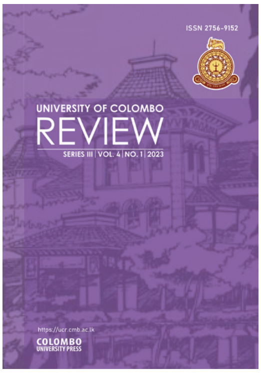

https://doi.org/10.4038/ucr.v4i1.116

University of Colombo Review (Series III),
Vol.4, No.1, 2023

Reading Supreme Courts from afar: Topic modelling judgements of the Supreme Courts of Sri Lanka and the United Kingdom

Sandani Yapa Abeywardena
Masters Student, Durham University; Affiliated Researcher, Digital Humanities Laboratory, University of Colombo

ABSTRACT

In legal scholarship, court judgments are pivotal in shaping jurisprudence. Situated within the field of digital humanities as it applies to legal texts, this article takes a closer look at the underlying themes in Supreme Court judgments by applying topic modelling to the judgments delivered by the Supreme Court of Sri Lanka (LKSC) and the United Kingdom Supreme Court (UKSC). Using two custom datasets curated by (web) scraping the respective LKSC and UKSC websites, this article employs Latent Dirichlet Allocation (LDA) with the Machine Learning for Language Toolkit (MALLET), a commonly used tool for topic modelling by digital humanists, to identify topics (themes) that represent the main areas of law in each jurisdiction primarily dealt with by the respective courts. 25 was selected as the number of topics after experimentation, and the topics identified in each jurisdiction were manually labelled. The results reveal the composition and evolution of judicial workloads, the shifting socio-political priorities in each jurisdiction, as well as the similarities and differences between the two courts. The findings have several implications for legal research and practice. These suggest that topic modelling can be used as a tool to organize and categorize judgments based on their themes, which can facilitate access and retrieval of relevant cases, and identify priority areas for judicial and legal training. They also challenge conventional legal taxonomies and classifications, and demonstrate the potential of computational methods for enhancing the understanding and analysis of law.

KEYWORDS:
Digital humanities; Latent Dirichlet Allocation (LDA); Judicial discourse; Supreme Court judgments; Topic modelling

Suggested Citation: Abeywardena, S.Y. (2023). Reading Supreme Courts from afar: Topic modelling judgements of the Supreme Courts of Sri Lanka and the United Kingdom. *University of Colombo Review* (New Series III), 4(1), 3-30

sandani.abeywardena@gmail.com

Introduction

> "Words are perhaps more important in law than in any other discipline"
>
> L.J.M. Cooray (1975, p. 533)

Language plays an undeniably significant role in judicial discourse. Much of judicial practice consists entirely of both oral and written language, from hearing evidence to writing judgments. Unsurprisingly, therefore, text corpora constitute the majority of legal corpora (Goźdź-Roszkowski, 2021) including corpora comprised of judgments. Such corpora can provide insight into judicial discourse in numerous ways. Situated within the field of digital humanities as it applies to legal texts, this article utilizes text mining and analytics to explore judicial discourse by topic modelling judgments delivered by the Supreme Courts of Sri Lanka and the United Kingdom. It explores how the respective judicial workloads are constructed in terms of the subject matter the court deals with, and how this workload has evolved (if at all) over time. The article further explores how topic modelling operates in relation to court judgments, specifically, how changing hyperparameters of the topic models (namely, number of topics) affects its output.

Topic modelling is an unsupervised machine learning method that aids in the discovery of topics contained in a selection of documents. It produces a "set of co-occurring words" that has the highest probability of occurring together in a given corpus which can be characterized as "topics" (Murakami et al., 2017, p. 243). For example, consider a corpus of research papers on sustainability. A topic modelling algorithm might find sets of co-occurring words such as "carbon", "greenhouse", "emissions", "geoengineering", and "pollutants" which highlights the presence of a theme in the corpus relating to climate change or climate science. This has led to topic modelling being described as "an attempt to inject semantic meaning into vocabulary" (Graham et al., 2012, para. 5) to uncover "evidence already in the text" (Brett, 2012, para. 23), and a form of "distant reading" (a term coined by Moretti (2000, 2013)) , "in the most pure sense: focused on corpora and not individual texts" (Meeks & Weingart, 2012, para. 1). While there are several algorithms and tools, this article adopts the Latent Dirichlet Allocation (LDA) algorithm. LDA, developed by Blei et al. (2003), is built upon Latent Semantic Analysis (LSA). While LSA groups documents by "semantic structures" in texts (Deerwester et al., 1990), LDA assumes that the words are drawn from an underlying (latent) topic (Blei et al., 2003), and has been used by scholars to explore various corpora from political agendas in Senate press releases (Grimmer, 2010), and academic discourse (Murakami et al., 2017), to policy analysis (DiMaggio et al., 2013) and twitter messages during the COVID-19 pandemic (Perera et al., 2022).

Topic modelling is also increasingly utilized in literary studies and digital humanities as well. Meeks and Weingart (2012) observe that introductions to topic modelling for (digital) humanists have increased in frequency since 2010. Indeed, the ability to conduct computational analysis of a large number of texts has the potential to shed light on patterns undiscernible through the close reading of an individual text. In this vein, Jockers and Mimno (2013) utilized topic modelling to extract themes (topics) from a corpus of 19th century literature comprised of 3200 British and American novels while Falk (2016) sought to combine both close reading and distant reading by topic modelling Amelia Opie's Adeline Mowbray. Similarly, Meeks (2011) used topic modelling as a tool to identify a definition of digital humanities from a corpus comprising of blogs and articles. Such work demonstrates the ability to explore topics that emerge in any text corpora, and open the possibilities for extending topic modelling to the legal domain.

While topic modelling has been deployed in numerous domains, exploration of its utility in law is ongoing with a few instances of topic modelling in similar settings existing in a few jurisdictions. For example, Carter et al. (2016) uses topic modelling to analyze 7476 judgments of the High Court of Australia (1903-2015) in an effort to examine the workload of the court. This appears to be the first instance topic modelling was applied to legal textual corpora. More recently, topic modelling has been deployed to analyze Brazilian and Czech Supreme Court judgments (Luz De Araujo & De Campos, 2020; Novotná et al., 2020). However, no such approach has been adopted towards the judgments delivered by the Sri Lankan Supreme Court, apart from using it as a tool in preliminary exploratory data analysis[1]. In the context of the United Kingdom, Izzidien et al. (2022) has examined the topics that arise in contract related judgments delivered by courts from 1709 to 2021. However, this is a general paper that does not focus solely on the Supreme Court, and further focuses only on contract related judgments[2]. Topic modelling, therefore, has so far been used on legal data in limited ways in Sri Lanka and the United Kingdom. Its use in other domains indicate that it can prove to be insightful in understanding how judicial discourse and workloads are constituted in these two jurisdictions.

Methodology

Dataset creation and pre-processing

Topic modelling has the potential to provide insight into judicial decision-making by uncovering broad thematic outlines of the contents of judgments. The respective Supreme Courts are selected (as opposed to lower court judgments) for two primary reasons. First, they are the highest courts in the judicial hierarchies

[1] See Hoole, E., Halliday, A., & Kumaraguru, Y. (2023). *Empirical study of Sri Lanka's apex courts: taking the first step with dataset collection and exploratory analysis.* See also, endnote iv.

[2] It is also a pre-print and has not been peer-reviewed at the time of writing.

in both jurisdictions. Since 2009, the UKSC is the final court of appeal in all civil matters in the UK, and for criminal matters in England, Wales, and Northern Ireland (Cowie, 2022). Similarly, the LKSC is the highest and final superior court in Sri Lanka and exercises final appellate and consultative jurisdiction in addition to jurisdiction in respect of constitutional, fundamental rights, electoral, and breach of parliamentary privilege matters (see Article 118, Constitution of Sri Lanka). Second, the UK legal system through colonization has also influenced the content and structure of the Sri Lankan judicial system (Cooray, 1975). Its influence is quite prominent in the domains of criminal, and administrative law, as well as the laws of procedure and evidence (Cooray, 1976). It also continues through the Civil Law Ordinance which provides that, in the absence of Sri Lankan legislation on the matter, English law applies in maritime and commercial matters in Sri Lanka[3]. Further, Therefore, a comparative examination of topic modelling of judgments from the two jurisdictions has the potential to provide insight into its respective judicial workloads.

However, in order to deploy topic modelling, it is necessary to have datasets (or corpora) of the judgments delivered by each Court in each jurisdiction. There are a number of publicly available datasets of judgments from various jurisdiction (for example, the SCOTUS dataset compiled by Alali et al. (2021), the Supreme Court of Nigeria Dataset compiled by Balogun et al. (2023), and the European Court of Human Rights dataset under the ECHR-OD dataset compiled by Quemy (2018)). However, at present, there are no publicly available datasets comprising the Supreme Court judgments in either Sri Lanka[4] or the United Kingdom.

Therefore, two collections of texts constituting judgments delivered by the Supreme Courts of each jurisdiction, were curated by scraping the respective websites of each court. The Supreme Court of Sri Lanka (LKSC) Dataset contains 933 judgments delivered by the LKSC and uploaded to the LKSC website from 2013-2020. The United Kingdom Supreme Court (UKSC) Dataset contains judgments delivered from 2009-2022. In total, 892 judgments were downloaded[5]. The respective datasets were then subject to text pre-processing.

[3] Sections 2, 3, and 4, Civil Law Ordinance, Sri Lanka. The Ordinance was adopted to "Introduce into Sri Lanka the law of England in certain cases, and to restrict the operation of the Kandyan law." Available at: https://www.lawnet.gov.lk/introduction-of-law-of-england-4/.

