

READING THE HIGH COURT AT A DISTANCE: TOPIC MODELLING THE LEGAL SUBJECT MATTER AND JUDICIAL ACTIVITY OF THE HIGH COURT OF AUSTRALIA, 1903–2015

DAVID J CARTER*, JAMES BROWN** AND ADEL RAHMANI***

I INTRODUCTION

In this article we apply the method of quantitative textual analysis known as ‘topic modelling’ to a significant Australian legal text corpus: that of judgments of the High Court of Australia from 1903 to 2015.¹ The High Court of Australia has been a perennial topic for study and analysis. It is the highest court in the Australian judicial hierarchy and the site of many of the most significant contests of legal doctrine and practice in Australian history. We find that the topic models generated by this research enable the development of a range of unique, novel and robust observations of the High Court’s judicial workload and the shifting make-up of its legal subject matter over time. Moreover, this article reveals the feasibility and value of topic modelling as a method for the study of legal texts and practices that might fruitfully complement other methods of legal scholarship.

To our knowledge this is the first time topic modelling has been applied in this way to the entirety of this legal textual corpus;² consequently, in this article we have focused more on presenting the method, deferring for the present

* Lecturer in Law, Faculty of Law, University of Technology Sydney.

** Professor of Official Statistics, Associate Head of School (Research), School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney; Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers.

*** Senior Lecturer, School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney.

The authors would like to thank Dr Anthea Vogl, Rachel Young and Matthew Sidebotham for their comments on drafts, as well as Dr Philip Chung (Austlii) and Michael Green SC (BarNet, Jade) for earlier discussion of the field more broadly. The usual caveat applies, views expressed herein are those of the authors.

1 This excludes special leave dispositions, transcripts and bulletins.

2 Seroussi and colleagues have applied topic modelling to a subset of this corpus, namely the judgments of Dixon CJ, McTiernan and Rich JJ in order to pursue author attribution studies of text attributed to those authors: see Yanir Seroussi, Ingrid Zukerman and Fabian Bohnert, ‘Authorship Attribution with Topic Models’ (2014) 40 *Computational Linguistics* 269. This work is a development of previous work on the question, undertaken using different approaches: see Yanir Seroussi, Russell Smyth and Ingrid Zukerman, ‘Ghosts from the High Court’s Past: Evidence from Computational Linguistics for Dixon Ghosting for McTiernan and Rich’ (2011) 34 *University of New South Wales Law Journal* 984.

extended analysis using its output.³ We limit the focus of this article to development of the topic model itself, and to providing some insight into the nature and workload of the Court. We understand this work to complement other forms of scholarship on the Court as an institution and on its workload. This includes the work of Groves and Smyth who chart the patterns of judicial writing on the High Court during the 20th century⁴ and, particularly, the ongoing work of Lynch and Williams in their annual statistical review series.⁵ So too does it complement the work of Smyth, who examines the changing patterns of the case load of the Australian state supreme courts during the 20th century, using a sampling and manual coding approach.⁶

We adopt a methodological position deliberately in alignment with the existing literature by Lynch and Williams, who position their statistical survey data of the Court's activity as 'intended to complement substantive analyses of the Court's work'.⁷ As such, we use the model to construct a perspective on the Court's judicial workload and the legal subject matter of that workload, venturing into more detailed analysis of specific periods, cases and the relationships to illustrate the results of our topic modelling. Specifically, we trace the changing length and number of cases published by the Court; we show how the Court's focus upon particular topics has changed over time; and we examine how this has occurred during particular periods of time during the 20th century. These tasks are undertaken both to show the validity of the models which we produce and to illustrate these important questions about the Court's activity. Finally, we briefly demonstrate the potential for this model to reframe the Court's own activity and that of the work of legal taxonomies through a description of how it is these key texts construct two visions of 'land' and 'territory'.

To begin, we situate the method within the frame of 'distant reading' and the digital humanities more broadly. Distant reading alters the balance between inclusion and exclusion of texts alongside the scale at which to read a collection. A distant reading paradigm argues for very broad inclusivity, reading thousands of texts at a time, and thus requires a practice of 'reading' which is other than the 'close' reading scale of attention to sentence, vocabulary, word order, narrative and style. This 'distant reading' is marked instead by techniques of reconfiguring large collections of textual material by computational processes. We argue that for legal scholarship, the practice of reading at a distant scale is more commonplace than it may seem at first glance, even though the digital humanities techniques we utilise are novel. With a shared bias towards text, the techniques

3 In any event, useful analysis of this data for the purposes of legal scholarship would require the use of a variety of other forms of legal method, a factor we discuss below. In this, we have modelled this approach on that used successfully by others in this field, notably Lynch and Williams.

4 Matthew Groves and Russell Smyth, 'A Century of Judicial Style: Changing Patterns in Judgment Writing on the High Court 1903–2001' (2004) 32 *Federal Law Review* 255.

5 See, eg. Andrew Lynch and George Williams, 'The High Court on Constitutional Law: The 2014 Statistics' (2015) 38 *University of New South Wales Law Journal* 1078, 1078.

6 Russell Smyth, 'The Business of the Australian State Supreme Courts over the Course of the 20th Century' (2010) 7 *Journal of Empirical Legal Studies* 141. See especially the discussion in relation to method: at 145.

7 Lynch and Williams, above n 5, 1078.

and methods of the digital humanities represent a natural source for legal textual analysis and scholarship. Over the remainder of the article, we explore this potentially fruitful method, applying it directly to legal text. Thus, in Part III of this article, we describe our research design for topic modelling the judgments of the High Court of Australia before presenting our results in Part IV and Part V. Throughout, we present insights, via a series of test cases, into the topic model and its usefulness for more developed legal analysis.

The results represent a unique, novel and robust contribution to the study of the High Court's judicial workload throughout its history. The method reveals new perspectives on the judicial workload, and its legal subject matter, not least across time. Based on the outcomes in our test cases, we believe topic modelling of legal texts represents an opportunity for new methodological problems and questions to be advanced, and for legal scholarship and analysis to render new insights whilst testing existing ones. Finally, we note that the techniques used herein have significant potential for extension into legal research, classification and search practices; a topic we will take up in future work.

II TOPIC MODELLING AND DISTANT READING

A Topic Modelling: Method and Use in the Digital Humanities

What Franco Moretti was describing when he popularised the concept of 'distant reading'⁸ was a maturation of earlier efforts at 'humanities computing', a field now generally known as the 'digital humanities'.⁹ This approach potentially offers both a radically different set of tools and a radically different methodological standpoint from which to engage with the literary and other text-based humanities.

