DATA40230 Digital Humanities: Practice and Theory

Final Summative Project

Contents

Reading Supreme Courts from a Distance:

Topic Modelling Judgements of the Supreme Courts of Sri Lanka and the United Kingdom

1. 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. This study utilises topic modelling of judgments delivered by the Supreme Courts of Sri Lanka and the United Kingdom to examine how the respective judicial workloads are constructed, and how it may shift over time. It also explores how topic modelling operates in relation to judicial discourse, specifically, how changing hyperparameters of the topic models (namely, number of topics) affects its output.

1.1. Topic Modelling

Topic modelling is an unsupervised machine learning method that aids in the discovery of topics contained in a selection of documents i.e. topic modelling can cluster phrases to best characterise the corpus. This has led to topic modelling being described as "an attempt to inject semantic meaning into vocabulary" (Graham et al., 2012) to uncover "evidence already in the text" (Brett, 2012). While there are several algorithms and tools, this study 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).

There are several tools that deploy LDA. However, some do not produce reasonably interpretable outputs. For instance, Tijare and Rani (2020) observe that models using Gensim LDA performs worse. Mimno (2022) posits that Gensim removes possibilities rapidly which may lead to a less effective model being selected before a more optimal model is found. Therefore, this exercise utilises **Ma**chine **L**earning for **L**anguag**e** **T**oolkit (MALLET). To develop its topic models, MALLET applies Gibbs sampling, a statistical method that efficiently generates a sample distribution (Graham et al., 2012).

1.2. Topic Modelling and Judicial Discourse

While topic modelling has been deployed in numerous domains, this study examines its utility in the legal domain. There are a few examples of topic modelling in similar settings in other jurisdictions. Carter et al. (2016) uses topic modelling to analyse 7476 judgments of the High Court of Australia (1903-2015), while more recently, topic modelling has been deployed to analyse 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 either the UK Supreme Court[1] or the Sri Lankan Supreme Court.

[1] Note that there is a 2022 Conference Paper titled "What goes on in court? Identifying contract-related topics decided by United Kingdom courts from 1709 to 2021 using machine learning" by Ahmed Izzidien, Holli Sargeant, and Felix Steffek available at https://www.cambridge.org/engage/coe/article-details/637c101621b45c8f0f245373. This is a general paper that does not focus solely on the Supreme Court, and is a pre-print and has not been peer-reviewed at the time of posting.

1.3. Selection of Jurisdictions

Topic modelling has the potential to provide insight into judicial decision-making by uncovering broad thematic outlines of the contents of judgments. While the respective Supreme Courts are selected (as opposed to lower court judgments) as they are the highest courts in the judicial hierarchies in both jurisdictions, the UK legal system, through colonisation, has also influenced the content and structure of the Sri Lankan judicial system (Cooray, 1975). Its influence continues through the *Civil Law Ordinance* which provides that, in the absence of Sri Lankan legislation, English law applies in maritime and commercial matters in Sri Lanka.[2] Therefore, a comparative examination of topic modelling of judgments from the two jurisdictions has potential to provide insight into its respective judicial workloads.

2. Datasets

2.1. Dataset Creation and Pre-Processing

The Supreme Court of Sri Lanka (LKSC) Dataset is an existing dataset personally collected by the author. It contains all 933 judgments delivered by the LKSC and uploaded to the LKSC website from 2013-2020. The United Kingdom Supreme Court (UKSC) Dataset was obtained by scraping the UKSC website for all judgments delivered from 2009-2022. In total, 892 judgments were downloaded.[3] 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. Thereafter, numbers, stop-words, and punctuation were removed, and the text was converted to lowercase using the little_mallet_wrapper.process_string function. For more information on the design of the scraper, and the code to convert text, see the accompanying notebook.

2.2. Limitations

The respective corpora represent only a component of the judicial workload. For example, judges 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.[4] However, this is attributable to the nature of the data because it is limited to the digitized judgments presently available.