[4] There is a commendable ongoing effort to compile a similar dataset comprised of the judgments of the apex courts in Sri Lanka. This dataset is yet to be made public. For more information, see Hoole, E., Halliday, A., & Kumaraguru, Y. (2023). *Empirical study of Sri Lanka's apex courts: taking the first step with dataset collection and exploratory analysis.* Computational Analysis of Apex Courts (Workshop), *International Conference on Artificial Intelligence and Law* (ICAIL).

[5] Important: though 1031 cases are listed on the UKSC website, only 892 judgments were available to be downloaded. This is partially due to connected cases being consolidated, and being disposed of with one judgment.

In order to prepare the datasets for topic modelling with MALLET, the text in each dataset must be represented in a manner that is comprehensible to the computer. First, as both datasets were in PDF format, and MALLET requires texts to be in .csv or .txt format, both datasets were converted to .txt using a custom written code. Next, numbers and punctuation were removed as these non-alpha numeric characters would distort the results of the topic model. Further, stop-words (e.g., "a", "also", "at") which are common function words were removed as these words provide little important information to the topic model.

Topic modelling with MALLET

> "In digital humanities research we use tools, make tools, and theorize tools not because we are all information scientists, but because tools are the formal instantiation of methods. That is why MALLET often stands in for topic modelling and topic modelling often stands in for the digital humanities."
>
> Meeks and Weingart (2012, para. 20)

This article utilizes **M**achine **L**earning for **L**anguage **T**oolkit (MALLET),[6,7] developed by Andrew McCallum and a team of collaborators at University of Massachusetts Amherst for topic modelling. To develop its topic models, MALLET applies Gibbs sampling, a statistical method that efficiently generates a sample distribution (Graham et al., 2012). MALLET is considered to be one of two popular tools for topic modelling among humanists (Brett, 2012; Meeks & Weingart, 2012). The other tool is Paper Machines developed by Jo Guldi and Chris Johnson-Roberson. While Paper Machines actually uses MALLET to deploy LDA topic modelling, as of May 2017, it is no longer maintained[8]. Meeks and Weingart (2012) argue that scholars can certainly use MALLET uncritically as a tool of natural language processing (NLP) and accept its results, but further, view the results critically with a focus on what it obscures as much as it reveals. This insightful observation highlights the potential MALLET has for humanists who may use it to derive themes in a given corpus, but also approach its results as well as the process, critically.

MALLET is argued to be a very useful tool for humanists which can, however,

[6] The (python) code to deploy the topic models using Mallet and the time series analysis was adapted from Melanie Walsh's *Introduction to Cultural Analytics & Python*. This is an open source project available at https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/08-Topic-Modeling-Text-Files.html and https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/11-Topic-Modeling-Time-Series.html respectively. The code also relies on the little_mallet_wrapper developed by Maria Antoniak.

[7] The full (python) code deploying the topic models and the time series analysis in this article is available at https://github.com/Sandaniy/Papers.

[8] See the Paper Machines: Unmaintained Notice on Github at https://github.com/papermachines/papermachines.

be difficult for humanists to deploy. Graham and Milligan (2013) observes that, for example, installing MALLET requires "tacit knowledge for computer scientists" which can be "completely opaque for humanists" (para. 7). MALLET is usually deployed using the command line, a user interface that may not be familiar for humanists. Further, MALLET requires installation of the Java Developer's Kit[9], and on a Windows device, MALLET must be unzipped only in the C: drive. An environment variable must also be specified to indicate to the computer where it is located[10]. Additionally, while MALLET can be run also on Jupyter Notebooks in addition to the command line, in order to do so, the path to the unzipped MALLET folder on the device must be specified. However, when MALLET was deployed following these steps, the model generated only the training text and not the topic keys or distributions. To resolve this, the source folder, in addition to the (bin) path was defined[11]. UTF-8 compatibility issues also proved to be a challenge for MALLET. Even though the files were converted to .txt format with UTF-8, when uploaded to a Pandas DataFrame[12] in python for text cleaning and manipulation, alien characters were present. These characters could not be processed by MALLET. Therefore, the texts were further processed by converting to ASCII encoding, and discarding the incompatible characters. Therefore, debugging some of these issues may require technical familiarity, and for a novice user, sustained effort.

Ethical considerations

Court judgments are generally public. In the United Kingdom, court judgments are considered "court records" as defined by the First Schedule of the Public Records Act 1958 and are considered public records. In Sri Lanka, Article 106(1) of the Constitution of Sri Lanka guarantees the right of the public to attend court sittings. In both jurisdictions, the judgments are publicly available via the Supreme Court websites, and are generally used as public legal jurisprudence. They also do not come within personal data categories.[13]

Further consideration is necessary in terms of the legality of web scraping. At present, the laws on web scraping are globally sparse. Though sparse, the decision

[9] Java can be downloaded at: https://www.oracle.com/java/technologies/downloads/#jdk20-windows

[10] Detailed guidance on how to do this is available at: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

[11] "Quick_train_topic_model issues #7" available at https://github.com/maria-antoniak/little-mallet-wrapper/issues/7

[12] Pandas is a python library that allows users to store (two-dimensional) data in a table with rows and columns (called a dataframe). It is a widely used library and useful for storing, cleaning, manipulating data, both textual and numeric.

[13] There are certain categories of cases that are protected in the UK (for example, where it involves a child). In such cases, the UKSC itself removes the name of the child in the judgment. This practice is not consistently followed in Sri Lanka.

of the US Court of Appeals protecting the scraping of public data (see hiQ Labs, Inc v. LinkedIn Corp), while not applicable in the UK or Sri Lanka, provides some direction on how legal systems may approach web scraping at present. It should be noted, however, that the law regarding this domain is subject to rapid change and development with the deployment of generative artificial intelligence (AI) in the last year. In PM v. OpenAI (2023), a class action lawsuit was filed against Open AI and Microsoft on the basis that its popular generative AI programs, ChatGPT and DALL-E used private information including personally identifiable information, which were scraped illegally and without adequate permission, as data to train the models. This was closely followed by J.L v. Alphabet Inc (2023) where a lawsuit was filed against Google, Google DeepMind, and Alphabet alleging copyright violation through unauthorized web scraping. Both these cases were filed in San Fransico in June and July 2023. The outcome of these cases could have significant implications for the use of web scraping as a tool to gather information, and the governance of such data.

Limitations

It is important to note that the respective corpora represent only a component of the judicial workload. For example, the LKSC also provide judicial opinions to the Parliament on upcoming bills, issue numerous orders and injunctions etc. which are not reflected in corpora comprised of judgments. Further, the size of the corpus, in comparison to other corpora which has topic modelling deployed, is quite small.[14] Large corpora is generally preferred as topic modelling is built for large collections of texts and Brett (2012) recommends texts "at least in the hundreds" (para. 9). In the present article, there are 892 UKSC judgments and 933 LKSC judgments, and it is noted that this limitation is attributable to the nature of the data because it is limited to the digitized judgments presently available. The article also opts not to remove domain-specific (technical) stop-words which may impact interpretability of the topics.

The Topic models

No. of Topics

A hyperparameter (an adjustable parameter in machine learning algorithms) in LDA is the number of topics the model should identify in the corpus i.e., if the number of topics is given as 10, the topic model will generate 10 topics. However, the suitable number of topics is generally unknown (Yau et al., 2014): a very small number may generate topics that are too broad to be meaningful, while a large

[14] For example, in *Mining the Dispatch*, Nelson utilised a corpus of 112,000 documents.

number may break down themes so much that it becomes meaningless (Sbalchiero & Eder, 2020). Therefore, selecting an appropriate number of topics remain one of the challenges of topic modelling (Weston et al., 2023), and the choice to select a specific number remains a "qualitative task" (Suominen & Toivanen, 2016, p. 2475). One of the approaches to identify the suitable number of topics has been, therefore, trial-and-error. Yau et al. (2014), for example, tested a range of topics and chose 50 as it was more manageable while Carter et al. (2016) produced models with 10, 15, 20, 50, and 100 topics, and selected 10 and 50. Similarly, Suominen and Toivanen (2016) selected 60 topics after trial-and-error. Following these practices, for this article, the models were trained for 5, 10, 25, 50, and 100 topics to explore how the model clusters words as you increase the number of topics. Admittedly, the selection of the "best" no. of topics is a subjective exercise, and it was observed that both 25 and 50 provided greater granularity in the topics identified while remaining interpretable. Therefore, 25 topics were selected for both the datasets prioritizing interpretability and ease, and the topic models retrained on each dataset.

Manually labelling Topics

While Mallet produces a given number of topics and the words that have the highest probability of occurring together, it does not label the topic. Therefore, a label was manually assigned to each topic. While some topics were easier to label (e.g., arbitration), others were generic as it referred to generic terms that recur in judgments (e.g., "petitioner", "respondent", "authority", "said", "filed", "may"). Further, some topics were mixed (e.g., "procedure", "section", "civil", "disciplinary"). Where identifiable, the topics were labelled in relation to a specific area (or sub-area) or law, and as generic if generic terms were present in the topic model.

Results and discussion

Overview of judicial workload

The LKSC delivered 933 judgments during 2013-2020 with the highest number of judgments delivered in 2017, and the lowest in 2013. The UKSC delivered 892 judgments during 2009-2022 with the highest number of judgments delivered in 2017, and the lowest in 2009. Coincidentally, both courts delivered the highest number of judgments in 2017.

