Finding the Topics of Case Law: Latent Dirichlet Allocation on Supreme Court Decisions

Ylja Remmits

June 2017

Supervisors

Dafne van Kuppevelt (Netherlands eScience Center)
George Kachergis (Radboud University)
Gijs van Dijck (Maastricht University)

Abstract

The law produces a large amount case law, which is still mostly processed by hand. The Case Law Analytics project aims to develop a technology that assists the legal community in analyzing case law. As a part of this project, this thesis explores the possibilities of finding accurate and useful legal topics with LDA and whether or not legal experts and people with a non-legal background agree in their judgments about this. To this end I investigated possible methods suited for evaluation of the model's results. I evaluated the topics as well as their assignment to the documents using human evaluation. I found that the topics evaluated to cohere most, are easy to label. Human subjects were also mostly able to differentiate between topics assigned to a document with high probability and topics that do not belong to this document. However less than half the topics were evaluated as coherent by the subjects and according to the subjects the main topic of a document was not found by the model for most of the documents. I also found that domain experts and non domain experts might evaluate topics differently. I argue that the usability of the results depends on the intended application and and introduce some complications specific to the legal domain, which should be taken into account as well.

1 Introduction

Each year the Dutch courts decide on a large number of cases. In 2015 alone more than 1.6 million court cases were handled by the courts [1]. The Dutch law knows several sources of law, e.a. legislation, treaties, case law. The Netherlands have no common law tradition, where case law directly defines law. However, the law system does ensure legal certainty and equality [1]. Because of the principles of legal certainty and equality, by deciding on cases, courts also develop law.

Apart from this, case law together with all other sources of law is used by legal researchers to determine reasonable interpretations of legislation to describe the current state of a certain sub-area of the law. Case synthesis is a method playing an important role in this type of research. Case synthesis is the method of synthesizing idea's from groups of cases in order to determine the state of prevailing law at a specific moment. It entails comparing case outcomes to the facts of the cases and existing legislation and court decisions [2].

The analysis of court decisions now mostly relies on manual human analysis. The law produces a vast amount of cases each year, which must be read through and classified by trained experts. Of the 1.6 million cases in 2015, for almost 150,000 of those the textual content of the decision is publicly available on www.rechtspraak.nl. Consequently the size of the data set is too large for complete manual analysis and thus case law is studied based on a fraction of the total amount of cases.

1.1 Automated legal analysis in literature

Automated legal analysis is increasingly researched in the field and these advances have the potential to greatly aid the process of case synthesis and are expected to have an influence on all aspects of law, from education to legal practice [3]. Being able to automate analysis has the potential of greatly decreasing the manual work needed and allows for a larger portion of available case law to be used in research and case synthesis.

Previous research took different approaches. Some tried to directly predict outcomes of cases, see for example Aletras et al. who used support vector machines to predict decisions of the European court of human rights [4] or Katz et al. who used a time evolving random forest classifier to predict the behaviour of the US supreme court [5]. Others used a network approach to gain new insights in the large number of documents and aid in case synthesis. Here data is represented as a network of vertices and edges, where a court decision or an article of legislation is a vertex and a citation is a an edge [6]. Fowler et al. used several centrality measures (centrality is a measure of importantness of a node

[1] https://www.rechtspraak.nl/SiteCollectionDocuments/factsheet-rechtszaken-2015.pdf

in a network) to determine authority of case law in the U.S. [7]. Winkels et al. used the in-degree centrality (the number of incoming edges for a node) of nodes in a network of court decisions and investigated this measure as a predictor of case authority of Dutch case law [8]. Winkels also attempted to create a legal recommender system, a system which recommends related case law with the use of betweenness centrality of nodes in a network (the number of times a node is on the shortest path between two other nodes) [9].

This thesis project is part of a larger project that aims to develop a technology that assists the legal research community in analyzing case law. The Case Law Analytics [2] project takes the network approach and focuses on references in court decisions to other decisions and the networks that can be drafted from these references. The project explores the possibilities of creating an interface which can be used by legal researchers to build, visualize and explore the networks of references and the possibilities of enriching the networks with metrics that provide insight in the networks, such as several measures of centrality, or coloring of clusters. These networks can be used to visualize and organize, determine cases with authority and determine relevance of case law.

Finding themes or topics of the decisions, such as endangerment ("gevaarzetting") and notice of default ("ingebrekestelling"), is relevant, as these topics can be used in information retrieval tasks and as labels to clusters in the networks described above. Correct assignment of themes can assist in information retrieval and clustering by revealing implicit thematical and semantical relations between court decisions which remain invisible when only relying on explicit references in the decision itself [10]. This last aspect might also be useful when drafting the networks of related court decisions, as the Case Law Analytics project is trying to accomplish, as well as in the context of recommender systems [9].

To investigate the possibility of automatically finding topics in court decisions, I will use a probabilistic topic model, Latent Dirichlet Allocation(LDA). Probabilistic topic models are a class of unsupervised algorithms developed to discover and annotate large set of documents with thematic information [10]. A variation of probabilistic topic models exist,appart from LDA some examples are probabilistic Latent Semantic Analysis (pLSI) and Non-negative Matrix Factorization (NMF). In addition to the mentioned models, there is a vast amount of extensions and variations to these models. All mentioned models have their own value, depending on the data and the goal for which the topics are trained. Stevens et al. compared the three mentioned algorithms and concluded that LDA learns the best descriptive topics and recommend LDA for applications where humans interact with the learned topics [11], which describes our application. LDA is also one of the early, simplest and the most well used of those models in humanities and linguistics, see for example Blevin who used LDA to explore the topics through time in 18th century midwife Martha Ballard's diary [12]. LDA views documents as a mixture of topics expressed in the documents

[2] https://www.esciencecenter.nl/project/case-law-analytics

and discovered topics have been found to correspond to human judgments of topics in some cases.

This project will investigate the possibilities of finding accurate and useful legal topics with LDA and whether or not legal experts and people with a non-legal background agree in their judgments about this. Towards this end I will also investigate which methods are the most suitable for the evaluation of the results of LDA.

2 Latent Dirichlet Allocation

LDA assumes a generative model of language which can be explained with the following analogy: Before an author starts writing (part of) a text, she knows what this text will be about, in other words she knows the themes of the text. For each theme there are certain words from our vocabulary that go with this theme. With the themes and the words related to these themes in mind, we construct language. Let's deconstruct this analogy into a full generative model of language.

First, we introduce some vocabulary following Blei et al. [13].

- The *corpus* D is the full data set of all texts, court decisions in our case. All words in this corpus define our vocabulary.
- A *document* is one text, one court decision. A sequence of N words.
- A *word* is the single basic unit of data. It is an item from the vocabulary.
- A *topic* is the abstract notion of a probability distribution over the vocabulary. In our analogy this the selection of words from our vocabulary that go with a theme.

