Research note: Measuring the evolution of CJEU case law using word embeddings

Silje Synnøve Lyder Hermansen

November 20, 2019

To study the conditions for judicial activism – self restraint, initiation or expansion of case-law – we need a consistent and comparable measure of court output. Can machine learning help us in this respect? This research note discusses ways of comparing court documents in order to obtain such measures. The focus is on word embeddings which allow researchers to identify similar – but not identical – terms and use them as building blocks for document-level comparisons.

Analyzing the evolution of case law requires researchers to read and compare judgments. It is a resource-intensive process both in terms of competence and number of working hours. Assessing the outcome of a case typically requires familiarity with legal precedent in the court, the doctrines underpinning rulings as well at the type of cases brought before the tribunal. In other words, – depending on the task – the researcher needs a good grasp of similar cases in the past (i.e. legal precedent) and/or the future (i.e. its value as a precedent) as well as the field-specific vocabulary.

The purpose of this resesarch note is to train a model to make such comparisons for us. To obtain word embeddings, I apply a word2vec algorithm on 30999 (machine-readable) legal texts produced by the members of the Court

of Justice of the European Union (CJEU) itself. I use these to construct a domain specific vocabulary of similar words. The similarities can then be used to identify and compare relevant documents based on their word occurences.

In the following, I give a brief overview the word2vec method and how I apply it. I then explore the results at the word-level before proceding to a verification of the model's validity through different document comparisons. I check the model's in-sample predictions, benchmark against human coders and common covariates of case law expansion.

The promise of word embeddings

While machines can readily flag identical text occurences, more complex models are required when prospecting for words, phrases or longer text chuncks that convey the same message but with a different wording. Word embeddings is a set of natural language processing (NPL) models that seek to relate word occurences with their neighbouring expressions. The underlying idea is that words in similar contexts carry similar meanings (Harris, 1954). The purpose is to create an algorithm able to identify words and convey information about their substantial meaning as well as semantic relationships.

More specifically, these models project words into a multidimensional vector space. Conceptually, they reduce the number of vectors from a high-dimensional space in which each unique word represents its own vector (a 'one-hot' encoding) to substantially fewer dimensions. Following such a mapping, all words can be located geometrically in relation to each other. Word similarity can conveniently be expressed as a cosine distance, for example.

Frequency-based models (such as Latent Semantic Indexing (Deerwester et al., 1990), Latent Semantic Analysis (Bellegarda and Member, 1998) or Latent Dirichlet Allocation (Blei et al., 2003)) have become increasingly popular approaches to document-level comparisons. However, these models are often computer intensive and/or fail to identify similar documents without

overlap in vocabulary. More recently, prediction-based models have proven both computationally more efficient and surprisingly apt at describing semantic and linguistic relations (i.e.: Word2Vec (Mikolov et al, 2013b) and Global Vectors/GloVe (Pennington et al., 2014)).

Among these, are the `word2vec` algorithms that rely on neural networks for estimation (Mikolov et al, 2013a; Mikolov et al, 2013b; Mikolov et al, 2013c). In 2013, the authors demonstrated the model's ability to perform analogical reasoning using simple vector arithmetics. Thus, their models are able to fill in the last component in syntactic analogies such as "quick"/"quickly" and "slow"/"slowly", but also linguistic relations such as "Germany"/"Berlin" and "France"/"Paris".

The reasons (and therefore the limitations) for the higher performance of these models are still poorly understood, however. One clear advantage is their computational efficiency. Researchers can train models on larger corpora of data using their desktop computer despite their relative complexity (Mikolov et al, 2013a). The initial results were, for example, based on a training set of 30 billion words (Mikolov et al, 2013b). The trained model has later been released for further use by researchers. However, there is a trade-off between the size of a corpus and its accuracy. In the following, I train models directly on texts produced by the CJEU for the purpose of constructing a vocabulary specific to that court.

Several approaches have been suggested to aggregate word vectors to document-level comparisons. Some applications build on word vectors from pretrained models to perform more traditional word comparisons (e.g.: Kusner et al., 2015, using Word Mover's Distance or Tsurel et al., 2017, using tf-idf). Other applications – such as the `doc2vec` method – proposes a simple extension of the word2vec model by adding document labels to the model and estimate them as additional parameters (Le and Mikolov, 2014). Empirical assessments have suggested that this approach performs particularly well when applied to longer documents (Lau and Baldwin, 2016). In the

following, I rely on results from an implementation of the doc2vec method to compare court documents.

Application to CJEU documents

The collection of texts was done using the `bs4` package in Python, while I relied on `gensim` for preprocessing and text analysis (Řehůřek and Sojka, 2010).

The corpus

To build a corpus, I scraped all available court judgments (13829 documents) and advocate general opinions (6037 documents) from the EUR-Lex website. Judgments and opinions furthermore include all parts of the original text: their original title, preamble, summary and decision on costs as well as the grounds of judgment. The size of the training corpus matters. Word2vec has been reported to perform substantially worse on smaller corpora. One study finds, for example, that word2vec only outperforms Latent Semantic Analysis when the corpus size exeeds a million words (Altszyler et al., 2017). In comparison, the present model is run on a corpus of approximately 110 million tokens with a vocabulary of 39784 unique words and phrases.

Word2vec models require very little data cleaning. For the purpose of this study, each document was simply converted to a list of lowercase tokens. Keeping texts intact have the advantage that results are less dependent on text pre-processing choices. This is a particular risk for unsupervised text models (Denny and Spirling, 2018). However, the additional noise of an overly complex vocabulary may also reduce the model performance, especially when training corpora are small to medium-sized. At this (early) stage of the analysis, I have prioritized consistency over performance. An alternative approach may be to lemmatize the text, as I am not interested in syntactic relationships as such.

Rather than conventional text pre-processing, the architects behind word2vec suggest simplifying the vocabulary by identifying common phrases through training an additional model prior to the tokenization. The model identifies bigrams in the corpus that appear frequently together but infrequently in other contexts (Mikolov et al, 2013b). In `gensim`, I applied the `Phrases()` function to the CJEU corpus and stored the tokens separately. The final skip-gram model is thus run on CJEU-specific tokens such as "european_union", "direct_effect", "fundamental_right" etc.