2.3. Ethical Considerations

Court judgments are considered public records, 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.[5] While laws on the legality of web scraping is sparse, the decision of the US Court of Appeals protecting the scraping of public data (see hiQ Labs, Inc v. LinkedIn Corp)[6], though not applicable in the UK or Sri Lanka, provides direction on how legal systems may approach web scraping in the future. Therefore, there appears to be no ethical or legal challenges to the proposed study.

[2] 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/.
[3] Important: though 1031 judgments were delivered by the UKSC, only 892 judgments were available to be downloaded.
[4] For example, in *Mining the Dispatch*, Nelson utilised a corpus of 112,000 documents.
[5] 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 necessarily followed in Sri Lanka.
[6] No. 17-3301 (N. D. Cal. Nov. 4, 2022).
Judgment available at https://storage.courtlistener.com/recap/gov.uscourts.cand.312704/gov.uscourts.cand.312704.404.0.pdf.

3. The Topic Models

3.1. Initial Challenges

Graham et al. (2012) observes that deploying MALLET requires "tacit knowledge for computer scientists" which can be "completely opaque for humanists." Indeed, MALLET requires installation of the Java Developer's Kit,[7] 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.[8] To run MALLET on Jupyter Notebooks, 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. Debugging this proved to be time consuming, and was finally resolved by defining the source folder, in addition to the (bin) path.[9]

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 dataframe, alien characters were present. These could not be processed by MALLET. Therefore, the texts were further processed by converting to ASCII encoding, and discarding the incompatible characters.

3.2. No. of Topics

A hyperparameter in LDA is the number of topics the model should identify in the corpus. However, the suitable number of topics is generally unknown (Yau et al., 2014) and the choice 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 & Toivanen, 2016) selected 60 topics after trial-and-error. Following this practice, the models were trained using the little-mallet-wrapper "quick_train_topic_model" function 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. It was observed that both 25 and 50 provided greater granularity in the topics identified, and 25 topics were selected for both the datasets prioritising interpretability and ease, and retrained using the little-mallet-wrapper "train_topic_model" function.

3.3. Manually Labelling Topics

While Mallet produces a given number of topics and the top words, it does not label the topic. Therefore, a label was manually assigned to each topic. Tables with these manual labels are included in the Appendix.

3.4. Code

The code to deploy the topic models and the time series analysis is adapted from Melanie Walsh's "Topic Modelling - Text Files" and "Topic Modeling – Time Series" in Introduction to Cultural Analytics & Python.[10] The code also relies on the little_mallet_wrapper developed by Maria Antoniak and her demo.

[7] Java can be downloaded at: https://www.oracle.com/java/technologies/downloads/#jdk20-windows
[8] Detailed guidance on how to do this is available at: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet
[9] "Quick_train_topic_model issues #7" available at https://github.com/maria-antoniak/little-mallet-wrapper/issues/7
[10] 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.

4. Analysis

4.1. Workload Overview

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 1031 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.