LDA assumes the topics are generated before the texts. Now for each document $\mathbf{w}$ in D we assume the following generative process:

1. Choose θ, the topic distribution over the document from a Dirichlet distribution $\text{Dir}(\alpha)$.
2. Choose length N.
3. For each word n from N in the document:
   (a) Choose a topic from z_n from the distribution in step 1.
   (b) Choose a word w_n for the corresponding topic z_n from $\boldsymbol{\beta}$, the distribution over words in the vocabulary, in essence the topics.

The Dirichlet distribution is a probability distribution over a set of probability distributions. A popular analogy for this distribution is the string cutting analogy. Say we want to cut a string with length 1 into k pieces, where k is set

prior. Each piece k_i can vary in length but has an average length also set prior. This average is α. The parameter k in our case is the number of topics.

This generative process is expressed in the following joint distribution:

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \boldsymbol{w_D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \Big(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\boldsymbol{\beta}, z_{d,n}) \Big) \tag{1}$$

Assuming this to be the model for the generation of the texts, we want to infer what the topics are in our corpus. In other words, given the corpus of documents, what are the topics, the topic distribution for each document and the topic assignments per word. This brings us to the key inferential problem of computing the posterior distribution of the hidden variables given the observed corpus. This posterior corresponds to the following equation [10]:

$$p(\boldsymbol{\beta}, \theta, \mathbf{z}|\boldsymbol{w_D}) = \frac{p(\boldsymbol{\beta}, \theta, \mathbf{z}, \boldsymbol{w_D})}{p(\boldsymbol{w_D})} \tag{2}$$

Inferring this marginal distribution $p(\boldsymbol{w_d})$ is intractable, but a variety of approximation methods including sampling based approximations and variational approximation [13] is developed to approach this posterior in practice.

LDA makes some important assumptions about corpora. It assumes documents to be a "bags of words" meaning that the order of the words in the document does not matter. It also assumes that the order of documents does not matter. These assumptions together are called the exchangeability assumption. The only thing that is taken into account by the model is whether or not a word is present in a document and with what frequency. Thus LDA aims to capture semantics and not syntax. Another assumption of LDA is that the number of topics is known a priori. This assumption forces the user of the model to set the number of topics prior to training the model. Extensions to LDA have been designed to address this problem, such as Hierarchical Latent Dirichlet Allocation [14].

3 Collecting and preprocessing the data

3.1 Rechtspraak.nl data set

De Rechtspraak is the official Dutch judiciary. They provide their data on rechtspraak.nl, a webservice which can be queried for meta data (ECLI's, European Case Law Identifiers) and court decisions. Not for all case law, the textual decision is published, for a substantial amount decisions only meta-data exists on rechtspraak.nl. Especially for older case law, the textual decision is not always openly accessible. The available collection is selected to be representative for

Dutch case law [3]. A comprehensive study on all aspects of the open accessibility of case law in the Netherlands was done by Opijen [15]. *rechtspraak.nl* can be queried for case law in an XML format. The full technical report can be found on rechtspraak.nl [4].

Even without the the full quantity available, the size of our available dataset is quite large. In the perspective of law development by courts, decisions of higher courts naturally have more authority, with supreme court decisions as the highest court in the Netherlands. Supreme court decisions are for this reason particularly of interest in the process of case synthesis. As a means of reducing the great amount of data to the most meaningful portion, the larger Case Law analytics project is mainly focused on the Dutch Supreme Court decisions. For this reason I chose to do the same.

On rechtspraak.nl there are about 30,000 decisions of the Dutch Supreme Court for which the textual content of the decision is available and not just the meta data. For each piece of case law all text between the $< uitspraak ><$ $/uitspraak >$ XML tags is used. These texts and the corresponding ECLI's are stored in a sqlite database.

All case law, with files smaller than 2 kilobytes of data (approximately 200 words) from my corpus. These very small pieces of case law generally provide no thematic information, but only a short procedural explanation. In case of an appeal at the Supreme Court, the Court does not decide on the facts again, but only decides on matters of law (i.e. whether the law was properly interpreted and applied). The Dutch law allows the Supreme Court to not give judgment in cases where the request is not a matter of law that is important for the unity or development of the law. In these cases the court will not judge on the case but it will limit itself to a reference to the law that allow this, with the very short documents as a result. These cases are called 81RO cases, named after the article of codified that allows this.

3.2 Preprocessing

All texts were preprocessed following common techniques in the Natural Language Processing field.

3.2.1 Tokenizing

As a first step all texts are tokenized, which is the splitting of texts into words, splitting of punctuation. Important here is that this is done in a manner compatible with the Dutch language leaving for example hyphens, which are a common

[3] https://www.rechtspraak.nl/Uitspraken-en-nieuws/Uitspraken/Paginas/Selectiecriteria.aspx
[4] https://www.rechtspraak.nl/SiteCollectionDocuments/Technische-documentatie-Open-Data-van-de-Rechtspraak.pdf

way of concatenating words in Dutch (e.a. procureur-generaal). I used the Treebank tokenizer in the NLTK [16] python package, which according to some small scale experimenting accomplishes this.

3.2.2 Stemming vs Lemmatization

One other common step is either stemming or lemmatization. The goal of both methods is to reduce derived or related forms of a word to a base form. Stemming commonly refers to the reduction of words to their stem, usually by simply removing specified prefixes and suffixes and removing the end of the word. Lemmatization is a more sophisticated and complex method of using morphological and part-of-speech information to find the dictionary form of a word, called a lemma [17]. Experiments with both lemmatizers and stemmers using real sentences from my corpus were conducted, since they are in Dutch I will give comparative examples in English.

An example of stemming would be reducing both 'walks' and 'walked' to their stem 'walk'. However, this is a rather crude instrument at moments: both 'marketing' and 'markets' to 'market' by the well-known Snowball stemmer [18]. Another problem with stemming for this project is the fact that stemmers tend to reduce the word to incomprehensible small part of words. The Snowball stemmer for example reduces 'variations' to 'variat' and 'quantitative' to 'quantit' [17]. These two problems make it hard, if not impossible for a human expert to translate these stems to a human readable version of the words. Because of the validation methods explained below, human readability is an important feature to preserve.

A solution to these problems is to use lemmatization, which is less efficient and more language specific. Pattern.nl is an open source python-based web mining and natural language processing module that has a lemmatization tool with, according to small scale experiments, acceptable accuracy [19].

Lemmatization by Pattern is not perfect: sometimes the chosen lemma is not an existing word, as we found with stemming. For example 'software' is lemmatized to 'softwaar' and 'kenteken' (license plate) to 'kenteek' (comparable to 'license plaat'). These errors are understandable, in the perspective of the Dutch language, because these forms exist in other words common in the language. Also, on the contrary to the errors made by our stemming algorithm, it is possible for any Dutch native to deduce the correct Dutch word from the shown lemma. This makes lemmatization a more suitable method for us to reduce the amount of words which have the same meaning. Before lemmatization the corpus contains 272797 unique tokens, afterwards the corpus contains 218981 tokens.