Choice of model

Word2vec models come in two versions: The continuous bag of words (CBOW) model predicts the probability of a word occurence from a window of context words, while the skip-gram model predicts the probability of context words given a word occurence. Although the former is faster to train, the latter performs better at predicting rare words. The skip-gram also gives additional weight to nearby words, while the word order is not taken into account using CBOW estimation. Here, I have opted for a skip-gram model with a vector size of 200 estimated on a window of 10 neighbouring words.

The architects behind the model point to several elements that contribute to both the efficiency and accuracy of its implementation (Mikolov et al, 2013b). Relying on negative sampling for estimation means that the model seeks to distinguish a target word from draws from a noise distribution using a simple logistic regression. The authors suggest that the number of negative samples should be larger for smaller data sets (5-20) (Mikolov et al, 2013b, p. 4). I have specified a negative sample size of 15.

The imbalance between rare and extremely frequent words also poses a challenge. I address this by pruning rare words while simultaneously sub-sampling the most frequent occurences. Frequent words contribute with little information to the overall estimation of word embeddings, while their omnipresence slows down the learning rate (Mikolov et al, 2013b, p. 4-5). In the

following, I subsample relatively agressively so that each word is discarded with a set probability defined by the word frequency and a user-specified parameter: $\frac{0.0005}{word\ frequency}$.

Word vectors are merely the building blocks in my attemt to compare documents. The model that I estimate is therefore a doc2vec function. Each document is tagged with EUR-Lex' identification number for all official EU documents (i.e.: the celex number). In this way, words and documents are mapped on to the same vector space.

The doc2vec models also come in two variations: The "distributed memory' (PV-DM) predicts the document's words from the tag, while the "distributed bag-of-words" (PV-DBOW) predicts documents from words. The distributed memory version is more complex (Le and Mikolov, 2014), but also less stable in its implementation. Moreover, in empirical assessments the simpler PV-DBOW implementation has outperformed the more complex estimation strategy (Lau and Baldwin, 2016). Here, I have opted for the distributed bag-of-words model. The learning rate is higher for the first documents in a pass. To ensure consistency, the order of the documents is thus shuffled. I then train the model over 100 iterations.

Robustness

I would like to make a more systematic assessment of the model's performance in future iterations. All suggestions on how to benchmark are welcome!

Following the training, all words and documents in the corpus can be mapped on to the same vector space, and cosine similarities can be caluclated. The cosine distance varies from 0 (no similarity) to 1 (completely similar).

Using these similarities, I start by a short assessment of the model's ability to infer its own word vectors before making a brief assessment of similari-

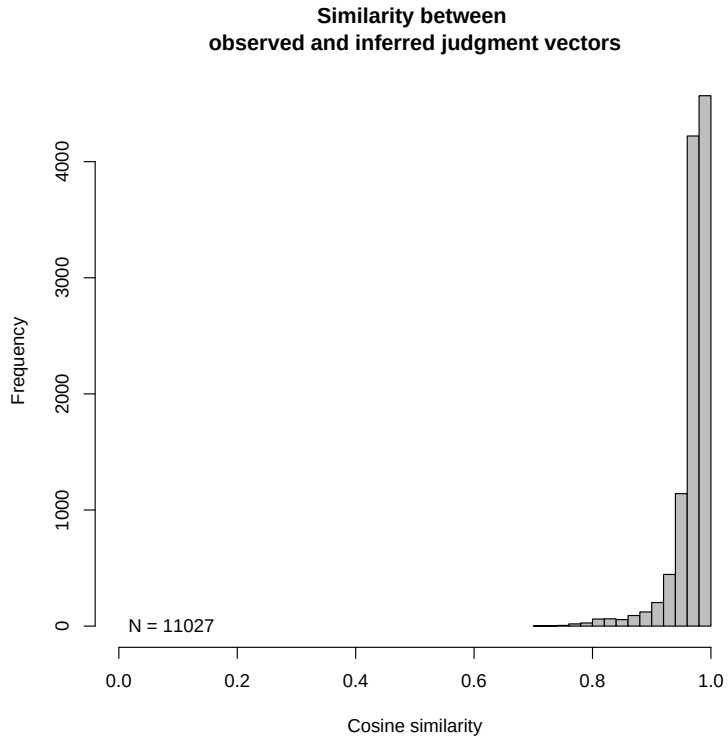

Figure 1: Similarity between document vectors calculated by the model and those inferred from the same model.

ties at the word level.[1] In the following section, I then proceed to a more qualitative assessment of document similarities using contextual variables.

In-sample prediction

The trained model can be used to infer vectors for new documents. In my first appraisal, I do an in-sample assessment of the similarity between vectors calculated by the model and those inferred from it. Ideally, these two would

[1]For the purpose of this paper, I estimated word vectors using skip-grams in parallel to the estimation of the document vectors. A faster implementation of the doc2vec is to let the model erase word weights during estimation of the model so that only document vectors remain.

be identical.

As is apparent from the histogram in figure 1, most inferred values for judgment texts tend to be relatively similar. Yet, there is room for improvement. Some 99% of the observations receive a similarity beyond 0.8, while 96% pass the 0.9 threshold.

Exploration of word vectors

Word similaritites are a function of the way words are used in the corpus. They provide the model with a vocabulary specific to CJEU jurisprudence.

Most similar words When quiering after the 11 most similar words to `january`, we obtain the following candidates with top scores:

```
[('july', 0.9058223366737366),
 ('february', 0.8992304801940918),
 ('april', 0.8952011466026306),
 ('december', 0.8948855400085449),
 ('march', 0.8915542364120483),
 ('october', 0.8884017467498779),
 ('september', 0.8734666705131531),
 ('november', 0.8719019889831543),
 ('june', 0.8510774374008179),
 ('august', 0.8459392189979553),
 ('dated', 0.6240531206130981)]
```

Each document is dated, providing ample training grounds for months to the model. The absence of the month of May also illustrates the limitations of the D-BOW method, however. As is apparent from table 1, May is indeed classified among other modal/auxiliary verbs.

Table 1 also reports other examples from quiry terms more specific to the CJEU. Together with the name of the current President Lenaerts are listed other members of the court with whom he has deliberated in the past.