As the context for the dominant application of topic modelling is the digital humanities, we believe it is best to present topic modelling in that same setting. Hence, in the following part we trace the basic contours of this collection of methods and the methodological tensions which arise in their application, showing how techniques and methods that emerged to study literary genre,¹⁰ geographical data,¹¹ Twitter,¹² the history of scholarship,¹³ and the collected

8 Franco Moretti, *Distant Reading* (Verso Books, 2013).

9 See generally Susan Schreibman, Ray Siemens and John Unsworth, *A New Companion to Digital Humanities* (John Wiley & Sons, 2016).

10 Sarah Allison et al, 'Quantitative Formalism: An Experiment' (Literary Lab Pamphlet No 1, University of Stanford, 15 January 2011) <<https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf>>.

11 Benjamin M Schmidt, 'Words Alone: Dismantling Topic Models in the Humanities' (2012) 2(1) *Journal of Digital Humanities* <<http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/>>.

12 See, eg, Wayne Xin Zhao et al, 'Comparing Twitter and Traditional Media Using Topic Models' in Paul Clough et al (eds), *Advances in Information Retrieval* (Springer, 2011) 338; Anastasia Giachanou, Morgan Harvey and Fabio Crestani, 'Topic-Specific Stylistic Variations for Opinion Retrieval on Twitter' in Nicola Ferro et al (eds), *Advances in Information Retrieval* (Springer, 2016) 466.

13 Andrew Goldstone and Ted Underwood, 'What Can Topic Models of PMLA Teach Us about the History of Literary Scholarship?' on Ted Underwood, *The Stone and the Shell* (14 December 2012)

works of St Thomas Aquinas,¹⁴ among other topics,¹⁵ might come to be applied to legal textual corpora.

The purpose of topic modelling is to ‘uncover evidence already in the text’,¹⁶ producing a model of the text corpora that ‘formalizes our definition of subject matter’.¹⁷ The topic model is produced by application of quantitative textual analysis, a computational process used to identify repeated occurrences of collections of words. Megan Brett describes the process of topic modelling by analogy with manual text analysis:

One way to think about how the process of topic modelling works is to imagine working through an article with a set of highlighters. As you read through the article, you use a different color for the key words of themes within the paper as you come across them. When you were done, you could copy out the words as grouped by the color you assigned them. That list of words is a topic, and each colour represents a different topic.¹⁸

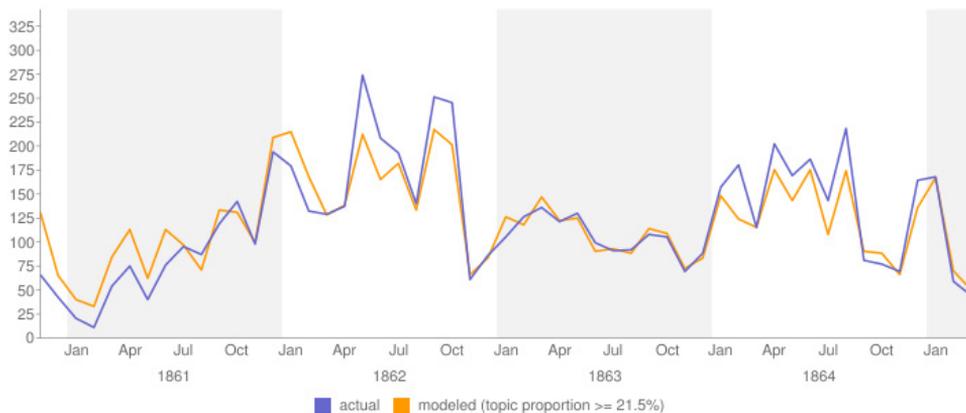
Thus, each topic is characterised by a series of words or ‘tokens’ that the topic model identifies as most likely to appear in documents belonging – in part – to that topic. For example, in work modelling press releases issued by United States (‘US’) senators, the topic model returned the tokens ‘border, homeland, immigr, patrol, secur, cross, agent, mexico, illeg, dh’¹⁹ as identifying a topic, in descending order of weight. The topic was labelled ‘border’ by scholars, a topic that related to ‘border security’. Ted Underwood illustrates this point with another example, highlighting how topic modelling is probabilistic, where different words have different probabilities of appearing in passages which discuss a ‘topic’: where ‘[o]ne topic might contain many occurrences of “organize,” “committee,” “direct,” and “lead.” Another might contain a lot of “mercury” and “arsenic,” with a few occurrences of “lead.”’ Thus, to quote Underwood further, ‘[t]opic modeling is a way of extrapolating backward from a collection of documents to infer the discourses (“topics”) that could have generated them’.²⁰

<<https://tedunderwood.com/2012/12/14/what-can-topic-models-of-pmla-teach-us-about-the-history-of-literary-scholarship/>>.

- 14 Steven E Jones, *Roberto Busa, S J, and the Emergence of Humanities Computing: The Priest and the Punched Cards* (Routledge, 2016).
- 15 See Schreibman, Siemens and Unsworth, above n 9, 303.
- 16 Megan R Brett, ‘Topic Modeling: A Basic Introduction’ (2012) 2(1) *Journal of Digital Humanities* <<http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/>>.
- 17 Michael A Livermore, Allen Riddell and Daniel Rockmore, ‘A Topic Model Approach to Studying Agenda Formation for the US Supreme Court’ (Virginia Law and Economics Research Paper No 2, University of Virginia School of Law, 10 July 2015) 10; for a more recent version of that paper see Michael A Livermore, Allen Riddell and Daniel Rockmore, ‘Agenda Formation and the US Supreme Court: A Topic Model Approach’ (SSRN Scholarly Paper, 29 February 2016).
- 18 Brett, above n 16.
- 19 Justin Grimmer, ‘A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases’ (2010) 18 *Political Analysis* 1, 15. See also Justin Grimmer and Brandon M Stewart, ‘Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts’ (2013) 21 *Political Analysis* 267.
- 20 Ted Underwood, ‘Topic Modeling Made Just Simple Enough’ on Ted Underwood, *The Stone and the Shell* (7 April 2012) <<https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>>. See also the fable-like explanation by Matthew Jockers, ‘The LDA Buffet Is Now Open; Or, Latent Dirichlet Allocation for English Majors’ on Matthew Jockers, *Matthew L Jockers* (29 September 2011)

Topic models have been ‘fit’ to a range of texts. One high-profile instance in this regard is the work of Robert K Nelson, who, in *Mining the Dispatch*, mined and then fit a model to the almost-complete run of the *Daily Dispatch*, a daily newspaper published in Richmond, Virginia, at one point the capital of the US Confederacy.²¹ Modelling the (now digitised) text from November 1860 through April 1865,²² Nelson constructed a view of the ‘social and political life of Civil War Richmond’,²³ a view unlikely to be achievable through human categorisation. Using his topic model, Nelson reports the number of items published in the newspaper where the proportion of the item was classified with the topic ‘fugitive slave ad’²⁴ as equal to or greater than 21.5 per cent. When overlaid on the actual human count of fugitive slave advertisements, the accuracy of the model in representing the underlying text is evident.