Table 1: Words removed appearing in more than 50%, 70% or 99% of the documents

	Fraction of documents		
	0.5	0.7	0.99
Words Removed	136143	136094	136094
Words Remaining	82688	82737	82737

3.2.3 Word removal

As Zip's law describes, most of the tokens in a text are accounted for by a few very high frequency words and there are many low frequency words. LDA assumes that related words will co-occur in texts. Words with a very high frequency, will co-occur with most other words, since they appear in most texts. From the assumption of LDA and this notion, one can see that LDA might assume semantic relationships between all these highly frequent words, which tell nothing about the actual structure of the corpus and relationship between documents.

The 100 most common words in the corpus were removed, following [20]. Upon inspecting these words, we find the most common verbs in Dutch such as 'have', 'read', 'see' and 'come' and large amount of procedural legal terms which are expected to be in supreme court decisions and appear to be non-informative, such as 'raad' (council), 'artikel' (article), 'voorzitter' (president) and 'beroep' (appeal). The full list is included in the appendix. In addition all words on a 50 word stop words list were removed. The list is added in appendix 1. 21 words in the stopword list were overlapping with the 100 most frequent words.

There is a large number of words appearing only once in the corpus. Co-occurrence is the basis of LDA, therefore these words do not contribute to the topic generation. The terms appearing only once in the corpus were removed, following the original experiments by Blei et al. [13]. However words appearing in only one document in the corpus, independently of how frequent this word is in that document, do not tell us anything about the relation of this document to other documents. For this reason a broader criterion is used and all words appearing in only one document are removed, which naturally includes all words with an absolute term frequency of one. 136095 words were removed that only appear in one document. Experiments were done removing all words appearing in more than 50, 70 and 90% of the documents and the results are shown in table 1. There are only 49 words appearing in more than 50% of the documents. There were no words appearing in more than 70% or 90% of the documents which explains why there is no difference between these cases.

After all word removal steps, the corpus has 19421 documents with 82737 unique words, an average document length of 657 tokens with 326 different words and a median document length of 472 tokens and 274 different words.

4 Methods

4.1 Training the model

Choosing the right number of topics can be a hard decision, which can strongly depend on the methods of validating these topics. In some of the previous research small experiments are done with different numbers of topics, before choosing the number of topics to compute, see for example [21]. However, as Chang et al. showed [22], statistical measures of model quality do not necessarily correlate to human evaluated model quality. Since the real world application is the main goal of this project and human evaluation is the main form of validation, experiments for different numbers of topics should be evaluated by humans. Consequently, each experiment with a different number of topics would require a number of subjects and a substantial time investment by the subjects, both of which were only available in a limited amount in this project. This limited the possibilities of doing extensive experimenting with different amount of topics. Chang et al. showed that more topics do not imply better results when evaluated by humans. Their model trained to yield 50 topics was evaluated to yield better topics than the model trained to yield 100 topics for two different corpora [22]. Quinn et al. who experimented with different numbers of topics and evaluated the results manually, also found that a relatively small number of topics was evaluated best (42 topics) [21].

Following these results, where a relatively small number of topics is evaluated best by human judgment, LDA is trained on 50 topics with Gensim [5]. Gensim is a software framework for topic modelling of large corpora, and offers an easy to use LDA implementation in the form of a python library [23]. Gensim's LDA implementation is based on Hoffman [24].

4.2 Model Evaluation in literature

Researchers commonly use some metric of model fit to evaluate topic models, for example perplexity or held-out likelihood. Metrics of model fit are useful to assess the reliability of the predictive model for the data. Most researchers do not address the validity of these measures, where validity is defined as the extent to which a measuring instrument measures what it is intended to measure [25]. Originally Blei et al. even warn to make "no epistemological claims regarding these latent variables (topics) beyond their utility in representing probability distributions on sets of words" [13].

In literature several researchers have addressed the validity of existing measures of evaluation for topic models and the relevance of research in this area. In his 2012 survey, Blei defines probabilistic topic models topic models to be a suite of algorithms aiming to "discover and annotate large archives of documents with

[5] Code can be found at: https://github.com/Ylja/CaseLaw

thematic information". He states that there is no evidence leading to the assumption that higher scores on statistical measures imply easier interpretation. He emphasizes that the development and validation of evaluation methods remains largely open for future research [10]. Van der Zwaan et al, emphasize that, while in application domains such as politics the importance of the validation of evaluation methods for topic modeling is increasingly recognized, it is also very relevant for computer scientists. Insight into which methods are successful and why, can help with the improvement and design or improvement of topic models and model checking problems [20]. Grimmer et al. [26] focus on the domain application of topic models in polictics. They state that "All quantitative models of language are wrong—But some are useful". Because of the powerful simplifying assumptions topic models make, information is always lost in some dimension. The generative model of language as described above is clearly different from a human's intuition about how we generates our language. Grimmer et al. emphasize that the only way to evaluate the "usefulnesss" is profound validation of the results. They suggest combining experimental, substantive and statistical evidence to demonstrate the conceptual validity of topic models [26].

Chang et al. address these problems by designing a method for measuring human interpretability and comparing this to the predictive log likelihood. They validate the assumed topic coherence and relevance of topic models using human experiments. For example in one of the experiments subjects were shown the top 10 words of a topic, plus a random word. Subjects then had to pick the intruding word. They found that the statistical measures tested are negatively correlated to the evaluations tasks they developed.

Quinn et al. present methods to qualitatively evaluate their topic model trained on political documents. They conclude that unsupervised methods and probabilistic topic models in general, essentially shift the burden of defining labels beforehand (as done in supervised models) to evaluation and validation of the results afterwards. They heavily rely on their own domain knowledge in their validation methods.

This research is focused on investigating the real-world applications of LDA for case law. To validate these results, following the above research, I mainly focus on human evaluation methods.

Grimmer and Steward state that the first step in validating the results of any topic model is to label the topics [26]. Quinn et al. also use this as their first step [21]. I will not follow their research in this. Labeling the topics requires a domain expert with knowledge of the model, unavailable in this project. In addition the task of labeling the topics is a highly subjective task and thus quite variable. With topics where there appears to be little cohesion between the words it is hard, if not impossible, to find agreement on their common label. The decision of whether or not 10 words form a coherent whole is a much easier decision than deciding on one label for these 10 words.

4.3 Measuring topic quality

To assess the quality of the topics, the main question is whether the words in a topic actually form a coherent theme. Human subjects are asked to evaluate the coherence of the 10 most probable words for all 50 topics. To this end I used an online questionnaire. For each topic subjects responded to the following statement: "These ten words form a coherent whole". They answer on a 5 point Likert scale: fully agree, agree, neither agree nor disagree, disagree, fully disagree, scored 1-5.