Similarly, we can quire after different quasi-constitutional concepts developped by the CJEU. The concept of `useful_effect` (in French, *effet utile*) is often related to a functional approach to interpretation postulating that legal provisions must be interpreted in a way that ensures their "effectiveness". That is, national legislation must not have the effect of depriving European acts of their meaning (Bengoetxea, 1993; Saurugger and Terpan, 2017).

```
[('deprive', 0.6135631799697876),
 ('deprived', 0.607200026512146),
 ('meaningless', 0.5553340911865234),
 ('effet_utile', 0.5551632046699524),
 ('rendered_ineffective', 0.5520445108413696),
 ('depriving', 0.539157509803772),
 ('redundant', 0.5198081135749817),
 ('raison_etre', 0.5181915760040283),
 ('undermined', 0.5179909467697144),
 ('tantamount', 0.5120956301689148)]
```

Analogies We can also test out the analogical capacities of the model. On the syntactic task "may"+"might"-"can" the models' first suggestion is:

```
[('could', 0.6113986968994141)]
```

The model also provides an accurate suggestion for the linguistic analogy "woman"+"wife"-"man":

```
[('husband', 0.6688438653945923)]
```

However, for more abstract and domain-specific tasks such as asking whether the "BVerfG" (German Federal Constitutional Court) has the same relationship with the German government as "Høyesteret" has to the Danish government, we have to go further down the list. We do, however, get the idea that courts and governments are related:

```
[('government', 0.5091025829315186),
 ('french_government', 0.46145492792129517),
 ('austrian_government', 0.45597782731056213),
 ('italian_government', 0.42836248874664307),
 ('federal_republic', 0.4280872344970703),
 ('belgian_government', 0.4223637580871582),
 ('polish_government', 0.40984803438186646),
 ('czech', 0.4029742479324341),
 ('danish_government', 0.4018552303314209)]
```

National courts can submit questions for preliminary rulings to the CJEU. This procedure also allows for national governments to submit amicus curia briefs ("observations") that are listed in the text. This is (presumably) the information that the model uses.

Similarly, the model has reasonable suggestions as to the denommination of the parties to a conflict: "plaintiff"+"defendant"-"applicant":

```
[('plaintiffs', 0.56504225730896),
 ('defendants', 0.5148352384567261),
 ('respondent', 0.4745844900608063),
 ('claimant', 0.4596676230430603)]
```

Beyond the syntactic suggestions, the model also suggests that the counterpart to a "defendant" in an appeal cases is a "respondent".

may	lenaerts	direct_effect	useful_effect	fundamental_right
can	prechal	direct_applicability	deprive	fundamental_rights
might	bonichot	horizontal_direct_effect	deprived	charter
could	von_danwitz	individuals	meaningless	enshrined
cannot	lenaerts_president	relied_upon	effet_utile	fair_trial
will	mengozzi	unconditional	rendered_ineffective	private_life
allows	azizi	disapplied	depriving	effective_judicial
shall	rosas	primacy	redundant	echr
authorises	bellamy	courts	raison_etre	family_life
must	dehousse	unimplemented	undermined	inviolability
allowing	edward	sufficiently_precise	tantamount	fair_hearing

Table 1: Quiry terms and 10 most similar words.

Exploration of document vectors

With the trained doc2vec model in hand, I then calculate similaritites between documents in the corpus. I consider to types of model assessments. First, I benchmark the results against data coded by humans. Second, I consider how document similarities correlate with known covariates to case-law evolution.

Comparison with human coders

One way of assessing whether the doc2vec similarities capture meaningful differences is to compare against human coders. When investigating whether the threat of override affects CJEU decision-making, Larsson and Naurin (2016) compiled a data set including human-coded positions of all relevant actors in preliminary reference cases brought before the Court of Justice (1997-2008). Each case was subdivided into issue areas and both the Court's position in the judgment and that of the Advocate General were coded. For the purpose of this study, I have only retained cases where there is complete agreement/disagreement between the two actors. The final data set therefore contains 964 document pairs.

I begin by calculating the similarity between the text of the judgment and that of the opinion. As is apparent from the histogram in figure 2, the spread in similarities between the two texts is moderate. Yet, the plots in figure 3 illustrate that the similarity scores can nevertheless inform a logit model predicting the human coders' decision. The likelihood that coders identify a complete agreement between the advocate general and the court judgment increases 2.1 times when similarity increases by 0.1.

Notwithstanding these promising results, the model also illustrates some of the limitations to the approach. In particular, it is unclear what the similarity reflects. While I have 200 document vectors, their similarity is unidimensional. Thus, it is unclear whether the score reflects development

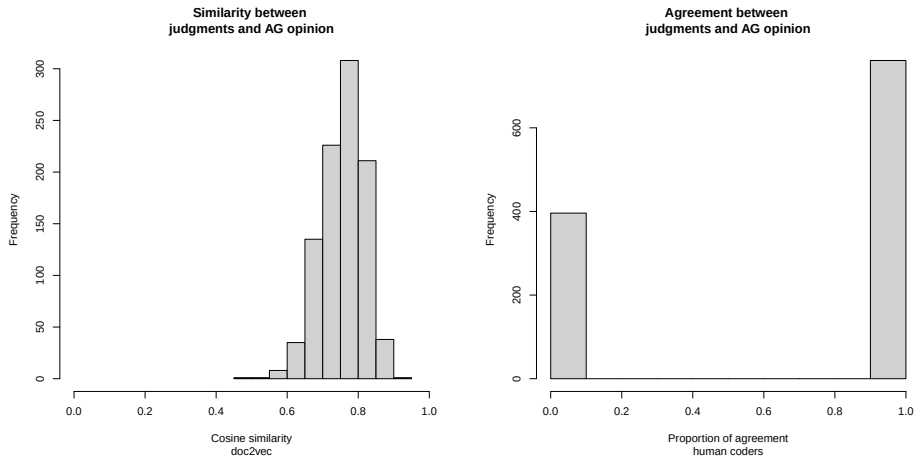

Figure 2: Cosine similarity from doc2vec and human-coded proportion of agreement between CJEU judgments and the advocate general's opinion in preliminary reference cases.