Figure 1: Modelled v Actual Count – ‘fugitive slave ad’ Items in the *Daily Dispatch*²⁵



<<http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/>>.

- 21 Robert K Nelson, *Mining the Dispatch*, University of Richmond Digital Scholarship Lab <<http://dsl.richmond.edu/dispatch/pages/home>>; Robert K Nelson, ‘Of Monsters, Men – And Topic Modeling’, *The New York Times* (online), 29 May 2011 <<http://opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/>>; Jeffrey W McClurken, ‘*Richmond Daily Dispatch*, 1860–1865 and *Mining the Dispatch*’ (2012) 99 *Journal of American History* 386.
- 22 Digitisation has been central to facilitating a range of legal scholarly methods, both topic modelling as well as historical and other methods: see, eg, David J Carter, ‘Correcting the Record: Australian Prosecutions for Manslaughter in the Medical Context’ (2015) 22 *Journal of Law and Medicine* 588 which utilises the National Library of Australia’s Trove Digitised Newspapers collection. See especially Katherine Biber’s exploration of, among other things, the impact of digitisation of records for legal and other scholarship: Katherine Biber, ‘In Jimmy Governor’s Archive’ (2014) 42 *Archives and Manuscripts* 270; see particularly at 277–8.
- 23 Robert K Nelson, *Mining the Dispatch: Introduction*, University of Richmond Digital Scholarship Lab <<http://dsl.richmond.edu/dispatch/pages/intro>>.
- 24 A topic which consists of the following predictive words: ‘negro years reward boy man named jail delivery give left black paid pay ran color richmond subscriber high apprehension age ranaway free feet delivered’: see Robert K Nelson, *Mining the Dispatch: Fugitive Slave Ads*, University of Richmond Digital Scholarship Lab <<http://dsl.richmond.edu/dispatch/Topics/view/15>>.
- 25 Used with permission (copy on file with authors).

Nelson describes his surface reading technique, and its graphical representation, as speaking to an important underlying social reality. This echoes a common trope in the field, where *purpose* is key to methodological concerns; for Nelson, his graphing is

an abstraction – a powerful and moving abstraction inasmuch as it evidences the courageous choices that many enslaved men and women made to attempt to escape their individual enslavement and to challenge and compromise the institution of slavery.²⁶

Nelson argues that the two sustained ‘spikes’ in the occurrence of the fugitive slave topic align with those periods where the Union army’s line approached Richmond, namely the summers of 1862 and 1864.²⁷ This analysis makes possible the development of Nelson’s concept of a ‘mobile North’,²⁸ a dynamic revealed in graphic form in the relationship between increased physical proximity of the free states of the North and the opportunity for enslaved African Americans to seize risky opportunities to escape when Yankee lines were close to Richmond.²⁹

Such powerful and moving abstraction is made possible by the *scale* at which these data are represented. And it is topic modelling deployed by digital humanists on large text corpora which is commonly used to achieve this *distant reading scale*. As to the effect of scale, ‘distant reading’ is one conception that has gained significant traction – approaching text not through ‘close’ reading of individual or small collections of texts, but rather attempting to understand text by aggregating very large collections, often spread over a time period. Distant reading is a method of ‘processing content in (subjects, themes, persons, places etc) or information about (publication date, place, author, title) a large number of textual items’³⁰ without actually *reading* the text itself. On this question of reading Moretti is clear, arguing that ‘we know how to read texts, now let’s learn how *not* to read them’.³¹ Moretti’s own choice of scale at which to read texts is not a scholarly practice resulting from the availability or utility of particular digital humanities techniques alone; rather, as he notes somewhat polemically, it is a choice about what counts as a text worthy of being read, a claim made in critique of canon formation:

the trouble with close reading (in all of its incarnations, from the new criticism to deconstruction) is that it necessarily depends on an extremely small canon. This may have become an unconscious and invisible premise by now, but it is an iron one nonetheless: you invest so much in individual texts *only* if you think that very few of them really matter.³²

26 Nelson, *Introduction*, above n 23. See also Nelson, ‘Of Monsters’, above n 21.

27 Nelson, *Introduction*, above n 23.

28 Ibid.

29 Ibid.

30 Johanna Drucker, ‘Distant Reading and Cultural Analytics’ in Johanna Drucker et al, *Introduction to Digital Humanities* (Coursebook, UCLA Centre for Digital Humanities, 2014) <http://dh101.humanities.ucla.edu/?page_id=62>.

31 Franco Moretti, ‘Conjectures on World Literature’ (2000) 1 *New Left Review* 54 (emphasis in original).

32 Moretti, *Distant Reading*, above n 8, 48 (emphasis in original).

Nor is it the end of attentiveness to text, so much as a question of text selection and use. On this point, Shawna Ross confirms as much for distant reading in the humanities, stating clearly that:

Even distant reading does not work by *not* reading ... Moretti means selective reading (reading only the titles, only the first paragraphs, or scanning for certain patterns), delegated reading (recruiting his graduate students), or mediated reading (using search tools to generate statistics and charts).³³

B Application to Legal Text

Livermore, Riddell and Rockmore argue that ‘studying the text of appellate court decisions is a mainstay of traditional legal scholarship’.³⁴ Yet what it means to read such texts, the question of canon and of scale – of what counts as a text worthy of being read and ‘how’ such texts are best read – is an already integral part of the very practice of legal reading. For law, textual selection and use is complex, with the development and contest of doctrine produced through the shifting of inclusion and exclusion in ‘canonical formations’, a process that results in binding precedent. The practice of reading at differing scales, too, is more commonplace than it may seem. Far from the tensions present in digital humanities methodological discussions, legal interpretation and scholarship *require and emerge from* the explicit and ongoing shift between attention to sentence, vocabulary, word order and style which exemplifies the ‘close’ reading scale, and the more ‘distant’ reading scale marked by techniques of reconfiguring large collections of cases by topic, drawing out and constructing legal themes or subject matter, alongside the reading of legal texts at the very great distance of subsequent social practice and effect. Thus, the tension between distant and close reading scales is, for law at least, a productive tension, essential to the very textual and social practices of lawyering in all its forms.³⁵

Although applying a distant reading paradigm to legal text is not a practice alien to legal scholarship, topic modelling presents a relatively new way of achieving this productive shift in scale. Whilst a less frequent subject of the technique, topic modelling has been very recently applied to socio-legal studies of legal and political transition in Myanmar,³⁶ to US Supreme Administrative Court decisions³⁷ and the *combined* case corpora of the US Supreme Court and

33 Shawna Ross, ‘In Praise of Overstating the Case: A Review of Franco Moretti, *Distant Reading* (London: Verso, 2013)’ (2014) 8(1) *Digital Humanities Quarterly* [6] (emphasis in original) <<http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html>>.