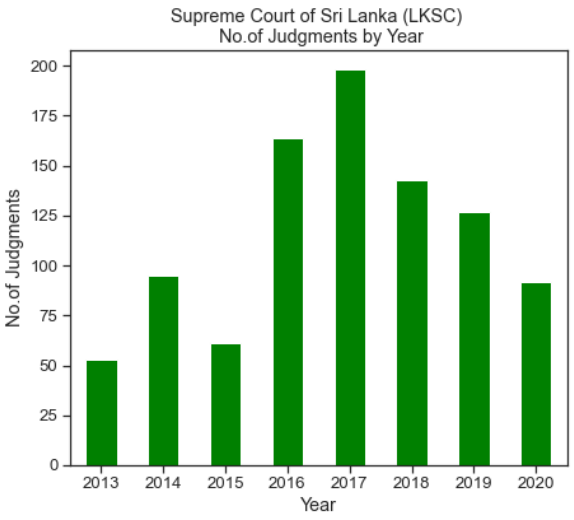

Figure 1. No. of Judgments by Year: LKSC

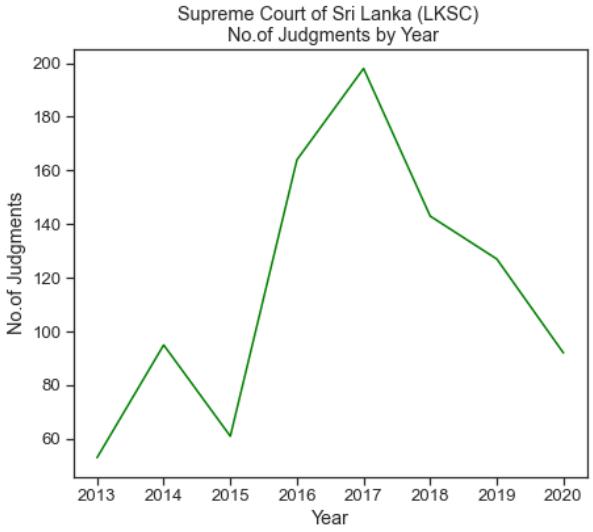

Figure 2. No. of Judgments by Year: LKSC

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.

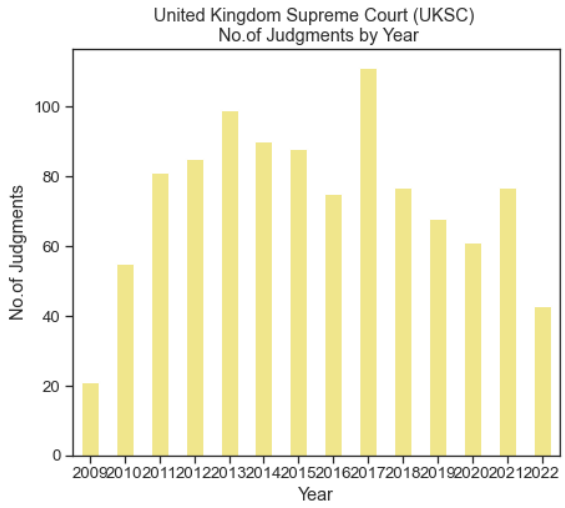

Figure 3. No. of Judgments by Year: UKSC

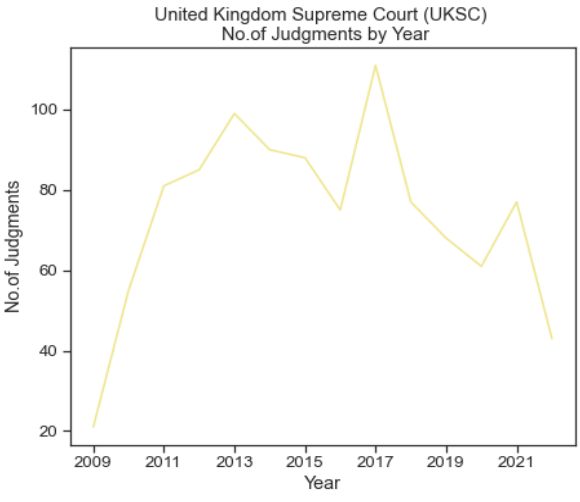

Figure 4. No. of Judgments by Year: UKSC

Even though the no. of judgments are 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.