4.4 Topic distribution quality

To assess the model's distribution of topics over the documents, the extent to which human subjects agree with the mixture of topics as a description of a document's content has be to measured. To this end Chang et al. developed the topic intrusion method. In this task subjects were shown 10 randomly picked documents together with four topics (the topics represented by their 10 most probable words). Three of the topics were the topics assigned to this document with the highest probability. The fourth is randomly picked from topics assigned to the document with a very low probability, this topic is the intruding topic. For each document these four topics are presented in a random order. Subjects are instructed to choose the topic which does not fit the document [22].

Diverging from Chang et al. subjects were provided with the full text. In the original experiment subjects did not see the full document: they were shown only the title and the first few sentences. This was done due to time constraints and was possible because their corpora (newspaper and encyclopedia) are structured in a way that the article gives an overview of the document in the first few sentences. My corpus is not structured in this way. A regular piece of case law is structured in several sections each providing profoundly new information about the case, for example procedural details, facts, considerations by the court or judgment. For this reason showing the first part of the document does not seem reasonable.

The topic intrusion task does not address the question of whether the topics learned do capture the main themes of the text. This problem is regularly addressed by comparing the topics model's results to the results of supervised methods, as done by Zwaan et al. [20]. However, since correct labels are unavailable, supervised methods cannot be used. In an attempt to address this, subjects were asked a second question, about each of the decisions they read: "Which of the following topics describes the main theme of the text?" Here the subjects were given four options, in a random order: the three topics with the highest probability according to the model and an option "None of the above".

5 Results

5.1 Topic Quality

Table 2: The five worst topics, translated to English [5], the original topics can be found in the appendix

.

Topic nr	Average	StDev	Topic
44	1.33	1.70	"page" 1988 1989 1987 1990 1984 1986 1991 "painting" "real"
7	1.70	1.01	a b "private company" c d "eg." e f "partnership" g
21	2.0	1.17	"foundation" "accountant" vie d'or "supervision" "official supervising body for insurance companies" dhow "actuary" edco "title"
28	2.0	1.45	the to and a or that for be as on
19	2.1	1.40	2011 2008 2009 2012 2014 2015 "the" 2.1 "hague" 2.3

A full list of topics is provided in the appendix. When inspecting the topics model we can already see some topics that do not provide any thematic information. These are topics for example 7, 19 and 44 (translated to English[6]):

7. a b b.v. c d bv e f "company" g

19. 2011 2008 2009 2012 2014 2015 den 2.1 haag 2.3

44. blz 1988 1989 1987 1990 1984 1986 1991 "painting" reaal

An other notable topic is topic 28 containing only English words. This makes sense taking into account the model of language employed, but other than telling us that this piece of case law contains English language, a quote perhaps, it gives us no information about the theme of the document.

The topics were evaluated by 10 subjects, 6 lawyers and 5 people with a non-legal background. One lawyer did not finish the survey, this response was excluded from the results.

All topics which were noted on first visual inspection were also evaluated as non-coherent by the subjects as shown in table 2.

The 5 topics that were evaluated as most coherent by the subjects are mostly easy to label, as I have done in table 3. It is remarkable that only one of these topics is a stricly legal topic (topic 34), while the four others might also appear in non-legal documents.

[6] Acronyms or (company)names can not always be translated to English, in this case the original word is presented without quotation marks

Table 3: The five best topics, translated to English[5], the original topics can be found in the appendix.

Topic nr	Average	StDev	Topic
25	4.6	0.67	Medical: "report" "competent" "hospital" "medical" "treatment" "her" "danger" "doctor" "examine" "psychiatric"
2	4.4	0.521	Car transport: "driver" 1994 "vehicle" "road" "car" "driving license" "motor vehicle" "reason" "accident" "speed"
0	4.4	0.93	Drugs: "opiates law" "quantity" "cannabis" "substance" "gram" "Dutch coffeeshop" "intentionally" "list" "cannabis plant" "find"
34	4.3	0.92	Legal procedure: "request" "court" "letter" "treatment" "document" "judge" "verdict" "appear" rv "court hearing"
18	4.3	0.67	Insurance: "insurance coverage" "insurance" "to insure" "insurance company" "insurance premium" "coverage" "policy of insurance" "damage" "claim

If we look at the top 5 best evaluated topics for each of the groups, there is a large overlap between the separate sets of best topics and the average set of best topics. Lawyers and non-lawyer both have the Car transport, Drugs and Medical topics in their top five. The non-lawyers share all topics with the average top 5 except for the Insurance topic, where the lawyers share all topics except for the Legal procedure topic.

It is remarkable that the top 5 topics, as shown in table 3, have much smaller standard deviations than the bottom 5 topics. A possible explanation for this is that people are better at judging coherence than incoherence. Another explanation might be that the question as it is phrased ("Do the 10 words form a coherent whole") is ambiguous, especially for the topics 44, 7, 28 and 19 which do not give any thematic information but do cohere in the formal sense of the question asked.

In figure 1 all scores for all topics are shown, sorted on average scores. Of all 50 topics, 24 topics are scored significantly higher than 3 ($p < 0.02$ in One-Sample T Tests). Topics are quite evenly spread over scores, but there are some differences between our tests groups.

Figure 2 shows a histogram of the differences between test groups. The average difference is significantly lower than zero (p=0.001 in a 1000 sample bootstrapped one-sample t-test), meaning the average lawyer scores are significantly lower than the average non-lawyer scores.