34 Livermore, Riddell and Rockmore, ‘Agenda Formation’, above n 17, 3.

35 These tensions have been explored in the field of empirical legal studies, however, we propose, tentatively, that digital humanities techniques, whilst related and even overlapping in methodological standpoint, differ from empirical legal studies in some material ways. Although topic modelling might well be interpreted as a tool of legal empiricism, and perhaps it should, to do so without further thought risks eliding the important historical separation, along with the methodological and textual biases which are different for both schools.

36 Melissa Crouch and Tim Lindsey (eds), *Law, Society and Transition in Myanmar* (Bloomsbury Publishing, 2014).

37 Ingo Feinerer and Kurt Hornik, ‘Text Mining of Supreme Administrative Court Jurisdictions’ in Christine Preisach et al (eds), *Data Analysis, Machine Learning and Applications* (Springer, 2008) 569.

Appellate Court,³⁸ and to Macey and Mitts' dataset of corporate veil-piercing cases,³⁹ a subset of 9380 US combined federal and state cases that reference either corporate veil piercing or other forms of disregarding the corporate form.⁴⁰ So, too, has Joshua Mitts utilised popular topics derived from Google's *n*-gram viewer to formulate a predictive model for regulatory issues that will produce market-disrupting events.⁴¹ The work of Daniel Taylor Young, which topic models constitutional change in the US, uses topic modelling to test 'empirically' a theory of such change developed by Bruce Ackerman using more traditional legal analysis.⁴² Livermore, Riddell and Rockmore also utilise a topic modelling approach in their developing project on agenda formation in the US Supreme Court.⁴³ *The Old Bailey Online*,⁴⁴ along with individual works that have flowed from that resource,⁴⁵ is a notable example of applying digital humanities methods, including topic modelling, to legal institutions and text. The now digitised records of proceedings of London's central criminal court, stretching from 1674 to 1913, have produced what project leaders describe as the largest body of texts 'detailing the lives of non-elite people ever published'.⁴⁶ Over 10 years of work on these texts has seen legal scholarship facilitated by a variety of digital humanities methods.⁴⁷ Individual works made possible by the project reveal information about the Court and its case load, including the overall case mix and the subject matter of cases.⁴⁸

38 Livermore, Riddell and Rockmore, 'A Topic Model Approach', above n 17.

39 Jonathan Macey and Joshua Mitts, 'Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil' (2014) 100 *Cornell Law Review* 99, 149.

40 Ibid 141.

41 Joshua Mitts, 'Predictive Regulation' (SSRN Scholarly Paper, 27 June 2014) <<http://papers.ssrn.com/abstract=2411816>>.

42 Daniel Taylor Young, 'How Do You Measure a Constitutional Moment? Using Algorithmic Topic Modeling to Evaluate Bruce Ackerman's Theory of Constitutional Change' (2013) 122 *Yale Law Journal* 1990.

43 See their draft paper: Livermore, Riddell and Rockmore, 'Agenda Formation', above n 17.

44 Tim Hitchcock et al, *The Old Bailey Proceedings Online, 1674–1913 (Version 7.2)* (9 August 2016) (March 2015) <<http://www.oldbaileyonline.org/>>.

45 See, eg, Krisda Chaemsaitong, 'Interactive Patterns of the Opening Statement in Criminal Trials: A Historical Perspective' (2014) 16 *Discourse Studies* 347; Caroline Anne Forell, 'Convicts, Thieves, Domestic, and Wives in Colonial Australia: The Rebellious Lives of Ellen Murphy and Jane New' (SSRN Scholarly Paper, 9 June 2012) <<http://papers.ssrn.com/abstract=2080526>>; Merja Kytö, 'Data in Historical Pragmatics' in Wolfram Bublitz, Andreas H Jucker and Klaus P Schneider, *Handbooks of Pragmatics* (De Gruyter Mouton, 2008) vol 8, 33; Ian Langford, 'Fair Trial: The History of an Idea' (2009) 8 *Journal of Human Rights* 37; C J Leppard-Quinn, *The Unfortunates: Prostitutes Transported to Van Diemen's Land 1822–1843* (PhD Thesis, University of Tasmania, 2013) <<http://eprints.utas.edu.au/17536/>>; Garthine Walker, 'Everyman or a Monster? The Rapist in Early Modern England, c1600–1750' (2013) 76 *History Workshop Journal* 5; Garthine Walker, 'Rape, Acquittal and Culpability in Popular Crime Reports in England, c1670–c1750' (2013) 220 *Past & Present* 115; T P Gallanis, 'The Mystery of Old Bailey Counsel' (2006) 65 *Cambridge Law Journal* 159; Drew Gray, 'Putting Undergraduates on Trial: Using the Old Bailey Online as a Teaching and Assessment Tool' (2014) 4 *Law, Crime and History* 104. See also a listing maintained by the Old Bailey project: Tim Hitchcock et al, *The Old Bailey Proceedings Online: Publications That Cite the Old Bailey Proceedings Online* (17 August 2016) Online <<http://www.oldbaileyonline.org/static/Publications.jsp>>.

46 Tim Hitchcock et al, *The Old Bailey Proceedings Online, 1674–1913*, above n 44.

47 Some 309 publications are tracked by the project as using the data from the Old Bailey project.

48 This first stage is that which we present here in relation to the High Court.

The application of topic modelling reported here complements other Australian work on both the High Court, and longitudinal studies of the state supreme courts completed by others. In relation to the High Court specifically, even though Lynch and Williams do not situate their work explicitly within a ‘distant reading’ paradigm, their work achieves as much and, as noted above, we support their framing of the ongoing provision of statistical survey data of the Court’s activity as explicitly ‘intended to complement substantive analyses of the Court’s work’.⁴⁹ This is important work which our application of topic modelling hopes to complement through provision of data regarding the Court’s changing patterns of case load over a longer period of time. In our own terms, we see the work presented here as useful for legal scholarship undertaken on a ‘mixed-scale’ between distant and close reading that we, like Lynch and Williams, believe ‘is important in indicating, and occasionally verifying, conjecture or theories about how the Court is functioning at any point in time, it can never, obviously, tell the whole story’.⁵⁰

III RESEARCH DESIGN: TOPIC MODELLING THE HIGH COURT

A Study Design

Our primary aim was to develop the topic model itself, testing the most appropriate number of topics for the corpus and thus proving the usefulness of the topic model(s) we developed by reference to their validity as judged by an experienced human interpreter. Our second aim was to establish the basic feasibility of topic modelling as a method for the study of Australian legal texts and the legal institutions that produce them.