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

Table 1. Summary Statistics of the Datasets

4.2. Impact of No. of Topics

Asthe no. 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', 'l
ord', 'whether', 'may', 'children', 'rights', 'also', 'person']

Topic 3

['law', 'court', 'article', 'state', 'act', 'section', 'jurisdiction', 'would', 'para', 'united', 'case', 'proceedings', 'pag
e', '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', 'judg
e', 'learned', 'company', 'case', 'agreement', 'applicant', 'action', 'appeal', 'judgment']

Topic 1

['plaintiff', 'defendant', 'court', 'land', 'respondent', 'district', 'appellant', 'case', 'evidence', 'judge', 'said', 'proper
ty', 'deed', 'high', 'action', 'appeal', 'title', 'learned', 'law', 'judgment']

Topic 2

['petitioner', 'petitioners', 'respondent', 'respondents', 'service', 'marked', 'public', 'said', 'application', 'marks', 'comm
ission', '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', 'le
arned', 'said', 'high', 'filed', 'judgment', 'judge', 'provisions', 'counsel']
```

Figure 6. LKSC 5-topic model: Topics

However, as the topics increased, majority developed greater specificity. 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', 'ac
tion', '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', 'offenc
e', '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', 'foreig
n', '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', 'dec
eased', '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', 'rul
e', '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', 'arbitrator
s', '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.

```
Topic 0

['plaintiff', 'defendant', 'judge', 'plaint', 'appeal', 'court', 'trial', 'action', 'district', 'high', 'learned', 'held', 'c
ase', 'issue', 'answer', 'evidence', 'stated', 'issues', 'also', 'dismissed']

Topic 1

['commission', 'member', 'public', 'secretary', 'colombo', 'ministry', 'department', 'service', 'road', 'general', 'nationa
l', 'director', 'chairman', 'excise', 'former', 'development', 'nawala', 'letter', 'mawatha', 'narahenpita']

Topic 2

['appellants', 'appeal', 'feet', 'stated', 'case', 'road', 'hewage', 'shop', 'room', 'matugama', 'wall', 'upon', 'substitute
d', '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', 'niya
gama', '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"]

4.3. Examination of Judicial Workloads: 25-Topic Models

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,

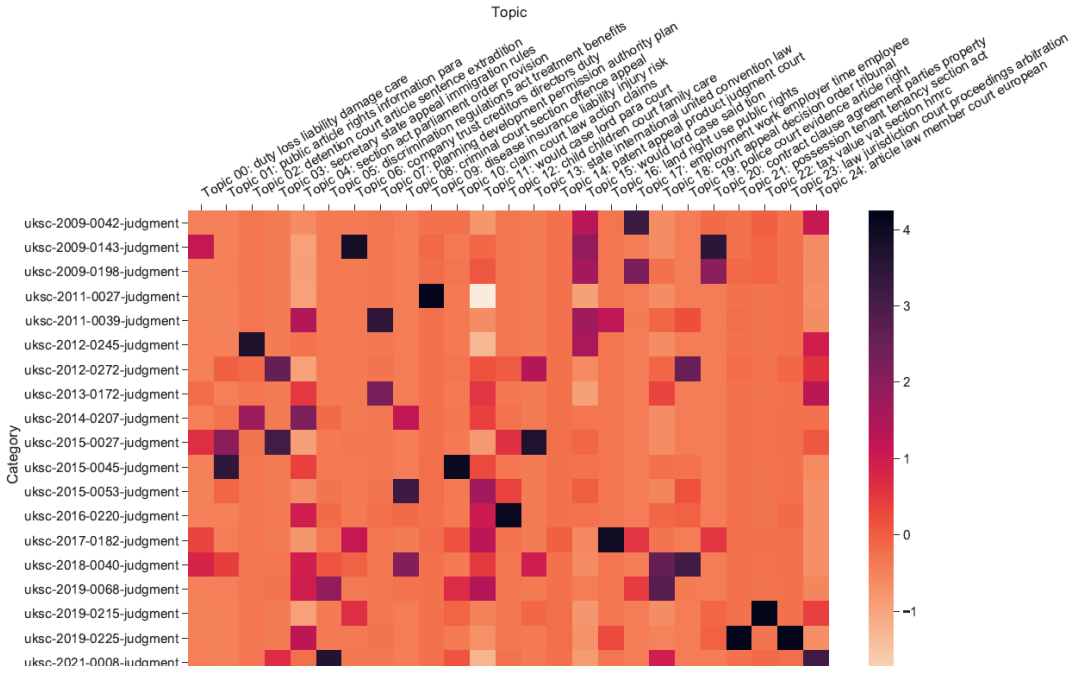

Figure 14. Heatmap of UKSC Dataset: 20 Random Judgments and Probabilities of Topics

