases. Because we don't capture these links, our calculated overlap sizes (Figure 1) represent a *conservative lower bound* on the actual number of overlaps. Nevertheless, it remains unclear what the maximum length of a path should be to still capture relevance between nodes.

It also remains an open question of how close we could ever get to reconstructing the citation network (purely from the content of court decisions). Some reasons include: not all court decisions are published online; not all relevant information about a case are published in the text; while the text does provide the legal arguments, topics and principles used in the case, it will often not depict *tacit* knowledge, information about the socio-economic and political climate in which the case was decided, nor the peripheral information about the parties involved; CJEU cases are substantively different from other court decisions in that they deal with fundamental EU law. E.g. two cases about free movement of goods can be textually quite different (one could be about wine and another about electrical appliances) but they might be similar in terms of related EU legislation concerning transportation of goods.

While we do not preprocess the texts (other than elimination of stopwords), this is more of a caveat than a limitation. The reason is that we plan to arrive at a computational signature for relevance that would be maximally explainable from an intuitive standpoint. We deliberately start with a naive implementation of algorithms so that they can be incrementally optimised systematically, thereby constructing a minimum viable algorithm. Finally, we adopted *citation* as the notion of relevance. However, this overlooks other notions of relevance (e.g. where cases are substantively related but the judge forgot to include a citation between them).

6. Conclusions & Future Work

We have presented an evaluation of selected state-of-the-art text similarity algorithms w.r.t. their ability to approximate relevance as captured by the CJEU citation network. We learned that we can approximate the CJEU citation network (at least for data protection, social policy and public health cases) using these algorithms with a completeness of up to 40%, with little to no preprocessing of the texts, and optimisation of the base algorithms. We also found that syntactic measures perform three times better than semantic measures overall for this task. Surprisingly, general word embeddings (GoogleNews) performed just as well as legal text word embeddings for the same task, while cosine similarity, as a document distance measure, performed poorly. We also observed that Word Mover's Distance was the most "consistent" document distance measure overall in that 75% of its overlapping cases came from the top 10-11 of its similarity list, and half of them came from the top 5. This is in contrast to all other methods tested, which had significantly more variance in the degree of textual similarity of the overlapping cases. Unsurprisingly, we also confirmed the generally acknowledged hypothesis that the higher the citation frequency of a case, the more likely it is to be cited. This was done by comparing the citation frequency of similar cases that are involved in a citation link vs. those that are not.

Our next steps will be to extend the study to understand if the findings we obtained generalise to: 1) other legal topics for cases in the CJEU network, and 2) other court decision corpora (e.g. ECHR decisions). We also plan to evaluate additional text similarity measures (both semantic and syntactic). From the semantic perspective, the recent *siamese networks* [25] appear to be promising, as well as the *Latent Dirichlet Allocation* (LDA) and *Latent Semantic Analysis* (LSA) methods for identifying abstract topics from text. Further syntactic approaches include *Dice's coefficient* and *Manhattan distance.* Finally, in this work, we adopted the notion of relevance captured by the CDCN. However, the presumption that citation frequency and centrality in the CDCN is a necessary condition for case relevance, is questionable if consistency of decision-making is the aim. Therefore, we would like to explore other definitions of relevance in future. One possible way to define relevance is to ask legal scholars which fragments of information in a case are most important to decide relevance. This information can be made machine processable through text annotation. We hope that these studies lead us closer to more reliable *computational signatures* of relevance for court decisions.

In the interests of promoting reproducibility, we have made all the data and software used to conduct our evaluation publicly available and accessible at the following digital object identifier (DOI) - (`http://doi.org/10.17605/OSF.IO/REBQX`). It is released under the GNU General Public License (GPL) v3.0 (`https://www.gnu.org/licenses/gpl-3.0.en.html`) which allows the distribution, modification and commercial use of the resources. However, it requires that all modifications made should be clearly stated, all source code for resulting works should be disclosed, and these works should also be released under the same license. The FAIR principles for data management [24] also advocate the interoperability and reusability of digital resources. Towards this, we have tried to document the resources we have produced in a manner that enables easier reproducibility of the study. We have used widely supported, platform-independent, data formats (CSV) and software standards (Python language with required libraries documented).