We produced a series of models with variable numbers of topics. We generated topics constituted by 10, 15, 20, 50 and 100 topics in aid of comparison and testing of the usefulness of each topic.⁵¹ We report on the 10 and 50 topic models developed to fit the corpus.

This question of the ‘usefulness’ of a topic model has a pedigree in legal scholarship. Formal models of subject matter, like those produced through the topic modelling process, are ubiquitous. Legal materials are commonly classified according to a range of taxonomies that are embedded in almost any encounter with legal text. In relation to the legal subject matter of a text, the definitional

49 Lynch and Williams, above n 5, 1078.

50 Ibid.

51 This question of the use of topic models is the subject of discussion in research currently underway. Usefulness is a central question in relation to the formation and use of taxonomies in law, of which a topic model forms one example. As the number of topics expand, there is a risk that the model comes to overfit the underlying material, yet, the topic model itself may still in some sense remain ‘useful’ for the human interpreter or for legal analysis of the underlying text. As such, it should be remembered that topic models, are just that: models. Our aim here has been to present varying forms of topic model to test their usefulness as a model of underlying textual reality and practice, rather than only those which we feel are wholly optimised from a technical standpoint which may well be a different undertaking.

scheme is settled: ‘constitutional law’, ‘criminal law’, ‘corporate law’, ‘administrative law’ and others. This taxonomy is embedded in editorial head notes, search strings, keyword searches and even the ‘Priestley Eleven’ core syllabus of the Australian law curriculum.⁵² This formal division of subject matter serves as the ‘bedrock’⁵³ of contemporary legal research, and structures legal thinking in accordance with the accepted taxonomical structure.⁵⁴ These questions are absolutely central to the development of topic models, a process that leverages quantitative techniques in an attempt to represent the underlying subject matter of a corpus of documents in an accurate and useful way.⁵⁵

Our corpus consists of 7476 decisions of the High Court of Australia (‘HCA’) spanning the years 1903–2015. It does not include transcripts, High Court Bulletins, special leave dispositions or other material also available in digital form. We appreciate that our corpus represents only one part of the judicial workload of the Court; however, these cases represent a manageable and meaningful corpus upon which to develop a topic model.⁵⁶ The decisions were sourced from the HCA’s own eResources repository of the Commonwealth Law Reports (‘CLR’) developed by BarNet for the Court,⁵⁷ as well as collections from the Australian Legal Information Institute (‘Austlii’).⁵⁸ Whilst the size of the corpus is by many traditional standards exceedingly large, for a topic modelling approach, the corpus is in fact quite small. For example, Nelson utilised over 112 000 items comprising approximately 24 million words in his *Mining the Dispatch* project,⁵⁹ whilst in the specifically legal context Young modelled some 24 934 documents, totalling 32 544 870 words,⁶⁰ and the ongoing work of Livermore, Riddell and Rockmore models the joint corpus of the US Supreme Court (7598 cases) and Appellate Court (a random sample of 25 000 cases in addition to 4180 cases referenced by the Supreme Court) between 1951 and 2007.⁶¹

52 See Jill McKeough, ‘Graduate Attributes – The Priestley Areas of Knowledge and the Broader Educational Context’ (Paper presented at the National Symposium on Internationalising the Australian Law Curriculum for Enhanced Global Legal Education and Practice, Canberra, 16 March 2012) 7–8 <http://curriculum.cald.asn.au/media/uploads/9_8_Graduate_attributes.pdf>.

53 Livermore, Riddell and Rockmore, ‘A Topic Model Approach’, above n 17.

54 See especially Peter Birks, ‘Equity in the Modern Law: An Exercise in Taxonomy’ (1996) 26 *University of Western Australia Law Review* 1; Emily Sherwin, ‘Legal Taxonomy’ (2009) 15 *Legal Theory* 25.

55 Our application of topic modelling means that the entire corpus (1903–2015) is regarded as a single entity with topics computed over the corpus. In this approach, as is common to most applications of topic modelling, topics are therefore assumed to be static across the corpus and across time.

56 Following the excellent work of Livermore, Riddell and Rockmore, we believe that inclusion of these materials, particularly special leave applications, represents a fruitful and worthwhile opportunity to understand the agenda ‘formation’ of the High Court: Livermore, Riddell and Rockmore, ‘Agenda Formation’, above n 17. This will be the subject of future research.

57 High Court of Australia, *High Court Judgments Database* <<http://eresources.hcourt.gov.au>>.

58 Early trials included use of the Linguistic Inquiry and Word Count (‘LIWC’) tool: James W Pennebaker et al, *Linguistic Inquiry and Word Count: LIWC 2015* (Pennebaker Conglomerates, 2015) <www.liwc.net>. See also Pennebaker’s fascinating work on pronouns which utilises the LIWC program: James W Pennebaker, *The Secret Life of Pronouns: What Our Words Say about Us* (Bloomsbury Publishing, 1st ed, 2011).

59 Nelson, *Mining the Dispatch*, above n 21.

60 Young, above n 42, 2021.

61 Livermore, Riddell and Rockmore, ‘Agenda Formation’, above n 17.

We generated a set of topics from our corpus using a machine learning algorithm called Latent Dirichlet Allocation ('LDA'). Here we explain the process of topic modelling as well as some of the specific features of LDA.

Traditional human interaction with documents (electronic or otherwise) proceeds by way of reading. Such reading shows us that in any collection of documents (corpus), words that relate to the same concept or discourse (topic) tend to co-occur more frequently within the same document than words that share little relationship to one another. Consider for instance an Australian newspaper. If we treat the newspaper as our corpus, and the articles within as our documents and we see the words 'bat' and 'innings' appearing several times in a document, we would expect to also see the word 'cricket'. Conversely, if we see the words 'company' and 'merger' in another document, we would not really expect to see the word 'cricket' in that same document. This seemingly trivial classification task relies on skills that humans take many years to develop. Furthermore, in a newspaper, we gain advantage from the fact that topics for our corpus have been preformed; they are the sections of the newspaper.