human rights, devolution, criminal, tort, administrative, labour, EU, and company law. It also includes tangential areas of law such as damages and enforcement of arbitral awards. In the heatmap in Fig.14, the model predicts that uksc-2021-008-judgment has a higher probability of belonging to Topic 24 while uksc-2009-0143-judgment 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 given topic. For example, uksc-2018-0080-judgment,[11] uksc-2021-0079-judgment,[12] and uksc-2009-0127-judgment[13] have more than 70% probability of belonging to Topic 16 (devolution). Indeed, each judgment relates to the Scottish devolution. 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, industrial, family, company, criminal, intellectual, 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. However, the LKSC model produced, on average, more generic topics with words like "may", "whether", "would", "however" etc.

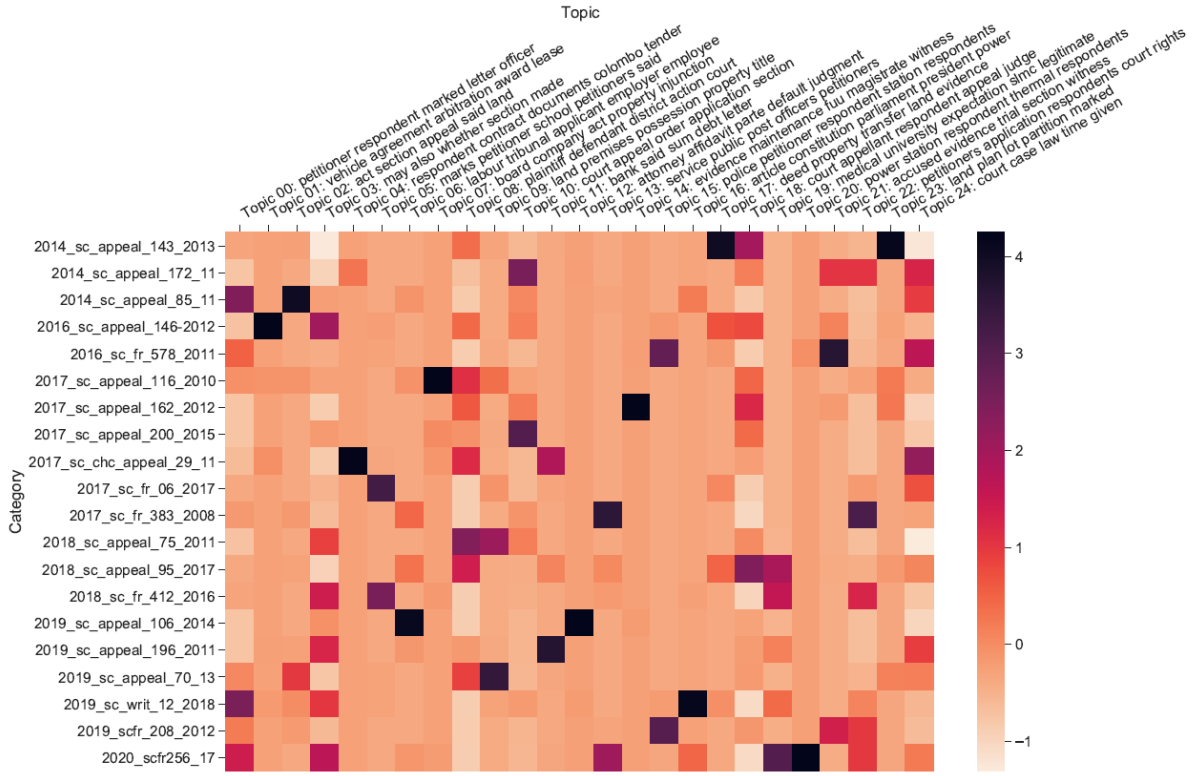

Figure 15. Heatmap of UKSC Dataset: 20 Random Judgments and Probabilities of Topics