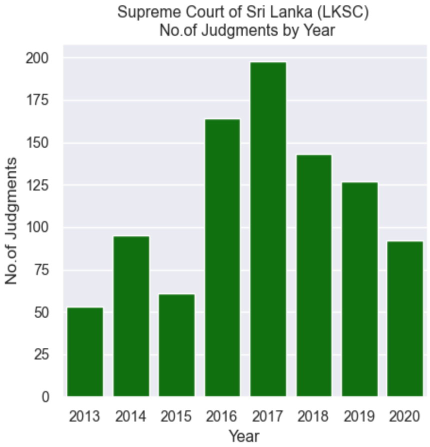

Figure 1. *No. of judgments by year: LKSC*

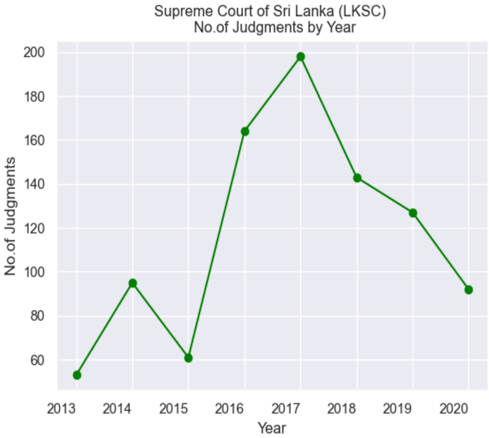

Figure 2. *No. of judgments by year: LKSC*

Since 2017, there has been a general decline in the no. of judgments delivered in both jurisdictions. Unsurprisingly, both jurisdictions delivered comparatively fewer judgments in 2020: the UKSC delivered only 43 judgments while the LKSC delivered 92 judgments, presumably due to challenges precipitated by COVID-19.

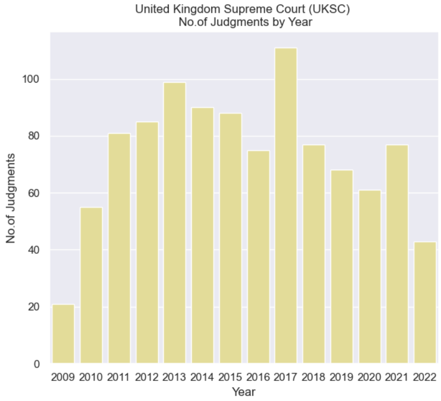

Figure 3. *No. of judgments by year: UKSC*

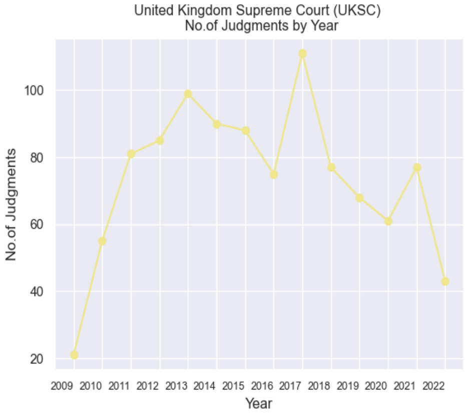

Figure 4. *No. of judgments by year: UKSC*

Even though the number of judgments is similar, on average, as seen in Table 1, UKSC judgments tend to be 4 times longer than the LKSC judgments. This significant difference in length is also reflected in the vocabulary size as well where the UKSC vocabulary is 35% larger than the LKSC vocabulary.

Table 1. Summary statistics of the datasets

Court	No. of Judgments	Mean No. of Words per Judgment	Vocabulary Size
LKSC	933	1659.5	50409
UKSC	892	7000.8	78293

Impact of number of topics

As the number of topics increased, there was increased granularity in topic identification. In the initial 5-topic model, topics were difficult to delineate except for Topic 3 (UKSC) which included matters of international and regional law. The remaining topics were either composed of general vocabulary relating to judicial proceedings (“evidence”, “case”, “action”) and parties (“appellant”, “respondent”, “defendant”), or collated several areas of law (e.g., company, constitutional, land) and provided little insight into the judicial workload.

Topic 0

[‘would’, ‘case’, ‘lord’, ‘section’, ‘para’, ‘court’, ‘act’, ‘appeal’, ‘page’, ‘land’, ‘use’, ‘may’, ‘one’, ‘whether’, ‘right’, ‘part’, ‘could’, ‘terms’, ‘also’, ‘made’]

Topic 1

[‘court’, ‘case’, ‘lord’, ‘would’, ‘appeal’, ‘para’, ‘section’, ‘evidence’, ‘act’, ‘whether’, ‘order’, ‘page’, ‘may’, ‘made’, ‘article’, ‘criminal’, ‘decision’, ‘proceedings’, ‘public’, ‘right’]

Topic 2

[‘would’, ‘court’, ‘case’, ‘para’, ‘state’, ‘secretary’, ‘section’, ‘appeal’, ‘child’, ‘act’, ‘article’, ‘page’, ‘decision’, ‘lord’, ‘whether’, ‘may’, ‘children’, ‘rights’, ‘also’, ‘person’]

Topic 3

[‘law’, ‘court’, ‘article’, ‘state’, ‘act’, ‘section’, ‘jurisdiction’, ‘would’, ‘para’, ‘united’, ‘case’, ‘proceedings’, ‘page’, ‘states’, ‘convention’, ‘member’, ‘may’, ‘within’, ‘international’, ‘courts’]

Topic 4

[‘would’, ‘lord’, ‘case’, ‘law’, ‘company’, ‘claim’, ‘section’, ‘para’, ‘court’, ‘liability’, ‘page’, ‘tax’, ‘may’, ‘appeal’, ‘act’, ‘made’, ‘loss’, ‘whether’, ‘contract’, ‘one’]

Figure 5. *UKSC 5-topic model: Topics*

Topic 0

[‘respondent’, ‘court’, ‘appellant’, ‘bank’, ‘defendant’, ‘plaintiff’, ‘high’, ‘labour’, ‘tribunal’, ‘said’, ‘evidence’, ‘judge’, ‘learned’, ‘company’, ‘case’, ‘agreement’, ‘applicant’, ‘action’, ‘appeal’, ‘judgment’]

Topic 1

[‘plaintiff’, ‘defendant’, ‘court’, ‘land’, ‘respondent’, ‘district’, ‘appellant’, ‘case’, ‘evidence’, ‘judge’, ‘said’, ‘property’, ‘deed’, ‘high’, ‘action’, ‘appeal’, ‘title’, ‘learned’, ‘law’, ‘judgment’]

Topic 2

[‘petitioner’, ‘petitioners’, ‘respondent’, ‘respondents’, ‘service’, ‘marked’, ‘public’, ‘said’, ‘application’, ‘marks’, ‘commission’, ‘colombo’, ‘sri’, ‘general’, ‘rights’, ‘article’, ‘court’, ‘constitution’, ‘lanka’, ‘school’]

Topic 3

[‘petitioner’, ‘police’, ‘respondent’, ‘accused’, ‘court’, ‘evidence’, ‘station’, ‘case’, ‘said’, ‘respondents’, ‘magistrate’, ‘appellant’, ‘also’, ‘person’, ‘officer’, ‘general’, ‘learned’, ‘made’, ‘law’, ‘taken’]

Topic 4

[‘court’, ‘appeal’, ‘section’, ‘respondent’, ‘order’, ‘application’, ‘law’, ‘act’, ‘case’, ‘petitioner’, ‘supreme’, ‘made’, ‘learned’, ‘said’, ‘high’, ‘filed’, ‘judgment’, ‘judge’, ‘provisions’, ‘counsel’]

Figure 6. *LKSC 5-topic model: Topics*

However, a majority of topics developed greater specificity as they increased. In the 10-topic models, more distinct topics (e.g., related to criminal, employment, land law etc.) are identified in both datasets.

Topic 2

[‘land’, ‘plaintiff’, ‘defendant’, ‘court’, ‘title’, ‘plan’, ‘respondent, ‘district’, ‘said’, ‘case’, ‘evidence’, ‘judge’, ‘action’, ‘high’, ‘lot’, ‘possession’, ‘partition’, ‘marked’, ‘deed’, ‘appellant’]

Figure 7. *LKSC 10-topic model: Topic 2[“Land law”]*

Topic 5

[‘court’, ‘case’, ‘lord’, ‘criminal’, ‘article’, ‘para’, ‘police’, ‘evidence’, ‘would’, ‘appeal’, ‘whether’, ‘section’, ‘offence’, ‘trial’, ‘page’, ‘right’, ‘person’, ‘may’, ‘said’, ‘act’]

Figure 8. *UKSC 10-topic model: Topic 5 [“Criminal law”]*

Single topics in the 5-Topic models were also split into two (e.g., international law divided into international and EU law, and domestic Parliamentary concerns in the UKSC dataset):

Topic 0

[‘section’, ‘act’, ‘court’, ‘law’, ‘would’, ‘order’, ‘parliament’, ‘lord’, ‘decision’, ‘para’, ‘public’, ‘page’, ‘may’, ‘case’, ‘rights’, ‘made’, ‘power’, ‘appeal’, ‘information’, ‘legislation’]

Topic 2

[‘law’, ‘article’, ‘state’, ‘court’, ‘jurisdiction’, ‘international’, ‘united’, ‘convention’, ‘states’, ‘proceedings’, ‘foreign’, ‘para’, ‘case’, ‘english’, ‘agreement’, ‘kingdom’, ‘arbitration’, ‘courts’, ‘would’, ‘within’]

Figure 9. *UKSC 10-topic model:*
Topic 0 [“Domestic Parliamentary Affairs including devolution”] and Topic 2 [“International law”]

This increasing granularity was present in the 25-topic and 50 topic models as well which in turn aided identification of sub-areas in specific area of law, and the identification of key areas that emerge in a specific area of law. For example, Topic 1 (25-topic model-LKSC) relating to land featured “possession”, “rent”, “permit”, and “occupation” which provide insight into key legal questions that recur in land law.