At its most basic level, topic modelling aims at capturing the essence of this classification process in a mathematical form that makes it possible for an algorithm to construct topics from a corpus. Several methods have been used to accomplish this task. In this work we use LDA.⁶² While we refer the reader interested in the technical details to the references cited, some features of LDA are worth discussion. Using LDA, the number of topics is selected a priori. The algorithm uses statistical inference to allocate each document across all topics. In contrast to the newspaper metaphor above, where an article can be said to belong to the sports section, or the business section, using LDA, a document is described as a probability distribution over all topics. In other words, for a given document (a judgment in this application of LDA), a set of weights describing the contribution to or presence of each topic in that document. In practice, some topic weights may be so small so as to be negligible. However, it remains true that all documents are *mixtures* of topics. This is an important feature, as it allows the model to capture and express more subtle nuances of a document. In a similar fashion, topics themselves are probability distributions over all words in the corpus. This means that a given word makes some contribution to *all* topics. This is particularly useful in handling polysemy. For instance, the word 'bat' in our newspaper metaphor, may appear in documents about cricket and in documents about nocturnal mammals. By requiring that words can have different weights in different topics, LDA facilitates the capture of semantics in a more realistic way.

62 Matthew D Hoffman, Francis Bach and David M Blei, 'Online Learning for Latent Dirichlet Allocation' in J D Lafferty et al (eds), *Advances in Neural Information Processing Systems 23* (Neural Information Processing Systems, 2010) 856 <<http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf>>; David M Blei, Andrew Y Ng and Michael I Jordan, 'Latent Dirichlet Allocation' (2003) 3 *Journal of Machine Learning Research* 993. For a more accessible discussion, see also David M Blei, 'Probabilistic Topic Models' (2012) 55(4) *Communications of the ACM* 77; David M Blei, 'Topic Modeling and Digital Humanities' (2012) 2(1) *Journal of Digital Humanities* <<http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/>>.

For our corpus, our goal was to fit a set of K topics to the 1903–2015 period. There is no real way to define a ‘correct’ value K a priori, so we settled on values that would fit our objective of reading the Court at distance. The number of topics needed to be large enough to capture some of the more important nuances of the work of the Court, but not too large for a reader to grasp at a glance.⁶³

Whilst the text required a series of pre-processing procedures, our aim was to reduce as far as possible human intervention in preparing cases for the topic modelling process. We started our analysis by transforming our corpus into a ‘bag of words’. In practice, the vocabulary is not merely the set of distinct words in the corpus. Many words, like ‘the’, ‘a’, ‘I’, ‘of’ or ‘by’, are very common but provide little insight. These ‘stopwords’ are filtered out.⁶⁴ It is also important to filter out frequent words that are specific to the corpus but hold little semantic information. For the HCA corpus, the names of the justices of the Court, or words like ‘plaintiff’ or ‘court’ fall into this category. We also excluded words that appear in fewer than 50 documents or which are present in more than 50 per cent of the corpus.⁶⁵

Finally, the vocabulary is not confined to single words but can be extended to include n -grams, sets of n words that appear consecutively in the text (after filtering out the stopwords). In this work we considered both single words and bigrams, which appear as pairs of words separated by an underscore in our results. We emphasise that no specific processing at the level of individual documents, such as removing the references, was performed. Our goal was to have a method that allows us to gain some insight into the corpus, without requiring a human reader to process the data first.

Upon completion of pre-processing we were left with a vocabulary of 44 759 tokens (words and bigrams). Computation was performed using the Python Gensim library,⁶⁶ which contains an efficient implementation of the LDA algorithm.⁶⁷ We typically streamed the corpus through the LDA algorithm 1000 cases at a time, and the whole corpus was processed 500 times to ensure that

63 For a deeper consideration of this issue, see Eleanor Rosch, ‘Principles of Categorization’ in George Mather, Frans Verstraten and Stuart Anstis (eds), *The Motion Aftereffect: A Modern Perspective* (MIT Press, 1998) 251.

64 This is a process which can make an immense difference to the quality of the model and its output. Here we erred on the side of reducing the number of stopwords to the very fewest possible.

65 The rationale for filtering out words that are either too frequent or not frequent enough across the corpus is the following. First, words that appear in only a handful of cases are unlikely to be representative of the larger focus of the Court. The threshold of 50 documents, which represents less than one per cent of the corpus, was chosen empirically to satisfy this requirement and reduce the size of the vocabulary. Second, filtering out words that appear in more than 50 per cent of the cases ensures that words that are too common to be helpful in discriminating between documents are also excluded. These two criteria were chosen in the light of our goal of reading the Court at a distance. Note that tightening the second criterion (eg, excluding words that appear in less than 20 per cent of the corpus) would result in more ‘targeted’ topics (topics with fewer words describing global concepts such as ‘land’), which might be useful in another context, but which we deemed to be less useful for a first reading at a distance analysis. Tokens were lemmatised using Python natural language toolkit WordNet Lemmatiser.

66 Radim Řehůřek and Petr Sojka, ‘Software Framework for Topic Modelling with Large Corpora’ in René Witte et al (eds), *Proceedings of Workshop on New Challenges for NLP Frameworks* (University of Malta, 2010) 46. We used Gensim version 0.12.2.

67 Hoffman, Bach and Blei, above n 62.

topics had converged. The maximum number of iterations (for expectation maximisation algorithm) is 1000.⁶⁸ The alpha parameter for the Dirichlet was set to 'auto' to learn a suitable prior from the data, whilst all other parameters were left to their default values. Typical computation time was 40 hours CPU time, with data (7476 documents) randomly shuffled to avoid topic bias towards earlier cases. Perplexity has been proposed as one approach to testing how well a fitted LDA model generalises to an unknown document.⁶⁹ Unfortunately, the current version of Gensim seems to have some issues with perplexity computations at this time,⁷⁰ and the relevance of perplexity as a measure of topic coherence for our purposes is unclear.⁷¹ As an alternative, we tested the models against new (unseen) documents, namely, judgments of the Court reported in 2016 and found that these models performed well.⁷²

IV THE JUDICIAL WORKLOAD OF THE HIGH COURT

In this section, we address the question of the judicial workload of the Court. This is of course a partial view, as our model has been fit to a limited number of cases and activities undertaken by the Court.⁷³ However, a range of descriptive data related to the judicial workload of the Court is produced in the process of topic modelling. One such data set is the number of published decisions of the Court over its lifetime, as shown in the following figure.⁷⁴ The Court's published output has fluctuated over time, with notable shifts in aggregate case publishing during the very early years of the Court, during the 1920s and again during the Second World War. We take these shifts as invitations for further analysis by other means.