[11] Relates to the "UK Withdrawal from the European Union (Legal Continuity) (Scotland) Bill."
[12] 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.
[13] Relates to the distinction of powers between the Scottish and UK Parliaments.

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.

4.4. 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. Topic 19, which involves judgments relating to the private medical education, have increased. The establishment (and subsequent abolishment) of the SAITM private medical university was the subject of constant review by various stakeholders due to various legal and regulatory questions, and was finally resolved only recently. An examination of the judgments with the highest probability of containing Topic 14, 80% of the top 10 judgments were fundamental rights applications, indicating that the issue demonstrates a strong human rights dimension.

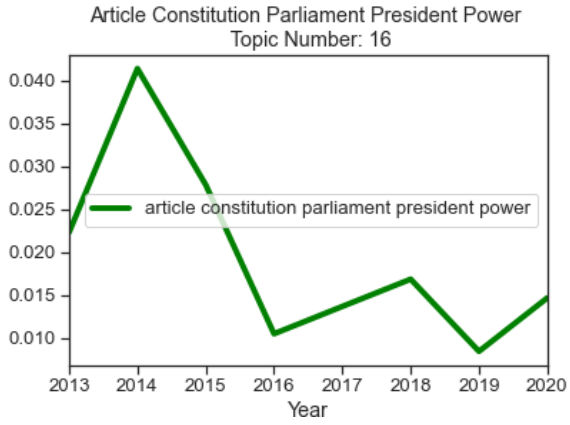

Figure 16.

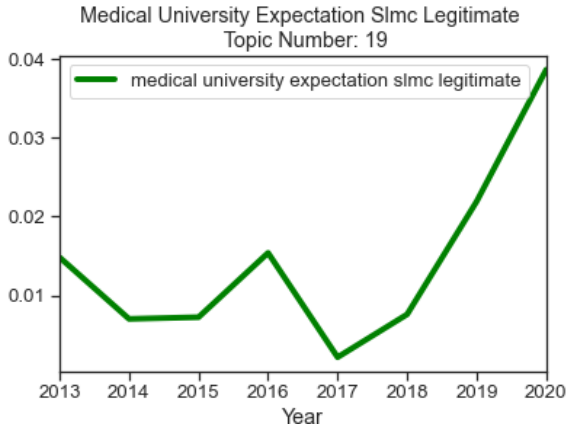

Figure 17.

Patterns of decision-making in the UKSC show that judgments relating to devolution (Topic 4) have been increasing over the last 2 years, while judgments relating to the European Union (Topic 24) have been declining (presumably with "Brexit"). 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.

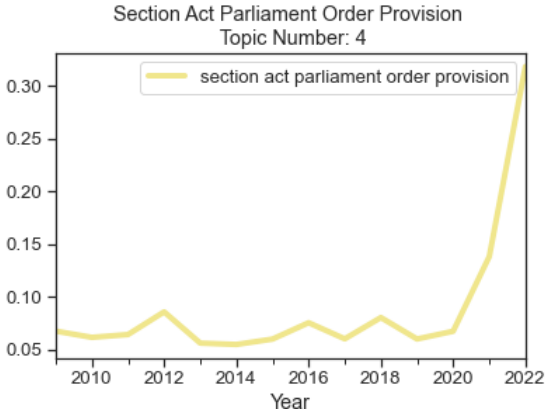

Figure 18.

Figure 19.

5. Conclusion

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. However, topic modelling, and in particular MALLET (and LDA), requires subjective decision-making to select the number of topics, and label them.