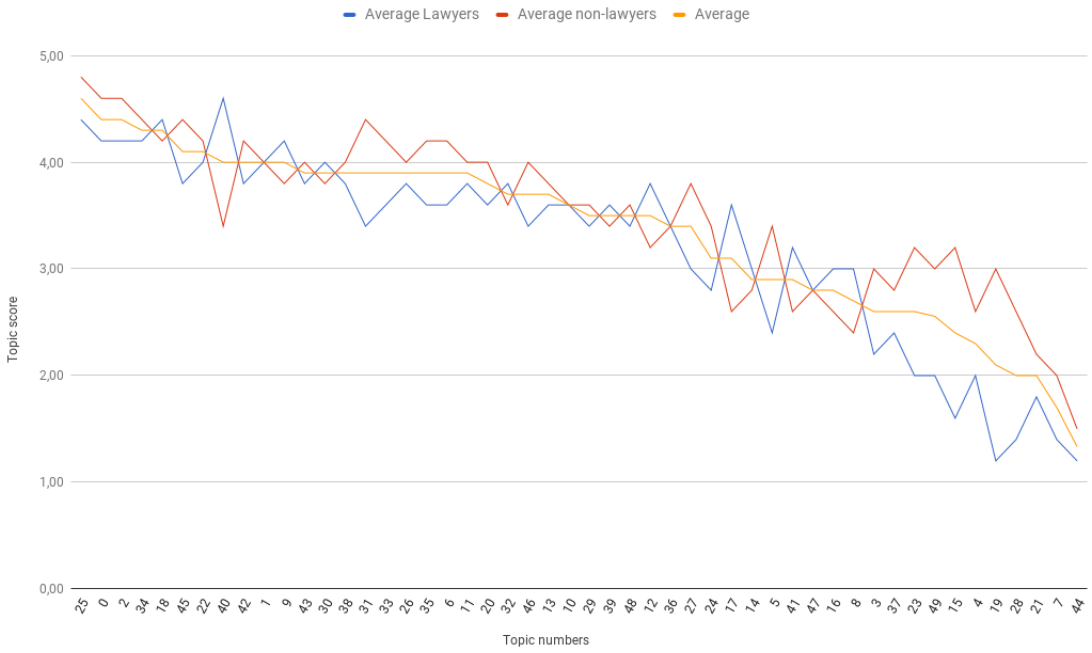

Figure 1: Scores for the topics sorted on average score

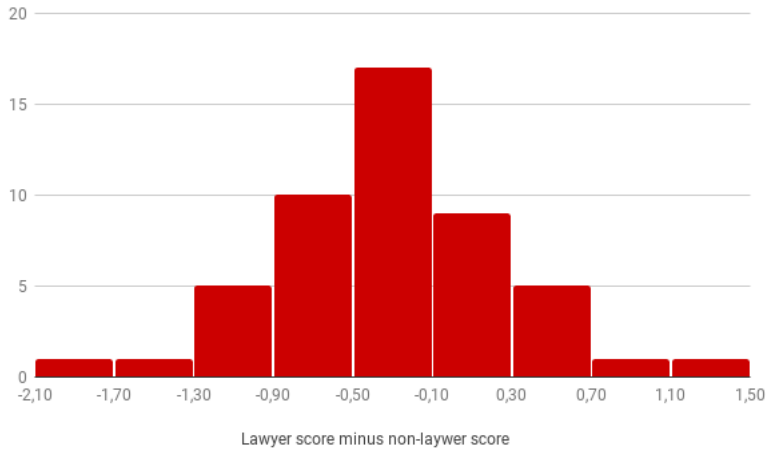

Figure 2: Histogram of differences between lawyer and non-lawyer scores. The average difference is lower than zero (p=0.001) meaning lawyer scores are on average lower than non-lawyer scores

Table 4: Topics with a significant($p < 0.05$ in a 1-tailed T Test for Independent Samples) difference between the score given by lawyers or non-lawyers) translated to English[5]

Topic nr	p	difference	topic
15	0.0095	-1.6	"Germany" "German" gmbh sypesteyn das "club" "Dutch Soccer federation" "Dutch" "stadium" glencore
19	0.02	-1.8	2011 2008 2009 2012 2014 2015 "the" 2.1 "hague" 2
23	0.006	-1.2	"part" rov "circumstance" "her" "complaint" "incorrect" "statement" "because" "insufficient" "but"
27	0.0325	-0.8	"suspect" "fact" "writing" "hearing" "means" sv "counselor" "propose" "on behalf of" "criminal case"
45	0.047	-0.6	"employee" "activity" "employer" "book of civil law" "labor contract" "sub-district court judge" "period" "January" "year"
40	0.017	1,2	"share" "company" "shareholder" "holding" "loan" "capital" "her" "fund" b.v. "interest"

When performing a T Test for independent samples for each of the topics [7], we find 6 individual topics where there is a significant difference in average score between lawyers and non-lawyers, shown in Table 4. What is remarkable when inspecting these topics is that 4 of these 6 topics are topics with a legal theme. This indicates that domain experts might score domain related topics differently than non-experts. An explanation for this might be that a non-lawyer might not understand the legal meaning of words, but does recognize them to be legal terms, thus judging a non-coherent topic full of legal terms to be coherent. The opposite can also be true which is visible in the one topic scored higher by the lawyers (topic 40), that because of their domain knowledge a lawyer sees coherence between words, lost to someone who is not a domain expert.

5.2 Topic Assignment Quality

The topic assignment was evaluated by 5 subjects, all with a legal background. One did not finish the survey, this response was excluded from the results. 10 documents were selected at random. One randomly selcted piece of case law was not used, as the Supreme Court ruled it to fall under article 81 RO as described in the data section. The topics 44, 7, 28, 19 were excluded since they provide no semantic information as described above. All selected pieces of case

[7] Independent-Samples T Tests assume normality of the data. This assumption might be violated for this data set.

law together with their topics and the randomly picked intruder are provided in the appendix.

Chang et al. have defined a quantitative measure of agreement between human judges and the models estimates, the topic log odds.

$$TLO_d^m = \frac{(\sum_s log\hat{\theta}^m_{d,j^m_{d,*}} - log\hat{\theta}^m_{d,j^m_{d,s}})}{S} \tag{3}$$

Here $\hat{\theta}^m_{d,j^m_{d,*}}$ is the probability mass assigned to the intruder topic and $\hat{\theta}^m_{d,j^m_{d,s}}$ the probability mass assigned to the topic selected as the intruder by subject s, S is the number of subjects [22].

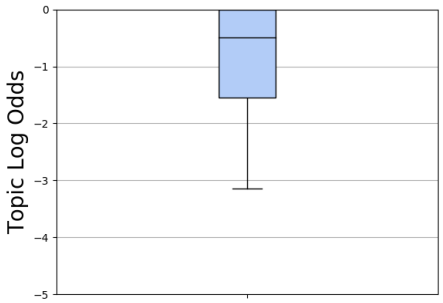

Figure 3: The topic log odds for LDA on supreme court case law. Higher is better.

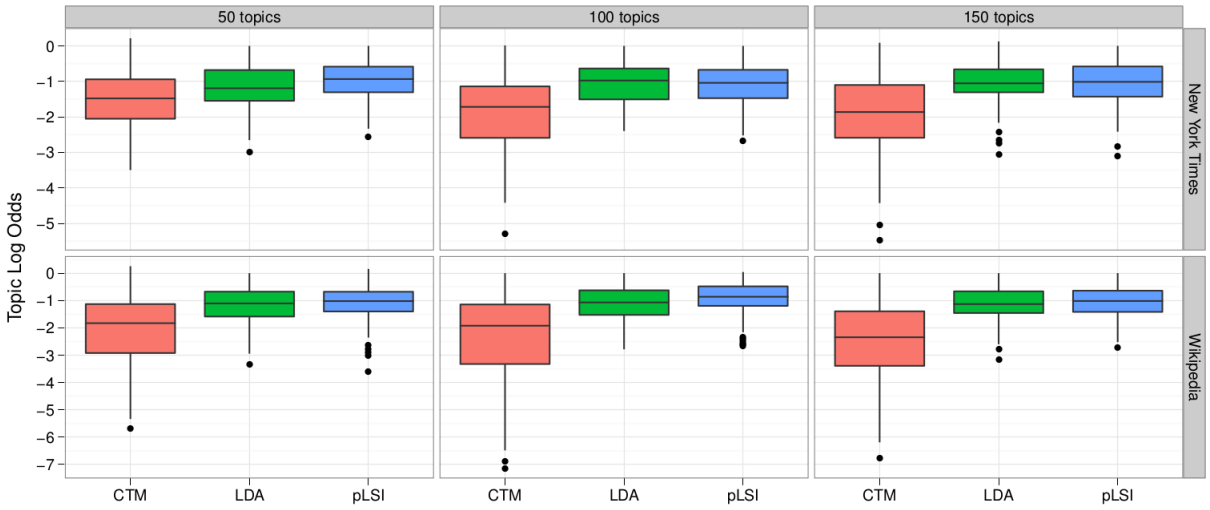

Figure 4: Results from Chang et al. The topic log odds for three models on two different corpora. Higher is better.