68 In our experience, increasing this setting from the default of 50 was more useful than increasing the number of passes beyond 500, with less impact upon the computation time.

69 Hoffman, Bach and Blei, above n 62; Blei, Ng and Jordan, above n 62.

70 Lev Konstantinovskiy, 'LDA: Increasing Perplexity with Increased No of Topics on Small Documents, Issue #701' on GitHub, *RaRe-Technologies/Gensim* (18 May 2016) <<https://github.com/RaRe-Technologies/gensim/issues/701>>.

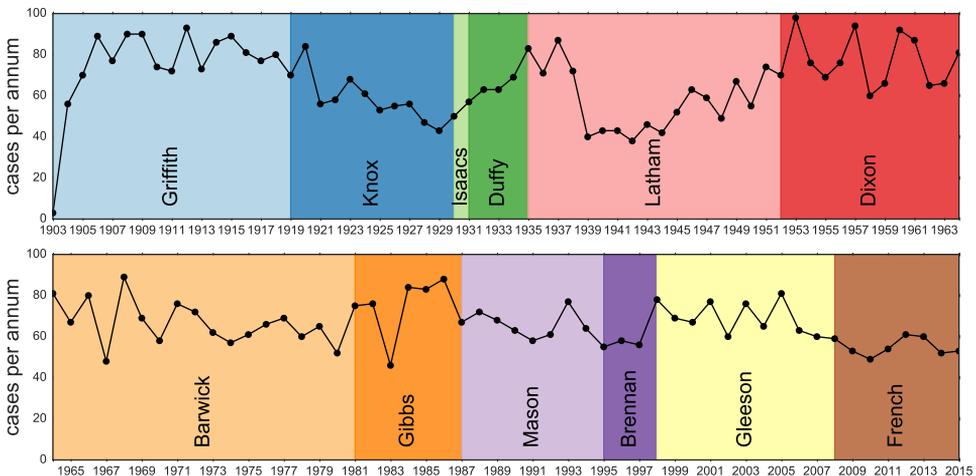
71 Jonathan Chang et al, 'Reading Tea Leaves: How Humans Interpret Topic Models' in Y Bengio et al (eds), *Advances in Neural Information Processing Systems 22* (Nural Information Processing Systems, 2010) 288.

72 Given the purpose of the study, we believe that this approach represents a potentially more suitable test of the models rather than perplexity. Results of this test are held on file by the authors, and form part of a planned publication. They may be obtained by contacting the authors.

73 Most notably limited by those cases which had been reported.

74 Coloured bands differentiate the results by the figure of the Chief Justice at the time, which for ease of reference has been 'rounded' to the year in which they served the majority of time in that position. Analysis of the Court grouped by the various 'natural courts' holds promise, and is the subject of ongoing research work.

Figure 2: High Court of Australia – Published Decisions 1903–2015



The first years of the Court brought an increase in the number of cases. This we believe to be due to the natural increase in any Court's volume at the advent of its life. The sudden 'dip' during the Second World War seems similarly unremarkable, no doubt the result of the diversion of attention to war-related issues that did not produce a significant workload for the Court.

The reduction in published cases during the Knox Court, however, requires further analysis. The timing seems unrelated to the advent or end of the First World War or to the generally accepted chronology of the Great Depression.⁷⁵ One potential explanation is that, after the early activity of the new Court, the Knox and Latham Courts represented a settling of the Court's workload, punctuated by the effects of the Great Depression during the Isaacs and Latham Courts. Certainly the shift during the Knox Court, from interpretation practices regarding the Constitution as 'political document' towards more text-based understandings of the Constitution as a 'legal document', has been documented by Anne Twomey.⁷⁶ How such a shift informs the pattern of lower judgment-making, if at all, is a question we provide some further commentary on below,⁷⁷ but one that requires more contextualisation than the topic model alone can offer.

Since 1953, there has been a progressive falling away in the number of cases published by the Court.⁷⁸ The more recent activity of the Court represents a relatively static period of output, beginning with the early years of the Mason Court (albeit with some year-on-year volatility during the Gleeson Court). The current French Court has published a number of cases broadly consistent with the

75 As beginning on or near the crash of Wall Street in 1929.

76 See Anne Twomey, 'The Knox Court' in Rosalind Dixon and George Williams (eds), *The High Court, the Constitution and Australian Politics* (Cambridge University Press, 2015) 98, 98.

77 See the 'Ten Topic Model' Part V(A) below.

78 Although this does not represent the complete judicial workload of the Court during this period, which, particularly recently has been growing significantly.

Latham Court during the Second World War and the very late years of the Latham and Brennan Courts. These three periods of the Court's history stand out against periods of higher case production. The number of the current Court's decisions is, however, as the Court itself notes, comparable with averages over the past 10 years.⁷⁹ The number of cases decided by the Court analysed here does not represent the entirety of the Court's judicial workload, with the increasing number of applications made for special leave to appeal now representing the greatest volume of matters filed with the Court.⁸⁰

This observation of a decline in the aggregate number of decisions finds further context in the trend towards shorter decisions that has marked the French Court, a trend that returns the Court to judgment lengths more consistent with the Court's history but not seen since the early period of the Brennan Court. Specifically, 2014 saw the shortest aggregate judgments (by average number of characters per judgment) by the Court, which is a return to a similar average judgment length 20 years prior, during the Mason Court in 1994.

The figure below represents the length of judgments by the average number of characters per judgment aggregated by year. Consistent with the aim of limiting human pre-processing, we have not differentiated between the judgments of individual justices on the Court, instead using single judgments as the finest level of detail. So too do these counts also include editorial comments and references. These data compare favourably with those produced by Groves and Smyth,⁸¹ in so far as our method produces similar trends to those reported there. However, the present data would be well served by a comparison with the raw data collated by those authors, who utilised a page count method of the CLR, whilst we utilised a count of characters.⁸² This could lead to a more accurate comparison with their data, and those of Lynch and Williams on the prevalence of concurring and joint judgments.⁸³

79 High Court of Australia, 'Annual Report 2014–2015' (Report, 12 November 2015) 19

<<http://www.hcourt.gov.au/assets/corporate/annual-reports/hca-annual-report-2014-15.pdf>>.