Table 2: Correlation between human coded and machine coded data

	Dependent variable:
	AG-court agreement (human-coded)
Text similarity (machine-coded)	7.552^{***}
	(1.294)
Constant	-4.396^{***}
	(0.964)
Observations	964
Log Likelihood	−491.457
Akaike Inf. Crit.	986.915
Note:	*p<0.1; **p<0.05; ***p<0.01

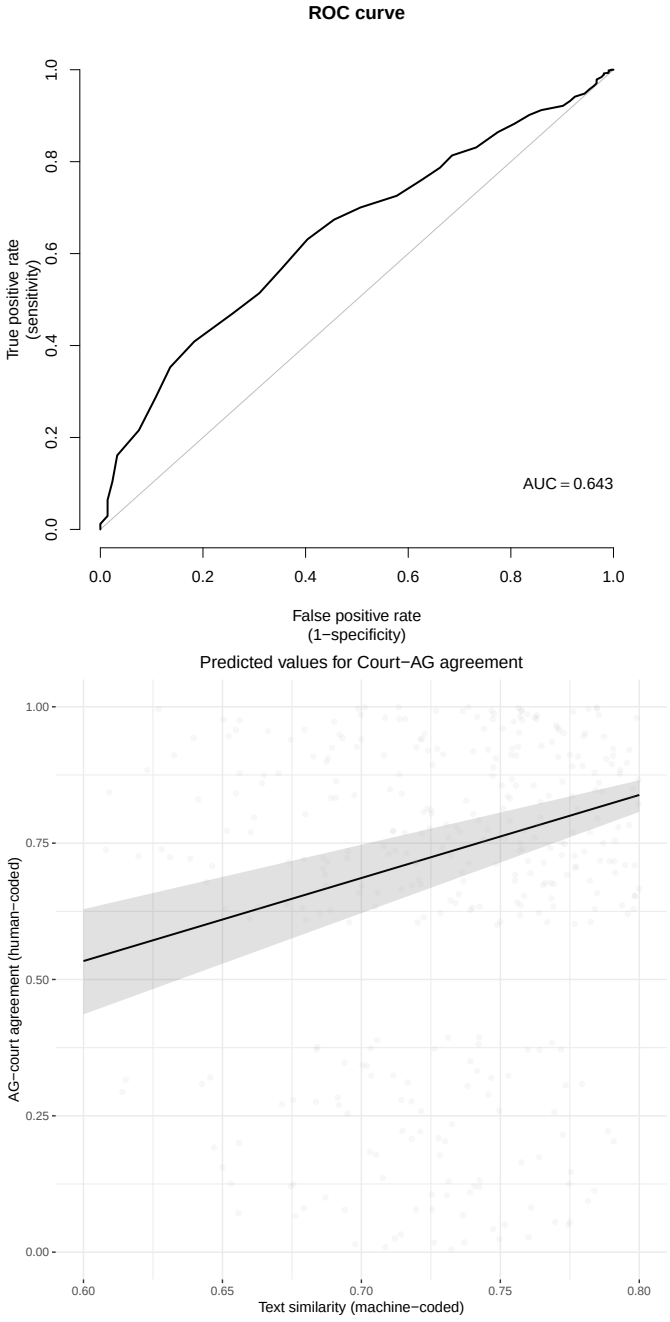

Figure 3: The machine-coded similarity between documents provides some information as to the degree of agreement between the Court and the Advocate General decided by human coders.

in the Court's legal doctrine, an expansion of its scope or other elements. In this case, the two variables are constructed in substantially different ways. While the human coders were instructed to code a binary agreement on the substance of the question, the machine-calculated similarity assesses two entier texts where agreement is merely a biproduct of the way in which facts and arguments are presented. Furthermore, as writing an opinion is a different excercise than writing a judgment, the advocate general's text is never entirely similar to the final judgment.

The moderate spread in the score also poses a challenge to the use of document similarities.

Similarity with previous cases

In a second attempt at validating whether doc2vec models can usefully describe text similarities, I identify and compare judgments from the institution's highest formation (i.e.: the Court of Justice) with the 10 most similar preceeding cases brought before the same Court. I the following, I therefore rely on a data set with 110345 observations of similarities between 11039 cases and their precedents. Figure 4 illustrates the distribution of similaritites. Once again, we see that the overall spread is moderate.

The idea behind the comparison is that cases which carry little similarity with preceeding judgments are attempts at expanding CJEU case-law. If doc2vec can meaningfully capture such attempts, I would also expect that the variable correlates with other elements related to case-law expansion. I start by providing bivariate statistics before proceeding to multivariate analyses.

Bivariate statistics

Higher-impact Court decisions tend to be identifiable by several characteristics. Complex cases tend to take longer time to process. The political salience

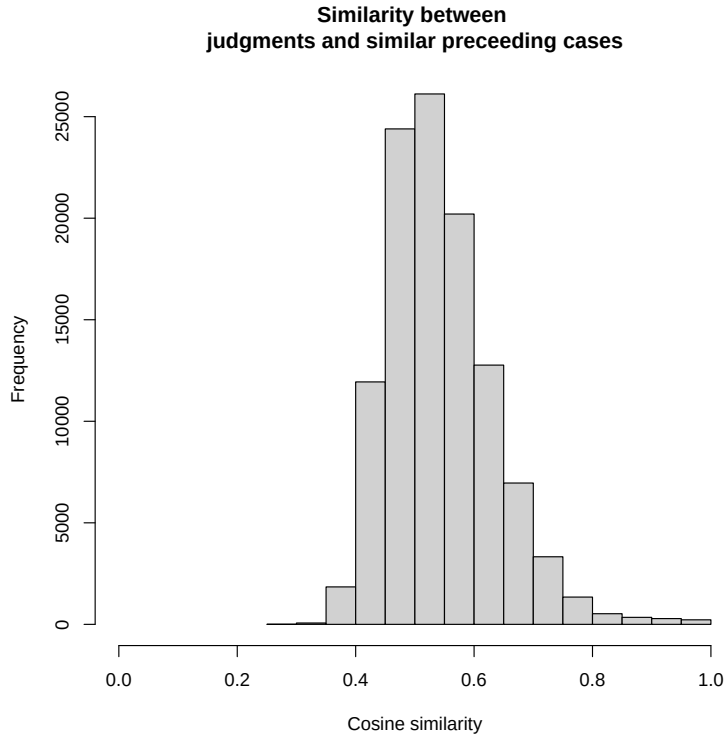

Figure 4: Cosine similarity from doc2vec estimation of all Court of Justice cases.