A box plot of the results is shown in figure 3. Higher values are better with an upperbound of 0. When comparing the results to those of Chang et al. in figure 4, the results are roughly comparable. The lower endpoint, first quartile and median value of the 50 topic LDA model trained on the case law corpus are higher than those values for the LDA models trained on the New York Times and Wikipedia.

There is an abnormality in the results that draws the attention. For one of the documents, two subjects pick one topic to both not fit and be the main topic of the document (document identifier ECLI:NL:HR:2011:BP3968). This might mean a mistake by the subject perhaps caused by alternations the question about best fit and worst fit, another explanation is that the questions are ambiguous or unclear.

Subjects were mostly able to chose the actual intruder. Table 5 shows the fraction of subjects picking the correct intruder for each document. In only one case they all picked a topic that was actually assigned to the document. The intruder in this case was quite a procedural broad topic. The other document for which more than half of the subjects could not identify the intruder was the document described above, where I suspect mistakes have been made.

When asked about the main topic of a document, for 8 out of 10 documents at least half of the subjects answered "none of the above" indicating that they felt like none of the three topics assigned to the document with the largest probability, was actually the main topic of the document. I will discuss possible reasons for these results below.

Table 5: The portion of subjects correctly identifing the intruding topic for each document

.

Item	Identifier	Portion of subjects correctly identifying intruder
1	ECLI:NL:HR:2014:2773	1.0
2	ECLI:NL:HR:2011:BP3968	0.25
3	ECLI:NL:HR:2011:BP4800	0.5
4	ECLI:NL:HR:2008:BD1842	0.0
5	ECLI:NL:HR:2001:AB2151	1.0
6	ECLI:NL:HR:2002:AD9487	1.0
7	ECLI:NL:HR:2014:3303	1.0
8	ECLI:NL:HR:1998:AA2396	1.0
9	ECLI:NL:HR:2014:1303	0.75
10	ECLI:NL:HR:2013:BZ7150	0.75

6 Discussion and conclusions

The law produces a large amount of data which is now processed by hand. This project attempted to aid the development of technology capable of assisting the legal community in analyzing case law. Finding themes and topics in case law can be relevant to label case law, find related case law or aid in corpus exploration. This research aimed to investigate the possibilities of finding accurate and useful legal topics with LDA and whether or not legal experts and people with a non-legal background agree in their judgments about this. To this end I investigated possible methods suited for evaluation of the model's results. I trained LDA with 50 topics and evaluated the topics themselves as well as their assignment to the documents using human evaluation, partly following Chang et al. I found that the topics evaluated to cohere most, are easy to label. Also subjects were mostly able to differentiate between topics assigned to a document with high probability and topics that do not belong to this document. However more than half the topics were not evaluated as coherent by the subjects and according to the subjects the main topic of a document was not found by the model for most of the documents. We also found that domain experts and non domain experts might evaluate topics differently.

Whether or not the results of a topic model are good enough to be used in practise, is firstly dependent on the goal. The results appear to be comparable to the results of Chang et al. [22]. However it is possible that the results can not be directly compared, since there are some differences in the setup (subjects reading only the intro versus subjects reading the full document) and the instructions to the subject (subjects were not presented the 'correct' answer after choosing). Also Chang et al. did not release exact results, apart form the box plots shown in figure 4. When comparability is assumed, the results appear to be in the same order of magnitude. Also for most documents, most subjects are capable of identifying the intruding topic. While these are promising results, it depends on the goal whether this is good enough.

When the goal of a project would be to use these topics to label clusters or nodes in a network, the topics should firstly be summarized into a single label per topic. This is not easy for the majority of the topics found. Secondly for the labels to be useful all or at least most relevant legal themes should be expressed in the topics. It can be concluded that this is not the case for legal experts judged the main theme to be missing from the top three topics for the majority of documents. However, when the goal is to use the topics as a measure of semantic relatedness of pieces of case law or for corpus exploration, this might very well be possible with the found results.

6.1 Model validity

When using probabilistic topic models intended for practical use, evaluation should always involves human subjects. As mentioned in chapter 4.2, unsupervised models shift the burden from work beforehand (e.g. labeling of data) to validation afterwards [21]. There are currently no complete methods described in literature for the validation of the models results. While Chang et al. attempted to define a method for this [22], my experiments show that good topic log odds do not imply that the result are of practical use.

Chang et al. also defined a task for measuring topic coherence. This is comparable to the topic intrusion task described above, but in this case involves an intruding word in a topic instead of an intruding topic for a document. This task seemed unfit from the first inspection of the topics, since all topics computed had one or a few seemingly intruding words. Since the results are binary: the subject chooses the true intruder or not, this would most probably result in low scores for almost all topics. For this research we wanted more than a binary result for each topic and because of the lack of resources it was unfeasible to use multiple tasks. For these reasons I decided not to use the the topic intrusion task.

Labeling the topics as Grimmer and Steward and Quinn et al. [26] [21] is hard. In small scale conversations with the subjects I found that subjects tend to focus on the first words of the topic and words that do not seem to fit the topic appear to have a lot of weight in their reasoning. Labeling the topics becomes much easier when all possible themes are known beforehand. In this case we can try and match the computed topics with the available themes as labels.

6.2 Required domain knowledge for human evaluation

To my knowledge no extensive study has been done on the effects of domain knowledge in the evaluation of the results of probabilistic topic models. This project also only briefly touches this subject. However it is notable that there is quite some difference between the two groups of human subjects, even though the sample sizes were small. From this small set of results it appears that having knowledge of the domain might influence the judgment. The extent of the influence of domain knowledge on the human judgment of topic model remains an interesting and open field of research.

6.3 Specific complications of topic modeling in the legal domain

Apart form jargon influencing the evaluated topic coherence, the legal domain might also have additional specifics that complicate topic modeling. Conversations with lawyers about their language, their field and their case law led to

some insights in the additional difficulties of attempting topic modeling in their complex field.

Firstly the topic of a piece of case law is quite depending on which lawyer is asked and the intention of this lawyer upon reading that particular document. When the lawyer is for example an attorney representing a client in a criminal law case, this lawyer might be looking for cases in which the facts are similar to the case of his client and thus the factual topic might be of most interest to him. If, however the lawyer is a legal researcher, involved in the process of case synthesis, this lawyer might not be as interested in the facts but more in the procedural topics of the case. As explained above the Supreme Court decides on different aspects of cases than the lower courts. In almost all cases the facts are stated, whether the Supreme Court is deciding or any lower Court. This means that the facts will always influence the topics computed and the distribution over topic for a specific document, even if this factual information does not provide the lawyer with what he is looking for.