Jupyter Notebooks (`https://jupyter.org`) are also used to enable inline documentation, plots and segmented running of code.

Acknowledgements. The authors would like to thank Seun Adekunle and Andreea Grigoriu of Maastricht University for helpful discussions and inputs to the methods.

7. References

[1] J. Fowler et al. (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis*, 15(3), 324-346.

[2] Y. Lupu and E. Voeten (2012). Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights. *British Journal of Political Science*, 42(2), 413-439.

[3] M. Derlén et al. (2014). Goodbye Van Gend En Loos, Hello Bosman? Using Network Analysis to Measure the Importance of Individual CJEU Judgments. *European Law Journal*, 20(5), 667-687.

[4] D. van Kuppevelt and G. van Dijck (2017). Answering Legal Research Questions About Dutch Case Law with Network Analysis and Visualization. In *JURIX*, 302, 95-100, IOS Press.

[5] M. van Opijnen and C. Santos (2017). On the Concept of Relevance in Legal Information Retrieval. *Artificial Intelligence and Law*, 25(1), 65-87.

[6] R. Mihalcea, C. Corley and C. Strapparava (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *AAAI*, 775-780, AAAI Press.

[7] A. Islam and D. Inkpen (2008). Semantic Text Similarity Using Corpus-Based Word Similarity and String Similarity. *ACM Transactions on Knowledge Discovery From Data*, 2(2), 10.

[8] W. H. Gomaa and A. F. Aly (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13), 13-18.

[9] O. Shulayeva, A. Siddharthan and A. Wyner (2017). Recognizing Cited Facts and Principles in Legal Judgements. *Artificial Intelligence and Law*, 25(1), 107-126.

[10] M. van Opijnen (2013). A Model for Automated Rating of Case Law. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, 140-149, ACM.

[11] S. Malmgren (2011). Towards a Theory of Jurisprudential Relevance Ranking. Using Link Analysis on EU Case Law, Graduate thesis, Stockholm University.

[12] A. Geist (2009). Using Citation Analysis Techniques For Computer-Assisted Legal Research In Continental Jurisdictions, Graduate thesis, University of Edinburgh.

[13] M. van Opijnen (2012). Citation Analysis and Beyond: in Search of Indicators Measuring Case Law Importance. In *JURIX*, 250, 95-104, IOS Press.

[14] A. Minocha, N. Singh and A. Srivastava (2015). Finding Relevant Indian Judgments Using Dispersion of Citation Network. In *WWW*, 1085-1088, ACM.

[15] R. Winkels et al. (2011). Determining Authority of Dutch Case Law. In *JURIX*, 235, 103-112, IOS Press.

[16] K. Sugathadasa et al. (2018). Legal Document Retrieval Using Document Vector Embeddings and Deep Learning. In *Science and Information Conference*, 160-175, Springer.

[17] K. Raghav, P. K. Reddy and V. B. Reddy (2016). Analyzing the Extraction of Relevant Legal Judgments Using Paragraph-level and Citation Information. *AI4JCArtificial Intelligence for Justice*, 30.

[18] Y. Panagis et al. (2017). Giving Every Case Its (Legal) Due. The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union Law. In *JURIX*, 302, 59-68, IOS Press.

[19] A. Tversky (1977). Features of Similarity. *Psychological Review*, 84(4), 327.

[20] J. Ramos (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 242, 133-142.

[21] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3, 1137–1155.

[22] I. Chalkidis and D. Kampas (2019). Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora. *Artificial Intelligence and Law*, 27(2), 171-198.

[23] M. Kusner, Y. Sun, N. Kolkin and K. Weinberger (2015). From Word Embeddings to Document Distances. In *International Conference on Machine Learning*, 957-966.

[24] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne and J. Bouwman (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data*, 3.

[25] J. Mueller and A. Thyagarajan (2016). Siamese Recurrent Architectures for Learning Sentence Similarity. In *AAAI*, 2786-2792, AAAI Press.