Topic 1

[‘premises’, ‘property’, ‘possession’, ‘title’, ‘owner’, ‘action’, ‘land’, ‘rent’, ‘tenant’, ‘person’, ‘house’, ‘death’, ‘deceased’, ‘permit’, ‘substituted’, ‘said’, ‘father’, ‘occupation’, ‘respondent’, ‘right’]

Figure 10. *LKSC 25-topic model: Topic 1 [“Land law”]*

Topic 4

[‘vessel’, ‘loss’, ‘clause’, ‘insurance’, ‘owners’, ‘insured’, ‘goods’, ‘damage’, ‘policy’, ‘court’, ‘cargo’, ‘appeal’, ‘rule’, ‘rules’, ‘ltd’, ‘insurers’, ‘owner’, ‘course’, ‘time’, ‘page’]

Figure 11. *UKSC 25-topic model: Topic 4 [“Maritime - Shipping law”]*

However, in the 100-topic mode, the outputs differ. The UKSC outputs provided even greater specificity: in Topic 41, key legal principles in arbitration from party autonomy (“choice”), “seat” of arbitration, to the final “award” and its “enforcement” were identifiable.

Topic 41

[‘arbitration’, ‘law’, ‘agreement’, ‘award’, ‘parties’, ‘contract’, ‘english’, ‘party’, ‘choice’, ‘arbitrator’, ‘arbitrators’, ‘international’, ‘proper’, ‘article’, ‘seat’, ‘enforcement’, ‘arbitral’, ‘convention’, ‘governed’, ‘clause’]

Figure 12. *UKSC 100-topic model: Topic 41 [“Arbitration law”]*

In the LKSC model, increased specificity pushes beyond identifiable areas (and sub-areas) of law. For example, Topics 0 and 2 are highly generic while Topic 10 (quite interestingly) contains words that are mostly place names. It appears, therefore, that MALLET is able, to an extent, understand words in context and identify place names.

Topic 0

[‘plaintiff’, ‘defendant’, ‘judge’, ‘plaint’, ‘appeal’, ‘court’, ‘trial’, ‘action’, ‘district’, ‘high’, ‘learned’, ‘held’, ‘case’, ‘issue’, ‘answer’, ‘evidence’, ‘stated’, ‘issues’, ‘also’, ‘dismissed’]

Topic 1

[‘commission’, ‘member’, ‘public’, ‘secretary’, ‘colombo’, ‘ministry’, ‘department’, ‘service’, ‘road’, ‘general’, ‘national’, ‘director’, ‘chairman’, ‘excise’, ‘former’, ‘development’, ‘nawala’, ‘letter’, ‘mawatha’, ‘narahenpita’]

Topic 2

[‘appellants’, ‘appeal’, ‘feet’, ‘stated’, ‘case’, ‘road’, ‘hewage’, ‘shop’, ‘room’, ‘matugama’, ‘wall’, ‘upon’, ‘substituted’, ‘two’, ‘main’, ‘street’, ‘questions’, ‘tac’, ‘amended’, ‘godellawaththage’]

Topic 3

[‘deceased’, ‘action’, ‘substituted’, ‘death’, ‘law’, ‘husband’, ‘substitution’, ‘person’, ‘case’, ‘ingratitude’, ‘original’, ‘place’, ‘cause’, ‘right’, ‘died’, ‘gift’, ‘donor’, ‘heirs’, ‘litis’, ‘contestatio’]

Topic 10

[‘thalgaswala’, ‘page’, ‘judgment’, ‘galle’, ‘appeal’, ‘nagoda’, ‘kahaduwa’, ‘mapalagama’, ‘maththaka’, ‘compensation’, ‘niyagama’, ‘central’, ‘kumara’, ‘employees’, ‘aluthihala’, ‘porawagama’, ‘road’, ‘manampitiya’, ‘kumari’, ‘respondents’]

Figure 13. *LKSC 100-topic model: Topic 0 [“generic”], Topic 1[“State administrative structures”], Topic 2[generic], Topic 3 [“Succession”], and Topic 10[“place names”]*

Examination of judicial workloads: 25-Topic models

Having selected the 25-Topic model for both jurisdictions for reasons of interpretability and ease as discussed above, the topics were manually labelled as seen in Tables 2 (LKSC) and Table 3 (UKSC) below.

Table 2. LKSC 25-Topic model with top words and (manual) labels

No.	Top Words of each Topic	Label
Topic 0	[‘petitioner’, ‘respondent’, ‘marked’, ‘letter’, ‘officer’, ‘dated’, ‘respondents’, ‘inquiry’, ‘authority’, ‘disciplinary’, ‘report’, ‘made’, ‘colombo’, ‘said’, ‘charge’, ‘general’, ‘decision’, ‘submitted’, ‘documents’, ‘issued’]	Generic; Administrative
Topic 1	[‘vehicle’, ‘agreement’, ‘arbitration’, ‘award’, ‘lease’, ‘section’, ‘party’, ‘policy’, ‘insurance’, ‘arbitral’, ‘person’, ‘said’, ‘parties’, ‘owner’, ‘jurisdiction’, ‘act’, ‘clause’, ‘tribunal’, ‘motor’, ‘driver’]	Arbitration
Topic 2	[‘act’, ‘section’, ‘appeal’, ‘said’, ‘land’, ‘order’, ‘provisions’, ‘state’, ‘law’, ‘made’, ‘council’, ‘commissioner’, ‘ordinance’, ‘application’, ‘shall’, ‘minister’, ‘gazette’, ‘respondent’, ‘court’, ‘authority’]	Administrative
Topic 3	[‘may’, ‘also’, ‘whether’, ‘section’, ‘made’, ‘would’, ‘must’, ‘upon’, ‘thus’, ‘stated’, ‘circumstances’, ‘law’, ‘case’, ‘view’, ‘aforesaid’, ‘however’, ‘set’, ‘issue’, ‘regard’, ‘act’]	Generic
Topic 4	[‘respondent’, ‘contract’, ‘documents’, ‘colombo’, ‘tender’, ‘lanka’, ‘electricity’, ‘loss’, ‘company’, ‘sri’, ‘goods’, ‘damages’, ‘customs’, ‘duty’, ‘board’, ‘marked’, ‘letter’, ‘procurement’, ‘bid’, ‘ltd’]	Contract; Administrative; Customs
Topic 5	[‘marks’, ‘petitioner’, ‘school’, ‘petitioners’, ‘said’, ‘education’, ‘circular’, ‘admission’, ‘application’, ‘children’, ‘college’, ‘child’, ‘board’, ‘vidyalaya’, ‘interview’, ‘clause’, ‘residence’, ‘respondent’, ‘grade’, ‘schools’]	Constitutional; Fundamental Rights
Topic 6	[‘labour’, ‘tribunal’, ‘applicant’, ‘employer’, ‘employee’, ‘order’, ‘workman’, ‘employment’, ‘termination’, ‘industrial’, ‘employees’, ‘compensation’, ‘services’, ‘president’, ‘respondent’, ‘evidence’, ‘act’, ‘application’, ‘company’, ‘disputes’]	Labor
Topic 7	[‘board’, ‘company’, ‘act’, ‘property’, ‘injunction’, ‘interim’, ‘conciliation’, ‘respondent’, ‘trade’, ‘name’, ‘work’, ‘order’, ‘rights’, ‘settlement’, ‘directors’, ‘word’, ‘section’, ‘high’, ‘colombo’, ‘companies’]	Company
Topic 8	[‘plaintiff’, ‘defendant’, ‘district’, ‘action’, ‘court’, ‘case’, ‘evidence’, ‘trial’, ‘defendants’, ‘judge’, ‘civil’, ‘plaint’, ‘plaintiffs’, ‘judgment’, ‘high’, ‘learned’, ‘parties’, ‘substituted’, ‘filed’, ‘appeal’]	Generic
Topic 9	[‘land’, ‘premises’, ‘possession’, ‘property’, ‘title’, ‘owner’, ‘rent’, ‘permit’, ‘tenant’, ‘section’, ‘person’, ‘said’, ‘ordinance’, ‘house’, ‘occupation’, ‘act’, ‘father’, ‘respondent’, ‘entitled’, ‘action’]	Land
Topic 10	[‘court’, ‘appeal’, ‘order’, ‘application’, ‘section’, ‘procedure’, ‘petitioner’, ‘respondent’, ‘supreme’, ‘made’, ‘filed’, ‘case’, ‘leave’, ‘act’, ‘high’, ‘civil’, ‘code’, ‘law’, ‘petition’, ‘shall’]	Generic; Civil procedure
Topic 11	[‘bank’, ‘said’, ‘sum’, ‘debt’, ‘letter’, ‘action’, ‘agreement’, ‘money’, ‘account’, ‘payment’, ‘loan’, ‘company’, ‘pay’, ‘guarantee’, ‘due’, ‘demand’, ‘commercial’, ‘marked’, ‘ltd’, ‘amount’]	Contract; Banking; Commercial