Secondly legal topics appear to be combinations of several broader topics. A document about a topic such as 'employer liability' might be assigned some probability for a topic such as 'employment' but might also be assigned some probability for 'insurance' or 'damage'. In the model resulting from my experiment, topics about employment and insurance are present, but it is hard, if not impossible to extract the specific legal term from these separate topics.

Lastly, a lawyer's perception of a topic is also greatly influenced by the articles of codified law that are mentioned in the court decision. These articles are often only mentioned by name and not fully cited. Therefore their content is not part of the model. Articles are cites in the form "article 6.2 of sr" where sr stands for a specific book in the Dutch law. In this research this would be tokenized into separate words. Keeping these expressions together, would require an extensive pipeline, such as designed by van Opijnen [27]. Using the mentioned articles, perhaps even the articles' texts in the topic model as tokens might be an interesting extension to investigate.

6.4 Recommendations for topic modeling in the legal domain

When attempting to use a topic model in the legal domain, the goal of the intended user should be considered. As described above this can influence of which courts one requires the case law and the sections of this case law to be used. Adding references to codified law as tokens, might also improve the results.

Exploring the possibilities of using probabilistic topic models to find related pieces of case law, would be an interesting open area of research. It can be evaluated whether two document have the same topics assigned to them by the model, are judged to be similar by humans. If this proves to be possible, it could

possibly be a great addition to the reference networks as drafted in Case Law Analytics projects as well as in the context of recommender systems [9].

Acknowledgement

I would like to thank Janneke van der Zwaan en Carlos Martinez Ortiz for their help and the fruitful discussions we had.

References

[1] J. C. Hage, "Recht, vaardig en zeker," 2009.

[2] J. K. Giofriddo, "Thinking like a lawyer: The heuristics of case synthesis," *Tex. Tech L. Rev.*, vol. 40, p. 1, 2007.

[3] J. O. McGinnis and R. G. Pearce, "The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services," 2014.

[4] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the european court of human rights: A natural language processing perspective," *PeerJ Computer Science*, vol. 2, p. e93, 2016.

[5] D. M. Katz, M. J. Bommarito II, and J. Blackman, "A general approach for predicting the behavior of the supreme court of the united states," *PloS one*, vol. 12, no. 4, p. e0174698, 2017.

[6] R. Whalen, "Legal networks: The promises and challenges of legal network analysis," *Michigan State Law Review*, vol. 2016, no. 2, p. 539, 2016.

[7] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, "Network analysis and the law: Measuring the legal importance of precedents at the us supreme court," *Political Analysis*, pp. 324–346, 2007.

[8] R. Winkels, J. d. Ruyter, and H. Kroese, "Determining authority of dutch case law," 2011.

[9] R. Winkels, A. Boer, B. Vredebregt, and A. van SOMEREN, "Towards a legal recommender system.," in *JURIX*, pp. 169–178, 2014.

[10] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[11] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler, "Exploring topic coherence over many models and many topics," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961, Association for Computational Linguistics, 2012.

[12] C. Blevin, "Topic modeling martha ballard's diary," 2010.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[14] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *Journal of the ACM (JACM)*, vol. 57, no. 2, p. 7, 2010.

[15] M. van Opijnen *et al.*, *Op en in het web: Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd*. Boom Juridische uitgevers, 2014.

[16] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[17] C. D. Manning, P. Raghavan, H. Schütze, *et al.*, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[18] M. Porter and R. Boulton, "Snowball stemmer," *On line http://www. snowball. tartarus. org.[Visited 15/11/2005]*, 2001.

[19] T. D. Smedt and W. Daelemans, "Pattern for python," *Journal of Machine Learning Research*, vol. 13, no. Jun, pp. 2063–2067, 2012.

[20] J. M. van der Zwaan, M. Marx, and J. Kamps, "Validating cross-perspective topic modeling for extracting political parties' positions from parliamentary proceedings," in *ECAI 2016: 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands-Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, vol. 285, p. 28, IOS Press, 2016.

[21] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to analyze political attention with minimal assumptions and costs," *American Journal of Political Science*, vol. 54, no. 1, pp. 209–228, 2010.

[22] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models.," in *Nips*, vol. 31, pp. 1–9, 2009.

[23] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Valletta, Malta), pp. 45–50, ELRA, May 2010. http://is.muni.cz/publication/884893/en.

[24] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, pp. 856–864, 2010.

[25] E. G. Carmines and R. A. Zeller, *Reliability and validity assessment*, vol. 17. Sage publications, 1979.

[26] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political analysis*, pp. 267–297, 2013.

[27] M. v. Opijnen, N. Verwer, and J. Meijer, "Beyond the experiment: the extendable legal link extractor," 2015.

Appendices

A Stop Word List

aan
af
al
als
bij
dan
dat
de
die
dit
een
en
er
had
heb
hem
het
hij
hoe
hun
ik
in
is
je
kan
me
men
met
mij
nog
nu
of
ons
ook
te
tot
uit
van
was
wat
we
wel
wij
zal
ze

B 100 Most Frequent Words

hebben
door
hoog
op
raad
1
raadsheer
2
3
voorzitter
niet
voor
maken
beroep
kunnen
tegen
uitspreken
4
worden
cassatie
daarvan
gehecht
deel
deze
instellen
geen
wijzen
beslissing
arrest
beoordeling
gerechtshof
geding
hof
uitspraak
ter
ten
vice-president
nr
naar
moeten
middel
zaak
openbaar
onder
der
zich
om
5
na
eerste
oordeel
advocaat-generaal
bestrijden
zullen
volgennen
grond
belang
meer
vernietigen
lid
nederlanden
3.1
geven
verklaren
bedoelen
advocaat
griffier
over
komen
oordelen
concluderen
mr.
stellen
volgen
doen
leiden
“
hiervoor
conclusie
hoger
artikel
brengen
zien
zouen
art
geval
staan
nemen
dienen
houden
3.2
waarnemen
nader
overwegen
feitelijk
behoeven
veroordelen
hierna

C Topics (in Dutch)