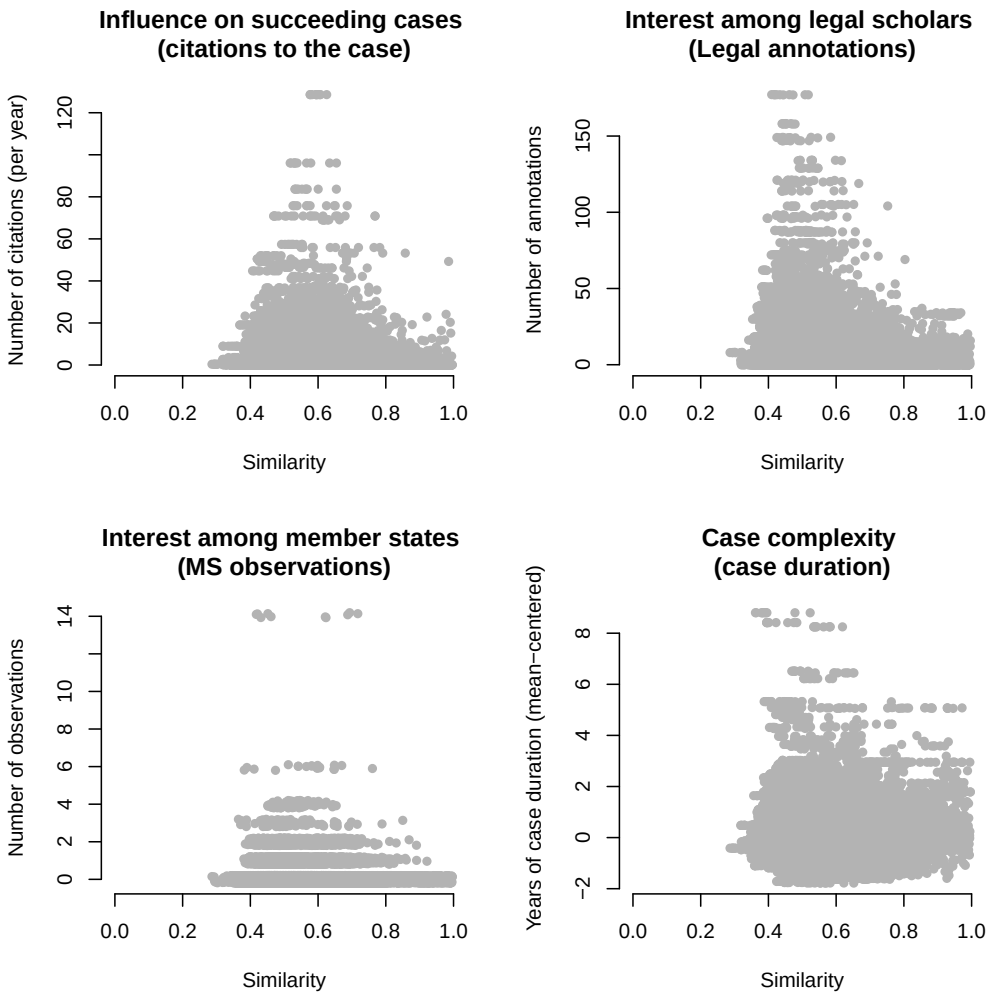

Figure 5: Bivariate relations between case similarity and factors capturing case-law expansion.

of preliminary reference cases is also reflected in the number of member states that submit observations. After the judgment, importance is reflected both by the interest it attracts in academic journals (i.e.: annotated cases) and by its value as a legal precedence (i.e.: the number of citations in later judgments).

The scatterplots in figure 5 already provide some indication to that effect. However, while the pairs of "case-preceding case" that obtain lower similarity are more likely to have higher values of interest (i.e.: high values on the y-axis), many low-similarity pairs show no such attributes (i.e.: low values on the y-axis).

Predictors of text similarity

There may be other reasons why cases are similar. In the regression models reported in table 3, I account for some of them. Thus, we see that – unsurprisingly – similarity is higher for preceding cases that are cited than those that are not. Similarity furthermore decreases as the time lapse between two cases increases, while judgments penned by the same judge-rapporteur tend to be more similar.

The first model in the table is fitted with random intercepts per year to account for variation over time in document similarity. Their point estimates are illustrated in figure 6. Here we can see that case similarity varied extensively from year to year during the first two decades of the Court's existence. In this period, the Court's case load was low, and several of the landmark judgments were rendered in the 1960-ies. The more notable part of the graphic shows the decrease in similarity around the Nice treaty (2003) and the subsequent EU-enlargments.

The second model in table 3 explores the effect of chamber size on case similarity. It stands to reason that larger chamber formations treat with higher-salience cases where the Court does not expect extant case-law to provide clear answers. In this model, the year intercepts are replaced by

Table 3: Text similarity as a function of case-related predictors

	Similarity with 10 preceeding cases (OLS)			
	Yearly rand. int.	Case-level rand. int.	Case-level rand. int.	Case-level rand. int.
	(1)	(2)	(3)	(4)
chamber_typeChamber of 3		0.042***		
		(0.002)		
chamber_typeChamber of 5 (<2003)		0.022***		
		(0.002)		
chamber_typeChamber of 5 (>2003)		0.023***		
		(0.002)		
chamber_typeFull court (>2003)		−0.016		
		(0.014)		
chamber_typeGrand chamber (>2003)		0.009***		
		(0.003)		
chamber_typePetit plenum (<2003)		0.012***		
		(0.003)		
case_duration.c			−0.006***	
			(0.001)	
N.observations.prop				−0.017***
				(0.002)
case_cited	0.059***	0.049***	0.049***	0.052***
	(0.001)	(0.001)	(0.0005)	(0.001)
time_between	−0.002***	−0.002***	−0.002***	−0.001***
	(0.0001)	(0.00004)	(0.00004)	(0.0001)
same_rapporteur	0.054***	0.050***	0.050***	0.039***
	(0.001)	(0.001)	(0.001)	(0.001)
same_rapporteur_nationality	−0.007***	−0.002***	−0.002***	−0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
Constant	0.537***	0.515***	0.537***	0.537***
	(0.003)	(0.001)	(0.001)	(0.001)
Observations	107,185	106,585	107,175	38,022
Log Likelihood	115,222.800	147,874.300	148,541.100	55,914.530
Akaike Inf. Crit.	−230,431.700	−295,722.600	−297,066.200	−111,813.100
Bayesian Inf. Crit.	−230,364.600	−295,598.100	−296,989.500	−111,744.700