Topic 12	[‘attorney’, ‘affidavit’, ‘parte’, ‘default’, ‘judgment’, ‘page’, ‘registered’, ‘tea’, ‘thalgaswala’, ‘law’, ‘proxy’, ‘complainant’, ‘company’, ‘date’, ‘sri’, ‘summons’, ‘evidence’, ‘appeal’, ‘galle’, ‘ahangama’]	Generic; Procedure
Topic 13	[‘service’, ‘public’, ‘post’, ‘officers’, ‘petitioners’, ‘member’, ‘commission’, ‘grade’, ‘class’, ‘department’, ‘scheme’, ‘interview’, ‘said’, ‘appointment’, ‘ministry’, ‘secretary’, ‘circular’, ‘sri’, ‘iii’, ‘examination’]	Administrative; Labor
Topic 14	[‘evidence’, ‘maintenance’, ‘fuu’, ‘magistrate’, ‘witness’, ‘trial’, ‘meusks’, ‘temple’, ‘thero’, ‘respondents’, ‘ska’, ‘wkqj’, ‘lrk’, ‘wxl’, ‘keye’, ‘marriage’, ‘lsh’, ‘slre’, ‘sabha’, ‘whs’]	Generic
Topic 15	[‘police’, ‘petitioner’, ‘respondent’, ‘station’, ‘respondents’, ‘petitioners’, ‘arrest’, ‘officer’, ‘article’, ‘officers’, ‘said’, ‘person’, ‘constitution’, ‘arrested’, ‘medical’, ‘hospital’, ‘magistrate’, ‘custody’, ‘rights’, ‘fundamental’]	Constitutional [Fundamental rights]; Criminal
Topic 16	[‘article’, ‘constitution’, ‘parliament’, ‘president’, ‘power’, ‘court’, ‘state’, ‘sri’, ‘lanka’, ‘general’, ‘respondent’, ‘jurisdiction’, ‘government’, ‘attorney’, ‘law’, ‘powers’, ‘election’, ‘petitioner’, ‘members’, ‘shall’]	Constitutional
Topic 17	[‘deed’, ‘property’, ‘transfer’, ‘land’, ‘evidence’, ‘trust’, ‘respondent’, ‘ordinance’, ‘notary’, ‘sale’, ‘interest’, ‘said’, ‘title’, ‘agreement’, ‘executed’, ‘public’, ‘gift’, ‘money’, ‘dated’, ‘loan’]	Land; Trust
Topic 18	[‘court’, ‘appellant’, ‘respondent’, ‘appeal’, ‘judge’, ‘high’, ‘learned’, ‘said’, ‘case’, ‘judgment’, ‘law’, ‘dated’, ‘supreme’, ‘evidence’, ‘appellants’, ‘question’, ‘referred’, ‘consider’, ‘counsel’, ‘questions’]	Generic
Topic 19	[‘medical’, ‘university’, ‘expectation’, ‘slmc’, ‘legitimate’, ‘degree’, ‘marked’, ‘medicine’, ‘council’, ‘section’, ‘saitm’, ‘court’, ‘respondent’, ‘ordinance’, ‘authority’, ‘education’, ‘universities’, ‘lanka’, ‘institute’, ‘registration’]	Fundamental rights; Administrative
Topic 20	[‘power’, ‘station’, ‘respondent’, ‘thermal’, ‘respondents’, ‘oil’, ‘water’, ‘cea’, ‘public’, ‘regulations’, ‘act’, ‘railway’, ‘boi’, ‘project’, ‘area’, ‘level’, ‘marked’, ‘part’, ‘persons’, ‘chunnakam’]	Environmental
Topic 21	[‘accused’, ‘evidence’, ‘trial’, ‘section’, ‘witness’, ‘prosecution’, ‘offence’, ‘code’, ‘sentence’, ‘penal’, ‘criminal’, ‘appellant’, ‘learned’, ‘magistrate’, ‘charge’, ‘deceased’, ‘complainant’, ‘attorney’, ‘counsel’, ‘conviction’]	Criminal
Topic 22	[‘petitioners’, ‘application’, ‘respondents’, ‘court’, ‘rights’, ‘colombo’, ‘fundamental’, ‘article’, ‘constitution’, ‘commission’, ‘petitioner’, ‘supreme’, ‘road’, ‘general’, ‘time’, ‘said’, ‘petition’, ‘filed’, ‘attorney’, ‘counsel’]	Constitutional; Fundamental Rights
Topic 23	[‘land’, ‘plan’, ‘lot’, ‘partition’, ‘marked’, ‘title’, ‘deed’, ‘share’, ‘surveyor’, ‘said’, ‘defendants’, ‘schedule’, ‘way’, ‘described’, ‘right’, ‘extent’, ‘corpus’, ‘plaint’, ‘district’, ‘road’]	Land
Topic 24	[‘court’, ‘case’, ‘law’, ‘time’, ‘given’, ‘one’, ‘even’, ‘would’, ‘supreme’, ‘taken’, ‘also’, ‘two’, ‘well’, ‘judge’, ‘fact’, ‘cannot’, ‘due’, ‘matter’, ‘could’, ‘without’]	Generic

Table 3. UKSC 25-Topic model with top words and (manual) labels

No.	Top Words	Label
Topic 0	[‘duty’, ‘loss’, ‘liability’, ‘damage’, ‘care’, ‘vessel’, ‘lord’, ‘negligence’, ‘liable’, ‘caused’, ‘ltd’, ‘tort’, ‘cause’, ‘appeal’, ‘act’, ‘policy’, ‘breach’, ‘owners’, ‘risk’, ‘law’]	Shipping; Tort

Topic 1	[‘public’, ‘article’, ‘rights’, ‘information’, ‘para’, ‘life’, ‘person’, ‘right’, ‘private’, ‘interference’, ‘data’, ‘protection’, ‘treatment’, ‘interest’, ‘convention’, ‘human’, ‘health’, ‘patient’, ‘court’, ‘relevant’]	Human Rights
Topic 2	[‘detention’, ‘court’, ‘article’, ‘sentence’, ‘extradition’, ‘decision’, ‘state’, ‘secretary’, ‘release’, ‘arrest’, ‘warrant’, ‘judicial’, ‘period’, ‘authority’, ‘case’, ‘board’, ‘prison’, ‘detained’, ‘person’, ‘prisoners’]	Immigration
Topic 3	[‘secretary’, ‘state’, ‘appeal’, ‘immigration’, ‘rules’, ‘decision’, ‘asylum’, ‘leave’, ‘home’, ‘application’, ‘tribunal’, ‘country’, ‘rule’, ‘person’, ‘united’, ‘appellant’, ‘remain’, ‘department’, ‘applicant’, ‘kingdom’]	Immigration
Topic 4	[‘section’, ‘act’, ‘parliament’, ‘order’, ‘provision’, ‘provisions’, ‘law’, ‘power’, ‘legislation’, ‘effect’, ‘part’, ‘statutory’, ‘made’, ‘within’, ‘subsection’, ‘scotland’, ‘scottish’, ‘would’, ‘sections’, ‘page’]	Constitutional [Legislative power]
Topic 5	[‘discrimination’, ‘regulations’, ‘act’, ‘treatment’, ‘benefits’, ‘jewish’, ‘grounds’, ‘asbestos’, ‘sex’, ‘religious’, ‘benefit’, ‘policy’, ‘persons’, ‘appeal’, ‘section’, ‘pension’, ‘social’, ‘status’, ‘group’, ‘women’]	Human Rights
Topic 6	[‘company’, ‘trust’, ‘creditors’, ‘directors’, ‘duty’, ‘insolvency’, ‘trustee’, ‘rule’, ‘assets’, ‘liability’, ‘companies’, ‘trustees’, ‘interests’, ‘ltd’, ‘companys’, ‘director’, ‘debt’, ‘liquidation’, ‘act’, ‘law’]	Company
Topic 7	[‘planning’, ‘development’, ‘permission’, ‘authority’, ‘plan’, ‘decision’, ‘local’, ‘secretary’, ‘council’, ‘site’, ‘public’, ‘application’, ‘state’, ‘policy’, ‘scheme’, ‘relevant’, ‘appeal’, ‘report’, ‘proposed’, ‘page’]	Administrative
Topic 8	[‘criminal’, ‘court’, ‘section’, ‘offence’, ‘appeal’, ‘defendant’, ‘evidence’, ‘conviction’, ‘trial’, ‘offences’, ‘justice’, ‘order’, ‘convicted’, ‘confiscation’, ‘prosecution’, ‘crime’, ‘article’, ‘lord’, ‘person’, ‘conduct’]	Criminal
Topic 9	[‘disease’, ‘insurance’, ‘liability’, ‘injury’, ‘risk’, ‘exposure’, ‘mesothelioma’, ‘lord’, ‘employers’, ‘asbestos’, ‘policy’, ‘caused’, ‘period’, ‘insurers’, ‘would’, ‘causation’, ‘insured’, ‘para’, ‘employer’, ‘fairchild’]	Tort; Labor; Insurance
Topic 10	[‘claim’, ‘court’, ‘law’, ‘action’, ‘claims’, ‘claimant’, ‘costs’, ‘appeal’, ‘proceedings’, ‘damages’, ‘claimants’, ‘defendant’, ‘rule’, ‘judgment’, ‘legal’, ‘time’, ‘case’, ‘limitation’, ‘solicitors’, ‘made’]	[Generic]
Topic 11	[‘would’, ‘case’, ‘lord’, ‘para’, ‘court’, ‘whether’, ‘one’, ‘could’, ‘may’, ‘question’, ‘law’, ‘page’, ‘said’, ‘view’, ‘first’, ‘must’, ‘however’, ‘made’, ‘also’, ‘see’]	[Generic]
Topic 12	[‘child’, ‘children’, ‘court’, ‘family’, ‘care’, ‘order’, ‘parents’, ‘mother’, ‘local’, ‘interests’, ‘father’, ‘home’, ‘authority’, ‘rights’, ‘appeal’, ‘best’, ‘judge’, ‘para’, ‘also’, ‘residence’]	Family
Topic 13	[‘state’, ‘international’, ‘united’, ‘convention’, ‘law’, ‘article’, ‘states’, ‘foreign’, ‘immunity’, ‘kingdom’, ‘jurisdiction’, ‘court’, ‘government’, ‘act’, ‘acts’, ‘within’, ‘armed’, ‘forces’, ‘authority’, ‘rights’]	International
Topic 14	[‘patent’, ‘appeal’, ‘product’, ‘judgment’, ‘court’, ‘use’, ‘claim’, ‘infringement’, ‘products’, ‘point’, ‘cat’, ‘patents’, ‘para’, ‘invention’, ‘competition’, ‘market’, ‘evidence’, ‘page’, ‘part’, ‘process’]	Intellectual Property
Topic 15	[‘would’, ‘lord’, ‘case’, ‘said’, ‘tion’, ‘court’, ‘must’, ‘para’, ‘may’, ‘one’, ‘question’, ‘fact’, ‘whether’, ‘could’, ‘issue’, ‘way’, ‘page’, ‘cases’, ‘right’, ‘first’]	[Generic]
Topic 16	[‘land’, ‘right’, ‘use’, ‘public’, ‘rights’, ‘act’, ‘property’, ‘owner’, ‘appeal’, ‘registration’, ‘statutory’, ‘value’, ‘authority’, ‘grant’, ‘council’, ‘lord’, ‘compensation’, ‘water’, ‘purposes’, ‘part’]	Land; Administrative, Human rights