Through topic modelling, 18 primary areas of law were identified in the UKSC, and 10 in the LKSC. While the workload of the jurisdictions has similarities (e.g. rights, company, criminal, and constitutional law feature 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, such as state school admission. Patterns in the judicial workload were also identifiable over time. Increasing focus on legislative power and declining focus on EU matters in the UKSC are observable patterns that indicate the shifting nature of the workload as well as how the cases before the Court may reflect social and political priorities of that jurisdiction.

The models also contained generic topics which could not be meaningfully labelled as an area of law. Though the generic topics seemingly do not aid analysis of the judicial workload, it indicates an aspect of judicial decision-making: 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.

Overall, the UKSC topics were easier to label in comparison to the LKSC topics. One of the reasons for this is perhaps the abundance of use of terminology referring to actors ("plaintiff", "defendant", "petitioner" etc) and processes that featured dominantly in the top words in the LKSC judgments. Another reason could be that this is evidence for topic modelling producing better results with large datasets, for in this case, even though there are more (933) texts in the LKSC dataset, the texts are considerably shorter (average 1604 words, in contrast to the UKSC dataset's 7003 words). In any event, such subjective decision-making in choosing the no. of topics, and manually assigning a label to topics results in the researcher "impact(ing) the results of the study" (Yau et al., 2014, p. 775), which in turn, poses challenges to reproducibility.

Topic modelling raises questions about how we perceive and categorise 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).[14] Therefore, topic modelling allows us to reconceptualise how we perceive legal discourse in both legal education and judicial training. The existing curricula is designed in more regimented demarcations of law as, for example, "family law", or "international law", but in practice, laws often overlap. This lends credence to "a taxonomy of [legal] practice" (Carter et al., 2016, p. 1338) that could exist in reconceptualising how judgments are categorised, and topic modelling could be an insightful tool in aiding this understanding.

Word count: 2967 [excluding references, annexures, and footnotes]

[14] 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.

Bibliography

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.

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

dannylesmy. (2022). quick_train_topic_model issues #7. In *little-mallet-wrapper*.

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.

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., 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

Luz De Araujo, P. H., & De Campos, T. (2020). Topic Modelling Brazilian Supreme Court Lawsuits. In *Legal Knowledge and Information Systems* (pp. 113-122). IOS Press.

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

Novotná, T., Harašta, J., & Kól, J. (2020). Topic Modelling of the Czech Supreme Court Decisions.*Proceedings of ASAIL 2020* ASALL,

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

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

References for Setting Up Mallet

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

dannylesmy. (2022). quick_train_topic_model issues #7. In *little-mallet-wrapper*.

References for Topic Modelling Code

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

Walsh, M. (2020a). *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. (2020b). *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. (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

Appendices

Appendix 1. LKSC 25-Topic Model with Top Words and 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; Taxation
Topic 5	['marks', 'petitioner', 'school', 'petitioners', 'said', 'education', 'circular', 'admission', 'application', 'children', 'college', 'child', 'board', 'vidyalaya', 'interview', 'clause', 'residence', 'respondent', 'grade', 'schools']	[State school admissions]
Topic 6	['labour', 'tribunal', 'applicant', 'employer', 'employee', 'order', 'workman', 'employment', 'termination', 'industrial', 'employees', 'compensation', 'services', 'president', 'respondent', 'evidence', 'act', 'application', 'company', 'disputes']	Labour
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']	Procedural

Topic 13	['service', 'public', 'post', 'officers', 'petitioners', 'member', 'commission', 'grade', 'class', 'department', 'scheme', 'interview', 'said', 'appointment', 'ministry', 'secretary', 'circular', 'sri', 'iii', 'examination']	Administrative; Labour
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']	[Private Medical University Establishment]
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

Appendix 1. UKSC 25-Topic Model with Top Words and 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; Labour; 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']	Labour
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