Note: *p<0.1; **p<0.05; ***p<0.01

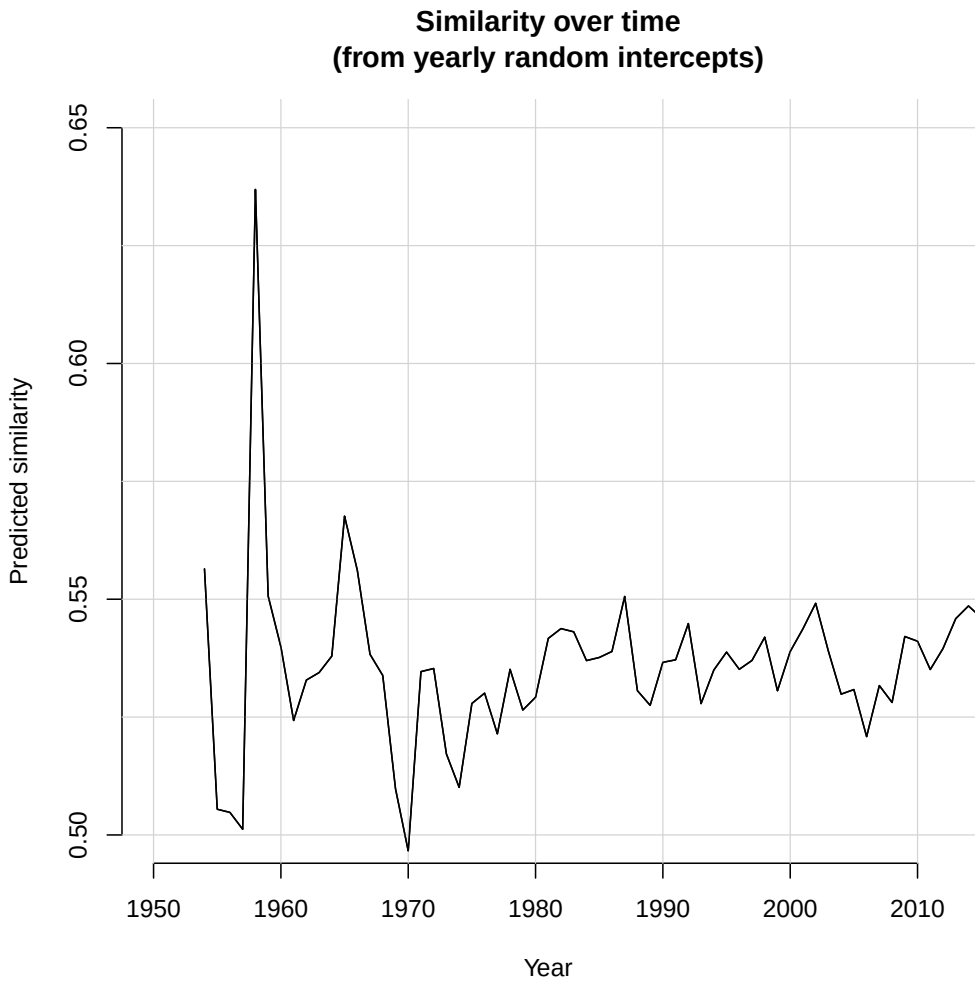

Figure 6: Variation in similarity over time when pairwise similarities are accounted for. Illustration of random intercepts.

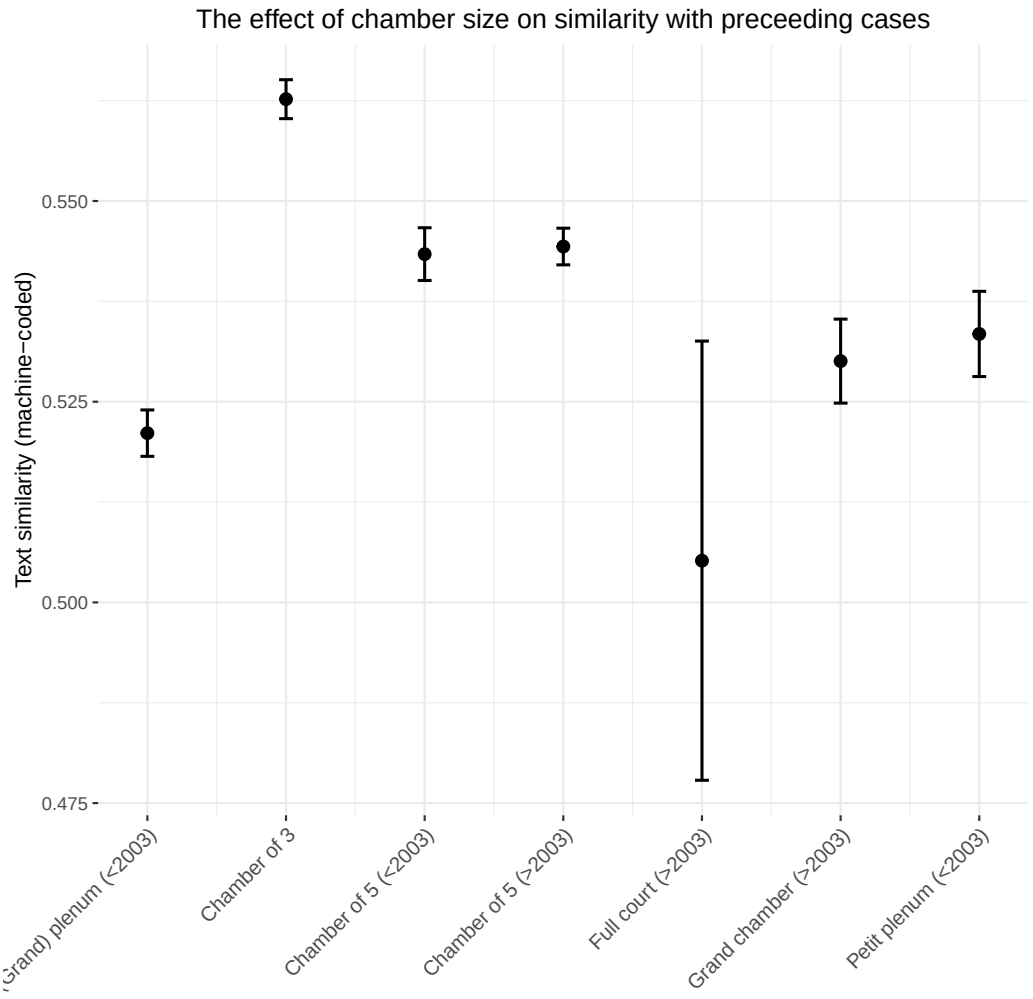

Figure 7: The larger chamber formations in the CJEU tend to render less similar judgments.