Topic 17	[‘employment’, ‘work’, ‘employer’, ‘time’, ‘employee’, ‘contract’, ‘employees’, ‘dismissal’, ‘tribunal’, ‘appeal’, ‘employers’, ‘workers’, ‘terms’, ‘regulations’, ‘worker’, ‘working’, ‘part’, ‘period’, ‘employed’, ‘noise’]	Labor
Topic 18	[‘court’, ‘appeal’, ‘decision’, ‘order’, ‘tribunal’, ‘review’, ‘proceedings’, ‘judicial’, ‘judgment’, ‘application’, ‘made’, ‘material’, ‘case’, ‘procedure’, ‘public’, ‘hearing’, ‘evidence’, ‘jurisdiction’, ‘courts’, ‘power’]	[Generic]
Topic 19	[‘police’, ‘court’, ‘evidence’, ‘article’, ‘right’, ‘para’, ‘rights’, ‘trial’, ‘investigation’, ‘convention’, ‘strasbourg’, ‘legal’, ‘case’, ‘whether’, ‘proceedings’, ‘time’, ‘judgment’, ‘act’, ‘made’, ‘given’]	Human Rights; EU
Topic 20	[‘contract’, ‘clause’, ‘agreement’, ‘parties’, ‘property’, ‘money’, ‘terms’, ‘payment’, ‘bank’, ‘interest’, ‘appeal’, ‘party’, ‘sale’, ‘would’, ‘pay’, ‘value’, ‘price’, ‘sum’, ‘services’, ‘part’]	Contract; Banking
Topic 21	[‘possession’, ‘tenant’, ‘tenancy’, ‘section’, ‘act’, ‘notice’, ‘landlord’, ‘accommodation’, ‘premises’, ‘order’, ‘part’, ‘right’, ‘court’, ‘housing’, ‘appeal’, ‘authority’, ‘building’, ‘code’, ‘occupation’, ‘para’]	Land
Topic 22	[‘tax’, ‘value’, ‘vat’, ‘section’, ‘hmrc’, ‘revenue’, ‘income’, ‘paid’, ‘scheme’, ‘amount’, ‘services’, ‘appeal’, ‘company’, ‘relevant’, ‘business’, ‘period’, ‘account’, ‘year’, ‘para’, ‘goods’]	Taxation
Topic 23	[‘law’, ‘jurisdiction’, ‘court’, ‘proceedings’, ‘arbitration’, ‘english’, ‘agreement’, ‘article’, ‘parties’, ‘foreign’, ‘contract’, ‘england’, ‘judgment’, ‘claims’, ‘courts’, ‘claim’, ‘appeal’, ‘case’, ‘rule’, ‘award’]	Arbitration
Topic 24	[‘article’, ‘law’, ‘member’, ‘court’, ‘european’, ‘state’, ‘directive’, ‘rights’, ‘national’, ‘right’, ‘states’, ‘united’, ‘kingdom’, ‘para’, ‘domestic’, ‘union’, ‘convention’, ‘regulation’, ‘decision’, ‘case’]	EU

Based on the 25-Topic model, the **UKSC workload** comprises 18 primary areas of law, namely, property, contract, intellectual property, insurance, immigration, international, customs, tax, family, human rights, devolution, criminal, tort, administrative, labor, EU, and company law. It also includes tangential areas of law such as damages and enforcement of arbitral awards.

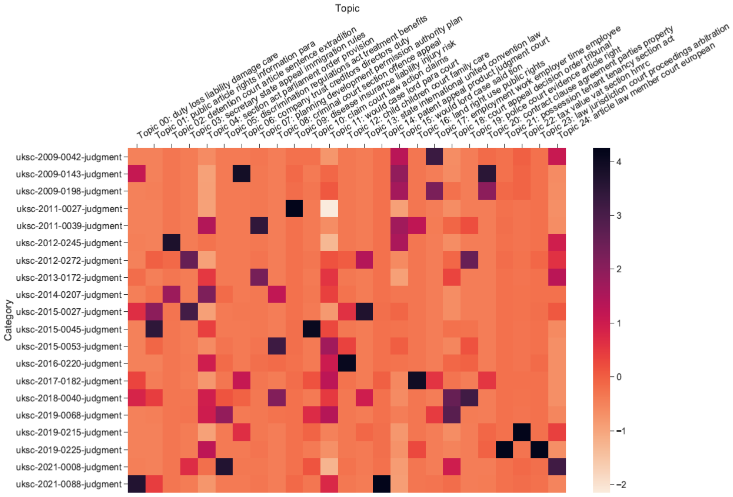

Figure 14. *Heatmap of UKSC dataset: 20 random judgments and probabilities of topics*

MALLET also provides the possibility of examining which topics a specific judgment would fall into as seen in the heatmap in Fig.14. Here, the model predicts that uksc-2021-0008-judgment ([2021] UKSC 53) has a higher probability of belonging to Topic 24 while uksc-2009-0143-judgment ([2009] UKSC 2) has the highest probability of belonging to Topics 6 and 22.

For greater insight into the judicial workload in a specific area, topic modelling can be used to identify which judgments are most significant in a selected topic. For example, uksc-2018-0080-judgment ([2018] UKSC 64)[15], uksc-2021-0079-judgment ([2021] UKSC 42)[16], and uksc-2009-0127-judgment ([2010] UKSC 10)[17] have more than 70% probability (amongst others) of belonging to Topic 16 (devolution).[18] Indeed, each judgment relates to the devolution of power to Scotland. A qualitative analysis of these judgments can provide insight into the nuanced dimensions arising in matters of devolution, from the legislative competence of the Scottish Parliament and the UK Parliament, the Scottish Parliament's ability to legislate for continuity of laws which are subject to EU law after the UK withdraws from the EU, to the legality of holding a referendum on Scottish independence.

The **LKSC workload** primarily relates to 10 areas of law: land, constitutional, labor, family, company, criminal, intellectual property, environmental, contract, and administrative law. Administrative law featured in three primary forms: inquiries (including disciplinary matters), state school admissions, and the establishment of a private medical university. This division of administrative law is quite insightful as they are not strict areas of law, but are domains that seemingly recur in the LKSC judicial workload. For example, the admission of children to state schools appears also as a distinct topic in a 10-Topic model indicating how pervasive these applications are in the LKSC's workload.

In the heatmap in Fig.15, the model predicts that 2014_sc_appeal_143_2013 is most probable to belong to Topics 17 and 23 while 2017_sc_appeal_162_2012 as belonging to Topic 14.

[15] Relates to the "UK Withdrawal from the European Union (Legal Continuity) (Scotland) Bill."

[16] Relates to two bills: the United Nations Convention on the Rights of the Child (Incorporation) (Scotland) Bill and the European Charter of Local Self-Government (Incorporation) (Scotland) Bill; considers whether the Scottish Parliament has the legislative competence to make specific laws.

[17] Relates to the distinction of powers between the Scottish and UK Parliaments.

[18] Note. this should not be understood as topics being mutually exclusive. Rather, topics are often overlapping labels. It is more than likely that a judgment will belong to more than one topic given the nature of legal and judicial reasoning.

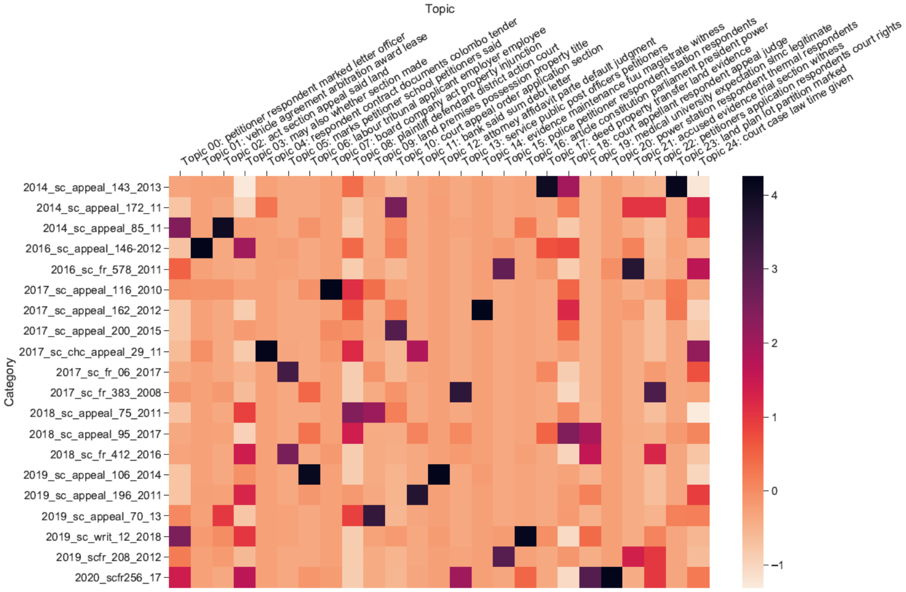

Figure 15. *Heatmap of UKSC dataset: 20 random judgments and probabilities of topics*

Evolution of judicial workloads over time

A time series analysis provides insight into how specific topics have been decided by the courts over time. In the LKSC, for example, Topic 16 which includes constitutional and electoral matters, demonstrate a general decline from 2014 onwards (see Fig. 16), while Topic 19, which involves judgments relating to university education, have increased (see Fig. 17). This reflects how the South Asian Institute of Technology and Management (SAITM) private university was the subject of increasing review by various stakeholders from around 2017. From the question of whether SAITM was a "Degree Awarding Institute" recognised under the Universities Act No. 16 of 1978, to the provisional registration of MBBS graduates of SAITM as medical practitioners under the Medical Ordinance, various legal and administrative questions were raised before the Supreme Court until 2019 when the LKSC directed that the petitioners be provisionally registered as medical practitioners (see S.C. F.R. Application No. 54/2019). This was followed by several applications from foreign MBBS graduates on related questions of eligibility to sit for the Examination for Registration to Practice Medicine (see S.C.F.R. Applications No. 399/2019, 145/2019, 149/2019). Further, an examination of the judgments with the highest probability of containing Topic 19 revealed that 80% of the top 10 judgments relating to university medical education were fundamental rights applications, indicating that the issue demonstrates a strong human rights dimension.