0. opiumwet hoeveelheid hennep stof gram coffeeshop opzettelijk lijst hennepplant aantreffen
1. eiser verweerder vordering verweerster eisere c. partij vonnis rechtbank schade
2. bestuurder 1994 voertuig weg auto rijbewijs motorrijtuig reden ongeval snelheid
3. belanghebbennen belanghebbende inspecteur staatssecretaris aanslag hofs financiën wet proceskost jaar
4. uur brand lood drank 8 alcohol fles onderzoek yukos plaats
5. ecli auto post personenauto gn bpm gebruiken gebruik kenteek indelen
6. bedrag voordeel verkrijgen kost betalen per bedragen jaar betaling totaal
7. a b b.v. c d bv e f vennootschap g
8. voorziening bijdrage casino kost verhaal kaart onderhoud gebied last hilversum
9. kind moeder vader ouder dochter zoon minderjarig gezag school haar
10. nederland nederlands ontvanger verdrag belasting buitenlands belgië 1990 toepassing recht
11. ondernemingskamer transactie rechtspersoon effect n.v. bestuur belegger commissaris informatie aandeel
12. 2016 besluit ambtenaar aanwijzing bevoegdheid wettelijk voorschrift uitoefening bevel dienst
13. organisatie vereniging deelnemer leed kansspel spelen klant deelnaam deelneming spel
14. nederlands gerecht curaçao aruba sint antillen gemeenschappelijk vonnis octrooi ep
15. duitsland duits gmbh sypesteyn das club knvb nederlands stadion glencore
16. beschikking man rechtbank vrouw verzoeker verzoek partij 2005 per verzocht
17. omzetbelasting belanghebbennen wet letter dienst levering verrichten aftrek naheffingsaanslag zake
18. uitkering verzekering verzekeren verzekeraar premie dekking polis 2006 schade aanspraak
19. 2011 2008 2009 2012 2014 2015 den 2.1 haag 2.3

20. 2013 nl jaar fiscaal rechtbank winst wet vennootschapsbelasting eenheid verlies
21. stichting accountant vie d'or toezicht verzekeringskamer dhow actuaris edco titel
22. geld voorwerp afkomstig voorhand contant onmiddellijk rekening betalen enig bankrekening
23. onderdeel rov omstandigheid haar klacht onjuist stelling omdat onvoldoend maar
24. merken c. inbreuk publiek gebruik onderdeel dwangsom recht product rov
25. rapport deskundig ziekenhuis medisch behandeling haar gevaar arts onderzoek psychiatrisch
26. richtlijn nederland europees lidstaat recht justitie nederlands nationaal verordening vraag
27. verdacht feit schriftuur terechtzitting middeel sv raadsman voorstellen namens strafzaak
28. the to and a or that for be as on
29. betrokkeneen verklaring getuige 2010 2005 2006 medeverdachte verdacht horen proces-verbaal
30. rechtbank justitie officier beschikking klager beslag sv onderzoek ministerie klaagschrift
31. slachtoffer haar jaar sr seksueel aangeefsteren handeling afbeelding tenlastelegging ander
32. rechtbank bezwaar awb termijn vergoeding indienen griffierecht belanghebbennen verzet beschikking
33. gemeente college burgemeester verordening schip wethouder onroerend gemeentewet netbeheerder gebruik
34. verzoek rechtbank brief behandeling stuk rechter vonnis verschijnen rv terechtzitting
35. onderneming waarde economisch koper onroerend verkoop verkopen maatschap verkoper vennoot
36. goedeer betaling cdw uitnodiging vervoer invoer douanerecht delta accijn douaneautoriteit
37. heisterkamp winkel kleding toestel a.h.t machine haaglanden kas substantie kledingstuk
38. rechtbank gemeente onderdeel huurovereenkomst onteigennen deskundigen provincie onteigening schadeloosstelling bestemmingsplan

39. 2010 partij recht bw verdeling notaris erflater tussen echtgenoot akte
40. aandeel vennootschap aandeelhouder holding lening vermogen haar kapitaal b.v. rente
41. overeenkomst sluiten dexia 2005 aangaan sns appartement training vve rechtshandeling
42. curator bank vordering faillissement betaling fw schuldeiser recht bw schuldenaar
43. aangifte jaar inspecteur navorderingsaanslaag boete gegeven belastingdienst verhoging awr inzake
44. blz 1988 1989 1987 1990 1984 1986 1991 schilderij reaal
45. werknemer werkzaamheid werkgever bw arbeidsovereenkomst kantonrechter verrichten periode januari jaar
46. woning pand a-straat perceel eigenaar gebruik 2005 eigendom plaats liggen
47. aanvrager herziening aanvraag aanvrage onderzoek hetzij bekend sv wezen omstandigheid
48. verdacht verbalisant slachtoffer mijn proces-verbaal gaan politie toen verklaring man
49. wet toepassing bepaling regeling indien wetgever zoals ii tweede kamerstukken

D Randomly selected documents, topic distribution and intruding topics

ECLI:NL:HR:2014:2773
Topic 27 with probability: 0.573145687517
Topic 48 with probability: 0.132596476419
Topic 0 with probability: 0.0785602654741
The intruder is topic 43 with probability: 7.60456273764e-5

ECLI:NL:HR:2011:BP3968
Topic 48 with probability: 0.306716017739
Topic 29 with probability: 0.210125413003
Topic 13 with probability: 0.130291858646
The intruder is topic 16 with probability: 2.11193241816e-5

ECLI:NL:HR:2011:BP4800
Topic 16 with probability: 0.705261365128
Topic 10 with probability: 0.0761404667971
Topic 34 with probability: 0.0350571825724
The intruder is topic 35 with probability: 0.000100502512563

ECLI:NL:HR:2008:BD1842
Topic 1 with probability: 0.559763997418
Topic 23 with probability: 0.183253667928
Topic 45 with probability: 0.049369955002
The intruder is topic 34 with probability: 3.46620450607e-5

ECLI:NL:HR:2001:AB2151
Topic 38 with probability: 0.547494275354
Topic 1 with probability: 0.281252912003
Topic 23 with probability: 0.0886392833495
The intruder is topic 45 with probability: 4.6511627907e-5

ECLI:NL:HR:2002:AD9487
Topic 34 with probability: 0.28028803025
Topic 30 with probability: 0.261734035857
Topic 27 with probability: 0.225112321842
The intruder is topic 17 with probability: 4.3956043956e-5

ECLI:NL:HR:2014:3303
Topic 27 with probability: 0.410842525105
Topic 1 with probability: 0.291668872155
Topic 6 with probability: 0.0699734087716
The intruder is topic 2 with probability: 4.62962962963e-5

ECLI:NL:HR:1998:AA2396
Topic 49 with probability: 0.204364962073
Topic 3 with probability: 0.170184775934

Topic 43 with probability: 0.156498832333
The intruder is topic 31 with probability: 1.73611111111e-5

ECLI:NL:HR:2014:1303
Topic 27 with probability: 0.467418611633
Topic 0 with probability: 0.381484940455
Topic 23 with probability: 0.0412636831355
The intruder is topic 5 with probability: 4.08163265306e-5

ECLI:NL:HR:2013:BZ7150
Topic 27 with probability: 0.271993503787
Topic 2 with probability: 0.234830843774
Topic 48 with probability: 0.150411208954
The intruder is topic 42 with probability: 2.849002849e-5

Removed because it was 81RO: ECLI: ECLI:NL:HR:2010:BO0187.json