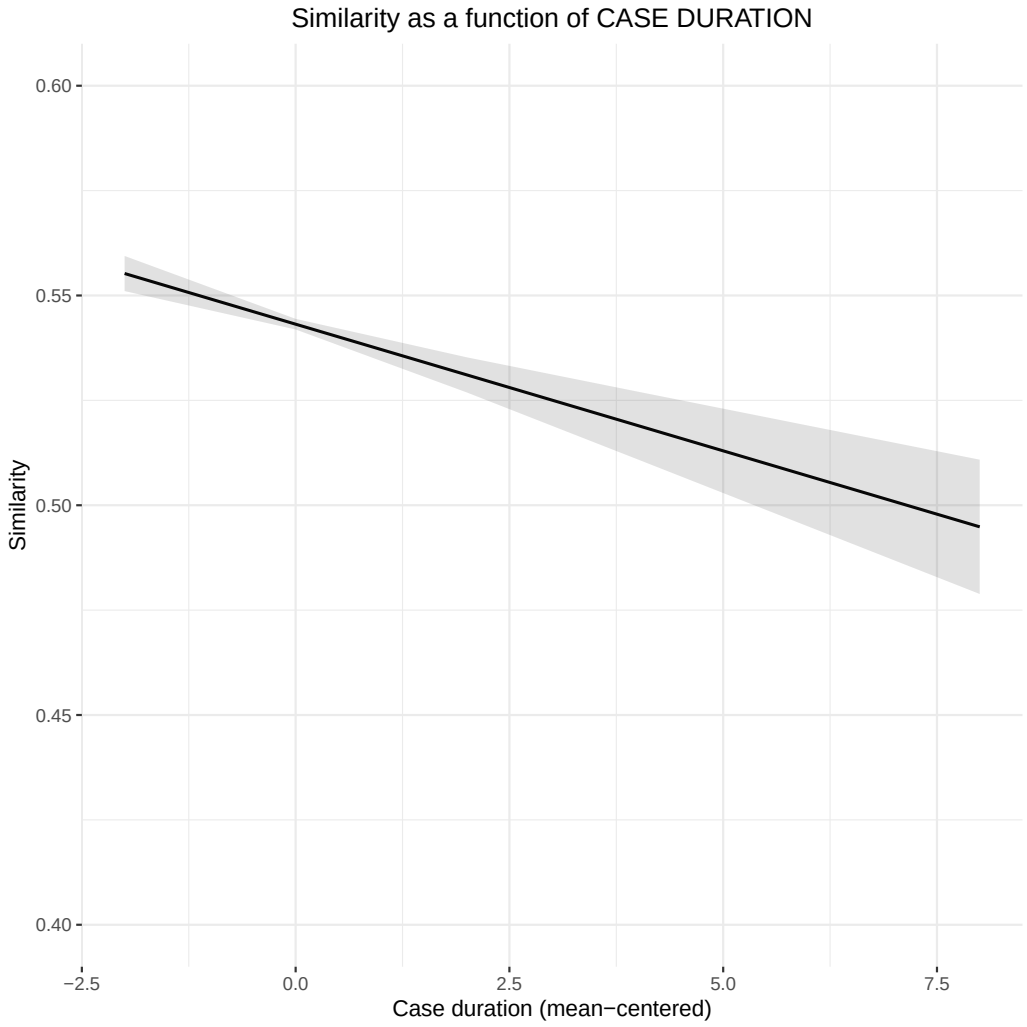

Figure 8: The longer the case has lasted, the less similar it is to preceeding cases.

case-intercepts to adjust the standard errors of case-level covariates (since each case is observed 10 times). The effects are illustrated in figure 7.

The chamber structure of the Court has changed over time. Overall, it can deliberate in 3 sizes: Small, medium and grand chamber formations.

Ever since the outset, lower-salience cases are decided in chambers of three judges. While the reliance on 3-sits has increased over time, these cases often include issues related to EU-staff. Comparing with judgments rendered by the plenary (prior to 2003) we see that the average similarity with preceeding cases tend to be somewhat higher (0.042 points) when rendered by the smallest formation.

In the 1970-ies member states also introduced a medium sized chamber formation; the chambers of 5 judges. Their use remained for a long time restricted by the Court's statutes. However, from the 1980-ies the Court informally relied on what was commonly labelled the "small plenary" for a number of medium-salience cases. The Nice treaty finally formalized the system by generalizing the use of 5-sits. In the model, we can see that all three formations have approximately equivalent effects. They tend to carry less similarities with their preceeding cases than judgments rendered by 3-sits, but are also more similar than judgments rendered by the Court's highest formations.

Formally, until 2003, the default formation was plenary deliberations. The Nice treaty effectively replaced it with the grand chamber (and at rare occasions) the full court. These are the formations which tend to render judgments with the lowest similarity on average.

Text similarity as predictor of influence

Finally, I consider text similarity with preceeding cases as a predictor of influence. Cases that make precedence contain two characteristics. On the one hand, they are dissimilar from the general Court output thus far. On the other hand, they tend to be discussed among academics and cited by judges.

Table 4: Text similarity as predictor of cases that set precedence

	The influence of a case (poisson reg.)	
	N. of annotations	N. of citations
	(1)	(2)
Similarity with 10 preceding cases (mean)	-4.583^{***}	-1.295^{***}
	(0.067)	(0.040)
Time (years)	0.062^{***}	0.059^{***}
	(0.002)	(0.001)
Time2	-0.001^{***}	-0.001^{***}
	(0.00003)	(0.00002)
Constant	2.993^{***}	2.533^{***}
	(0.041)	(0.024)
Observations	10,741	10,741
Log Likelihood	−55,676.120	−144,495.800
Akaike Inf. Crit.	111,360.200	288,999.700
Note:		*p<0.1; **p<0.05; ***p<0.01