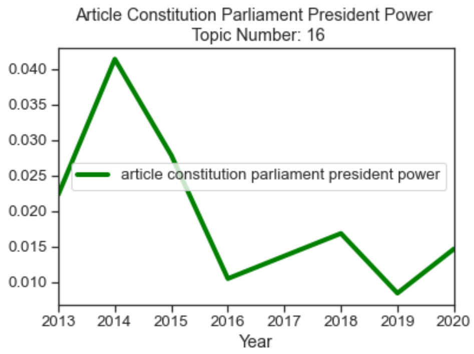

Figure 16.

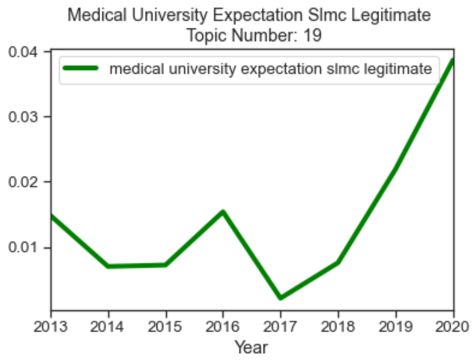

Figure 17.

Patterns of decision-making in the UKSC show that judgments relating to devolution (Topic 4) as well as judgments relating to the European Union (Topic 24) have been increasing over the last 2 years. A closer examination shows that several recent judgments (delivered in 2021-2023) relate to questions of devolution, and the withdrawal of the UK from the European Union ("Brexit"). While the UK formally exited the European Union in January 2020, questions regarding the impact of the formal Withdrawal Agreement on devolved powers were referred to the UKSC (see [2023] UKSC 5). Other questions regarding devolution include whether specific Bills would be within the legislative competence of the Scottish Parliament or the Northern Ireland Assembly (see [2021] UKSC 42), [2022] UKSC 32), and whether the Scottish Parliament can legislate to hold a referendum on Scottish independence (see [2022] UKSC 31). Whether a second referendum on Scottish independence[19] could or should be held has been a question that has been gaining traction over the last few years (Torrance, 2023). This not only demonstrates the shifting patterns of the judicial workload, but also reflects the social and political priorities of the jurisdiction (and its impact on the judicial workload) at a given point in time.

[19] For a detailed overview of the question of Scottish independence, including referendum proposals and negotiations from 1999 – 2014, new referendum developments, and the political steps taken towards a second referendum, see Torrance, D. (2023). *Scottish independence referendum: legal issues*. House of Commons Library. https://researchbriefings.files.parliament.uk/documents/CBP-9104/CBP-9104.pdf.

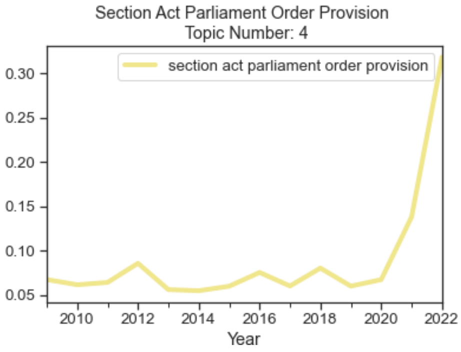

Figure 18.

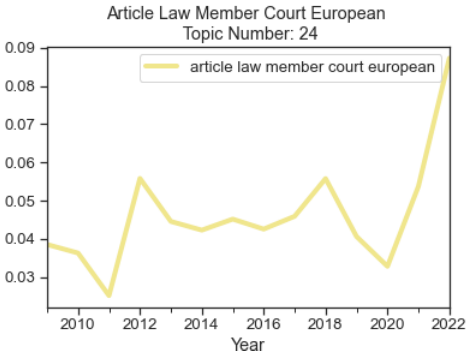

Figure 19.

Conclusion

Topic modelling is a means of “distant” reading hundreds of judgements at once: it provides a bird’s eye view of judicial writing and thinking by aggregating and analysing large amounts of legal data. Topic modelling is a useful tool of text analysis that can provide deeper insight into both the judicial workload and its practice through identifying nuanced dimensions of a specific area of law, to common legal questions that arise in specific domains. While this article engaged in a comparative, exploratory analysis of judicial discourse through text analytics, its findings have ramifications for understanding judicial workloads, how judicial workloads evolve over time, and traditional legal taxonomies, and demonstrates the potential for computational analysis of law.

Topic modelling sheds light on the composition of judicial workloads. Through topic modelling, 18 primary areas of law that the courts predominantly deals with were identified in the UKSC, and 10 in the LKSC. While the workloads of the jurisdictions have similarities (e.g. rights, company, criminal, and constitutional law feature prominently in both jurisdictions), the UKSC’s focus on immigration, international law, and the EU is unique. Similarly, the LKSC workload included environmental law, and specific domains within administrative law and fundamental rights, such as state school admissions. This makes clear that while the Supreme Court is the highest court in each jurisdiction, and though the Courts have differing jurisdictions, the nature of the questions of law and its volume may also differ due to the socio-political context of each country.

Patterns in the judicial workload were also identifiable over time. Increasing focus on legislative competence (relating to devolution) and EU-related matters in the UKSC are observable patterns that indicate the shifting nature of the UKSC workload. It also reflects the shifting social and political priorities of that jurisdiction. Therefore, topic modelling can be useful to identify and analyze legal trends, and possibly shifts in questions that arise in a specific area of law, for both lawyers as well as judges.

These findings have several implications that can aid organizational and training capacities within the administration of justice. First, it has the potential to assist in organizing and categorizing judgments based on underlying themes present in each case. Given that neither UK nor Sri Lankan apex court judgments are labelled with specific areas of law, topic modelling combined with other computational tools can aid in automatically classifying or categorizing judgments efficiently (see, for example, the work of Howe et al. (2019) and (Li et al., 2015)), allowing both judges and lawyers to access recently delivered judgments belonging to a specific area of law (e.g., land law) quickly. Second, topic modelling can be used to identify key judgments under each theme, which can make teaching and learning more efficient and effective. Third, the findings of this article demonstrate that not all areas of law come before the apex courts in equal measure. While further research is required, it is possible that the data would support a similar conclusion if this exercise is carried out with judgments of the lower courts. It may also be useful in instances where there are limited resources (including time), for example, in judicial or legal training programs. While it is important for both judges and lawyers to be skilled in tackling *any* question of law, limited resources can result in a need to prioritize the nature of training topics. Topic modelling can be used to identify key concepts that emerge frequently or are increasingly becoming prevalent, and target training towards such areas, and could be particularly useful to identify priority areas when designing continuing judicial education curricula and/or programs.

This article further sought to view the process of topic modelling beyond being a mere tool of NLP as suggested by Meeks and Weingart (2012). In addition to asking the questions of how these topics models provide insight into the judicial workload and how it may aid judicial processes, the article also examined what the subjective decisions taken to deploy the models reveal about the findings themselves. For instance, the models also contained generic topics which could not be meaningfully labelled as an area of law, presumably due to leaving in technical stop-words specific to legal language. At first blush, the generic topics seemingly do not aid analysis of the judicial workload. However, by leaving in such stop-words, an important aspect of judicial decision-making is highlighted: these are frequently used words in legal reasoning (e.g., “would” – when discussing outcomes of a case) and is a core aspect of judicial language. These are also words used to refer to specific evidence (e.g. “fact”, “question”) and its evaluation. Similarly, words like “may,” “must,” “could” are used to discuss both hypothetical circumstances (a staple in judicial judgments) and discretion (of the court and other actors). Therefore, while it is not thematic in a strict sense, it is reflective of an importance aspect of judicial power, and by extension, its workload, by highlighting the nature of judicial reasoning and how it is represented through language. It is noted, however, that subjective decision-making in choosing to leave in technical stop-words, selecting the optimum no. of topics, and manually assigning a label to topics based on the author's subjective evaluation, results in the author "impact(ing) the results of the study" (Yau et al., 2014, p. 775), which in turn, poses challenges to reproducibility.

Topic modelling also raises questions about how we perceive and categorize laws. For example, the heatmaps in Fig.14 and Fig.15 indicate how some judgments have varying degrees of probability of containing more than one topic. Similarly, in both jurisdictions, only one topic related strongly to criminal law, thus challenging the broader, more general division of judicial practice into "criminal" and "civil" law (which has influenced the design of the judicial hierarchy).[20] Therefore, topic modelling allows us to reconceptualize how legal discourse is perceived in both legal education and judicial training, as the existing curricula envisions rigid demarcations of law (for example, "family law", or "international law"). But in practice, laws often overlap when considering a single legal dispute. This finding lends credence to "a taxonomy of [legal] practice" (Carter et al., 2016, p. 1338) that could exist in reconceptualizing how judgments are categorized, and topic modelling could be an insightful tool in aiding this understanding.