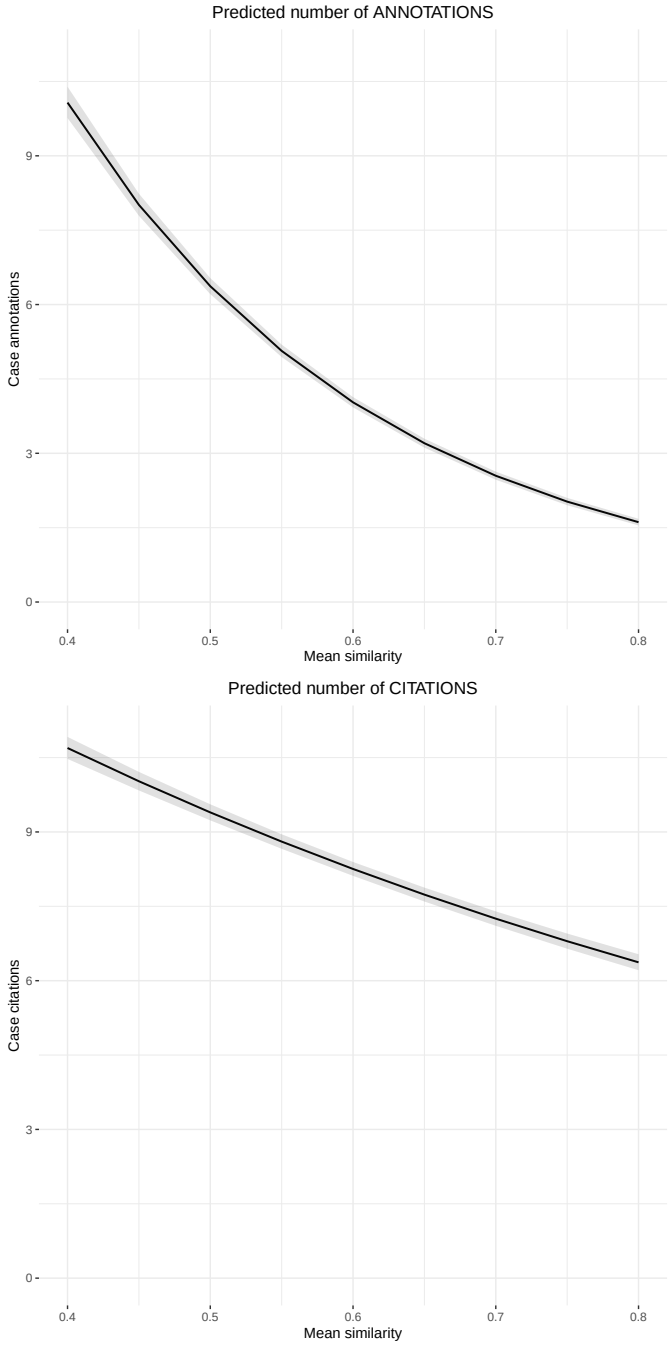

Figure 9: The effect of similarity with previous judgments and case interest in the legal community (number of case annotations) and for the court itself (number of citations).

Table reports the results from two poisson regressions. This time, cases are only observed once in the data, and time-varying elements are controlled for by a (curvilinear) time-trend. Similarity is calculated as the mean similarity with the 10 most similar preceeding cases. Results are reported in table 4 and figure 9.

The results provide ample support for the assumption that text (dis-)similarity is predictive of influence. A decrease of 0.1 points in similarity is predicted to decrease the number of published annotations by 63.2%. The same is observable for the number of citations to a case. A 0.1 increase in similarity leads to a predicted 87.9% decrease in number of citations compared to other cases concluded the same year.

Conclusion

In this research note I have trained a doc2vec model on a corpus of CJEU texts and investigated if the resulting similarities can be used for meaningful comparisons between documents. The results are promising, albeit not satisfactory.

I have verified the reliability and validity of the measure in different ways. First, I found that the model is able to predict similar document vectors as those on which it was trained. This in-sample prediction still has margins of improvement however. In future iterations, I would like to hold back a number of documents during the initial training in order to do an out-of-sample comparison of document vectors.

Second, I have benchmarked the results against trained human coders and found that document similarity to some degree is able to discriminate between cases in which the advocate general has suggested a different conclusion than the Court in preliminary reference cases. In this verification strategy, I held the litigation-specific substance of the case constant, since the two texts treat with the same preliminary reference questions. The results suggest that the

model can meaningfully identify differences in legal arguments at least when holding the topic constant.

Third, I have performed a more qualitative model assessment. I have investigated whether (dis)similarity with previous cases is related to common covariates of new and original Court cases. I have found that similarity is consistently correlated with these elements. Cases that are decided by a larger number of judges, take a longer time to decide or receive more attention from member states, legal scholars and other judges tend indeed to be less similar to their preceeding texts.

Overall, the results are nevertheless attenuated by the moderate spread in the estimated text similarities. In future iterations I would like to train the model on a larger corpus. Possible candidates are the case summaries published on EUR-Lex, the (English) case annotations already referred to in this research note as well as the text of the European laws cited in the cases' grounds of judgment. The latter training may also provide similarities that can be used to measure the effect of new legislation on the evolution of case-law.

References

Altszyler, E., M. Sigman, S. Ribeiro, and D. F. Slezak (2017, November). Comparative study of LSA vs Word2vec embeddings in small corpora: A case study in dreams database. *Consciousness and Cognition 56*, 178–187.

Bellegarda, J. R. and S. Member (1998). A multispan language modeling framework for large vocabulary speech recognition. In *IEEE Transactions on Speech and Audio Processing*.

Bengoetxea, J. (1993). *The Legal Reasoning of the European Court of Justice: Towards a European Jurisprudence*. Clarendon Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003, January). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Deerwester, S., S. T. Dumais, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

Denny, M. J. and A. Spirling (2018, April). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis 26*(02), 168–189.

Harris, Z. S. (1954, August). Distributional Structure. WORD *10*(2-3), 146–162.

Kusner, M. J., Y. Sun, N. I. Kolkin, and K. Q. Weinberger (2015). From Word Embeddings To Document Distances. In *Proceedings of the 32 Nd International Conference on Machine Learning*, Lille, France, pp. 10.

Larsson, O. and D. Naurin (2016). Judicial Independence and Political Uncertainty: How the Risk of Override Affects the Court of Justice of the EU. *International Organization 70*(2), 377–408.

Lau, J. H. and T. Baldwin (2016, August). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin, Germany, pp. 78–86. Association for Computational Linguistics.

Le, Q. V. and T. Mikolov (2014, May). Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013, January). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013, October). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Mikolov, T., W.-t. Yih, and G. Zweig (2013, June). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 746–751. Association for Computational Linguistics.

Pennington, J., R. Socher, and C. Manning (2014, October). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics.

Řehůřek, R. and P. Sojka (2010, May). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta. ELRA.

Saurugger, S. and F. Terpan (2017). *The Court of Justice of the European Union and the Politics of Law* (1 ed.). The European Union Series. London: Palgrave.

Tsurel, D., D. Pelleg, I. Guy, and D. Shahaf (2017). Fun Facts: Automatic Trivia Fact Extraction from Wikipedia. pp. 345–354. ACM Press.