The use of digital methods also raises questions about legal and data ethics as well. In the United Kingdom, where the judgment relates to a child, the name of the child is anonymized in the judgment. For example, in UKSC-2009-0075, the Judgment was titled "I (A Child)". There is also a Practice Guidance issued by the President of the Family Division in 2018 which provides guidelines on removing personal and geographical indicators in judgments (as well as explicit descriptions of sexual abuse of children) (McFarlane, 2018). This practice is not followed consistently in Sri Lanka. While there have been a few instances of anonymization (for example, see S.C. Appeal 32/2020 delivered in October 2020), some judgments involving children still contain the name of the child (for example, see S.C. Appeal No. 89A/2009, S.C. Appeal No. 17/2013, and more recently S.C. Appeal No. 239/2017 (decided in 2022) where the name of the victim is not redacted or anonymized though they were under 16 years of age). Once such judgments are publicly uploaded to the LKSC website, this data gains an additional degree of visibility as it can be accessed, viewed, and analyzed by any person with digital access.

Finally, it is worth noting that deploying topic modelling in this manner demonstrates some of the difficulties a humanist would encounter when engaging with digital methods. To a degree, engaging in the digital humanities requires a certain awareness and familiarity with not only digital methodologies, but fundamentals of programming in order to navigate some of the technical challenges that might

[20] For example, in Sri Lanka, the Courts of First Instance are divided into "criminal" and "civil" courts: the District Court handles the civil matters, and the High Courts are divided into "criminal" and "civil" as well.

arise when doing so. This makes a compelling case for introducing humanities students to computing for greater collaboration that provides new avenues to revisit traditional humanities questions, and to discuss the ethical dimensions of digital methodologies.

Conflict of interest

The author has no conflict of interest to declare.

Acknowledgements

The author acknowledges Chevening Scholarships, the UK government's global scholarship program, funded by the Foreign, Commonwealth and Development Office (FCDO) and partner organizations which funded the author's studies at Durham University, during which this article was produced. The author is deeply grateful to Prof. Peter Heslin, Durham University, for his guidance in navigating the Digital Humanities and combining DH approaches with legal data, and is further thankful to the two reviewers whose comments made this article more robust.

References

Alali, M., Syed, S., Alsayed, M., Patel, S., & Bodala, H. (2021). JUSTICE: A benchmark dataset for Supreme Court's judgment prediction. *arXiv*. https://doi.org/10.48550/arXiv.2112.03414

Antoniak, M. (2021a). *little-mallet-wrapper*. Github. Retrieved 25th March 2023 from https://github.com/maria-antoniak/little-mallet-wrapper

Antoniak, M. (2021b). *little_mallet_wrapper demo*. Github. Retrieved 25 March 2023 from https://github.com/maria-antoniak/little-mallet-wrapper/blob/master/demo.ipynb

Balogun, J. N., Ogor Folashade, Ayankoya, Ernest, O., Kasali, F., & Christopher, A. (2023). *Appeal Cases heard at the Supreme Court of Nigeria Dataset* (Version 1) [Dataset] Mendeley Data https://doi.org/10.17632/ky6zfyf669.1

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Brett, M. R. (2012). Topic Modeling: A basic introduction. *Journal of Digital Humanities*, *2*(1). https://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Carter, D. J., Brown, J., & Rahmani, A. (2016). Reading the High Court at a distance: Topic modelling the legal subject matter and judicial activity of the High

Court of Australia, 1903–2015. *University of New South Wales Law Journal*, *39*(4), 1300-1354.

Constitution of the Democratic Socialist Republic of Sri Lanka1978 (LK)

Cooray, L. (1976). The administration of justice in Sri Lanka. *Hong Kong Law Journal*, *6*(67), 67-88.

Cooray, L. J. M. (1975). Common law in England and Sri Lanka. *The International and Comparative Law Quarterly*, *24*(3), 553-564. http://www.jstor.org/stable/758782

Cowie, G. (2022). *The UK Supreme Court* (Research Briefing, Issue. https://commonslibrary.parliament.uk/research-briefings/cbp-9536/#:~:text=The%2012%2Dmember%20Court%20is,Wales%20and%20in%20Northern%20Ireland.

dannylesmy. (2022). quick_train_topic_model issues #7. In *Maria Antoniak: little-mallet-wrapper*. Github: https://github.com/maria-antoniak/little-mallet-wrapper/issues/7.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics*, *41*(6), 570-606.

Ebeid, I., Arango, J. S. M., & Xu, X. (2016). *Mallet vs GenSim: Topic modelling evaluation report*. University of Arkansas at Little Rock.

Falk, M. G. (2016). *Faraway, so close!: Reading Adeline Mowbray closely using topic modelling* Digital Humanities 2016: Conference Abstracts, Krakow. https://dh2016.adho.org/abstracts/337

Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, *34*(5), 1515-1540. https://doi.org/10.1007/s11196-021-09860-8

Graham, S., & Milligan, I. (2013). Review of MALLET, produced by Andrew Kachites McCallum. *Journal of Digital Humanities*, *2*(1). https://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham/

Graham, S., Weingart, S., & Milligan, I. (2012). *Getting started with topic modeling* *and MALLET*. Programming Historian. Retrieved 19 March 2023 from https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

hiQ Labs, Inc v. LinkedIn Corp (2022) N. D. Cal. 17-3301

Howe, J. S. T., Khang, L. H., & Chai, I. E. (2019). Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. *arXiv*. https://doi.org/arXiv:1904.06470

Izzidien, A., Sargeant, H., & Steffek, F. (2022). *What goes on in court? Identifying contract-related topics decided by United Kingdom courts from 1709 to 2021 using machine learning.* Language Sciences Annual Symposium 2022, Cambridge. https://www.cambridge.org/engage/coe/article-details/637c101621b45c8f0f245373

J.L v. Alphabet Inc (2023), N.D. Cal. 3:23-cv-03440.

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics*, *41*(6), 750-769.

Li, X., Ouyang, J., & Zhou, X. (2015). Supervised topic models for multi-label classification. *Neurocomputing*, *149*, 811-819. https://doi.org/https://doi.org/10.1016/j.neucom.2014.07.053

Luz De Araujo, P. H., & De Campos, T. (2020). Topic modelling Brazilian Supreme Court lawsuits. *Legal Knowledge and Information Systems 334,* 113-122.

McFarlane, A. (2018). *Practice Guidance: Anonymisation and Avoidance of the Identification of Children and the Treatment of Explicit Descriptions of the Sexual Abuse of Children in Judgments intended for the Public Arena*. www.judiciary.uk. Retrieved 5 May 2023 from https://www.judiciary.uk/wp-content/uploads/2018/12/anonymisation-guidance-1-1.pdf

Meeks, E. (2011). *Comprehending the Digital Humanities.* Stanford University Libraries. https://dhs.stanford.edu/comprehending-the-digital-humanities/

Meeks, E., & Weingart, S. (2012). The Digital Humanities contribution to topic modeling. *Journal of Digital Humanities*, *2*. https://journalofdigitalhumanities.org/2-1/dh-contribution-to-topic-modeling/

Mimno, D. (2022). *Why I don't recommend stochastic variational Bayes for topic models*. Retrieved 21 February 2023 from http://mimno.org/notebooks/Variational_Bayes_LDA.html

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*(54).

Moretti, F. (2013). *Distant reading*. Verso Books.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora*, *12*(2), 243-277. https://doi.org/10.3366/cor.2017.0118

Novotná, T., Harašta, J., & Kól, J. (2020). Topic modelling of the Czech Supreme Court decisions. *CEUR Workshop Proceedings,* Automated Semantic Analysis of Information in Legal Text (ASAIL) 2020, Online.

Perera, S., Perera, I., & Ahangama, S. (2022, 23-24 Feb. 2022). Exploring Twitter messages during the COVID-19 pandemic in Sri Lanka: Topic modelling and emotion analysis. 2022 2nd International Conference on Advanced Research in Computing (ICARC), Belihuloya, Sri Lanka.

PM v. OpenAI (2023), N.D. Cal. 3:23-cv-03199.

Public Records Act 1958 (UK)

Quemy, A. (2018). European Court of Human Rights open data project. *arXiv*. https://doi.org/arXiv:1810.03115

Sbalchiero, S., & Eder, M. (2020). Topic modeling, long texts and the best number of topics. Some problems and solutions. *Quality & Quantity*, *54*(4), 1095-1108. https://doi.org/10.1007/s11135-020-00976-w

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464-2476. https://doi.org/https://doi.org/10.1002/asi.23596

Tijare, P., & Rani, J. P. (2020). Exploring popular topic models. *Journal of Physics: Conference Series*, *1706*. https://doi.org/doi:10.1088/1742-6596/1706/1/012171

Torrance, D. (2023). *Scottish independence referendum: legal issues*. House of Commons Library. https://researchbriefings.files.parliament.uk/documents/CBP-9104/CBP-9104.pdf

Walsh, M. (2020a). *Introduction to cultural analytics & python*. Retrieved 19 March 2023 from https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/08-Topic-Modeling-Text-Files.html

Walsh, M. (2020b). *Introduction to cultural analytics & python*. Retrieved 1 April 2023 from https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/11-Topic-Modeling-Time-Series.html

Walsh, M. (2020c). *Introduction to cultural analytics & python*. Retrieved 18 March 2023 from https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/07-Topic-Modeling-Set-Up.html

Weston, S. J., Shryock, I., Light, R., & Fisher, P. A. (2023). Selecting the number and labels of topics in topic modeling: A tutorial. *Advances in Methods and Practices in Psychological Science*, *6*(2). https://doi.org/10.1177/25152459231160105

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767-786. https://doi.org/10.1007/s11192-014-1321-8