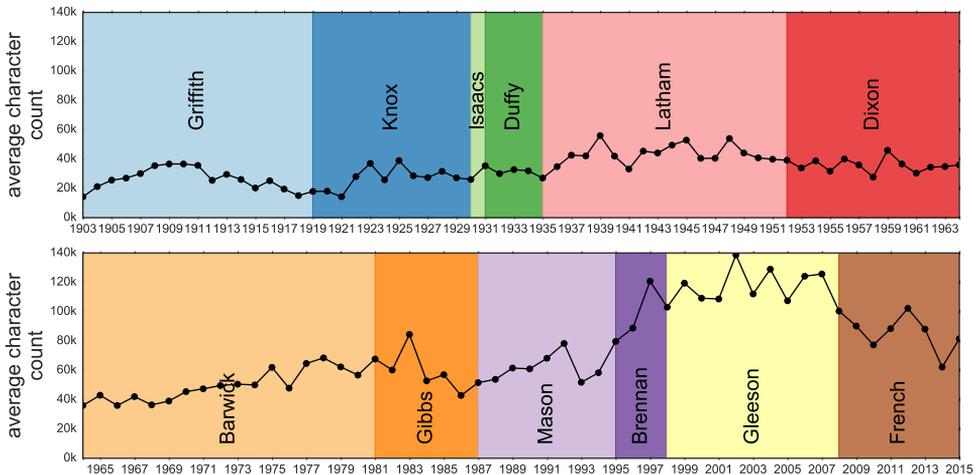
80 This question is the subject of ongoing research by colleagues Anita Stuhmcke and Pamela Stewart, Faculty of Law, University of Technology Sydney. See Pam Stewart and Anita Stuhmcke, 'High Court Negligence Cases 2000–10' (2014) 36 *Sydney Law Review* 586; Pamela Stewart and Anita Stuhmcke, 'Lacunae and Litigants: A Study of Negligence Cases in the High Court of Australia in the First Decade of the 21st Century and Beyond' (2014) 38 *Melbourne University Law Review* 151. The Court has recently introduced changes to its process for special leave applications. These changes to special leave announced in March 2016 now mean that all applications (represented or unrepresented) will be reviewed by a Panel of Justices on the papers to determine whether an oral hearing is warranted: see Andrew Phelan, 'Changes to High Court Procedures for Considering Applications for Special Leave' (Press Release, 16 March 2016) <http://www.hcourt.gov.au/assets/corporate/policies/Special_Leave_Changes.pdf>.

81 Groves and Smyth, above n 4, 259.

82 See *ibid* 258.

83 Lynch and Williams, above n 5.

Figure 3: Average Number of Characters per Decision (Aggregated by Year)



Whilst there has been a general trend towards shorter judgments during the French Court, there is also a clear ‘uptick’ in the Court’s judgment length during the period 2011–13. This has been the subject of some commentary, notably, in the work of Lynch and Williams, who point to a reduction in unanimous decisions (a low of 13 per cent in 2012) and an increase in split decisions (as high as 50 per cent in 2011).⁸⁴ They note that from 2011 until his retirement from the bench in 2013, Heydon J took to issuing separate judgments in aid of the process and principle of judicial independence within the Court.⁸⁵ This analysis is borne out in the results presented here.

Looking beyond the French Court, the average length of the Court’s judgments remained relatively stable for the first five or six decades of the Court’s life, with a slight upward trend in aggregate judgment length. This relative stability ends with an upward trend in aggregate length during the 1970s, with the Brennan and Gleeson Courts producing the lengthiest judgments in the Court’s history. Whilst the causes of this shift from the 1970s onwards are likely related at least partly to the advent of computer and word processing technology, more detailed understanding of the historical work practices of the Court, and its individual judges, would be required to confirm such a hypothesis.

V LEGAL SUBJECT MATTER OF THE HIGH COURT

In what follows, we provide an account of the development and results of the core of our topic modelling process. We do so by presenting the results of our 10

⁸⁴ Lynch and Williams, above n 5, 1080.

⁸⁵ See *ibid* 1080, n 3: Dyson Heydon, ‘Threats to Judicial Independence: The Enemy Within’ (2013) 129 *Law Quarterly Review* 205.

and 50 topic models. We then test the topic model by applying its output to a number of questions about the Court's workload.

A Ten Topic Model

Table 1 provides a list of the topics generated by the 10 topic model. The words (more correctly 'tokens') listed for each topic appear in descending order, from the highest to least weighted words in the distribution.⁸⁶ The order of the topics themselves is random. The 'Topic Label' is the label generated by researchers to describe the topic during analysis.

Table 1: 10 Topic Model

Topic	Words	Topic Label
0	'minister', 'tribunal', 'australian', 'review', 'protection', 'conduct', 'submission', 'convention'	Government action in relation to immigration
1	'child', 'trust', 'property', 'trustee', 'estate', 'wife', 'death', 'husband'	Beneficiaries' rights to property in an estate
2	'trial', 'offence', 'criminal', 'jury', 'accused', 'crime', 'police', 'prosecution'	Trial process for criminal matters
3	'trade', 'licence', 'board', 'mark', 'sale', 'patent', 'price', 'business'	Trade licencing, regulation and IP
4	'damage', 'negligence', 'injury', 'loss', 'liability', 'care', 'reasonable', 'trial'	Damage to persons through injury
5	'company', 'income', 'share', 'commissioner', 'money', 'assessment', 'business', 'payment'	Company financial flows
6	'jurisdiction', 'federal', 'judicial', 'applicant', 'hearing', 'officer', 'federal_court', 'document'	Jurisdictional divisions and actions
7	'award', 'employee', 'dispute', 'employer', 'industrial', 'employment', 'compensation', 'work'	Employment entitlements and disputes
8	'land', 'contract', 'agreement', 'property', 'lease', 'title', 'sale', 'owner'	Land contracts and agreements
9	'constitution', 'parliament', 'government', 'regulation', 'territory', 'legislative', 'federal', 'constitutional'	Constitutional actors and relationships

These labels were generated by the research team through an iterative process, beginning with an intuitive label based on the dominant words in each topic and then refined by analysis of the cases the model categorised as

⁸⁶ As described above at Part III(A) we note that each topic is a mixture of *all of the words* in the corpus' vocabulary. For each topic, most words have a negligible weight. We thus list the terms most associated with a topic, that is, here the eight terms which make the largest contribution to the topic. So in that sense, these most dominant words are understood to 'form' the topic.

containing/reflecting the greatest alignment with that topic.⁸⁷ For example, Topic 1 consisted of the words ‘child’, ‘trust’, ‘property’, ‘trustee’, ‘estate’, ‘wife’, ‘death’ and ‘husband’. The most highly weighted word in the distribution was ‘child’, which was followed immediately by four words of legal terminology or mechanism until the reoccurrence of words relating to a legal subject (wife, husband) separated by the word ‘death’. This topic is one of the more varied in the 10 topic model; unlike others, it involves a mixture of legal subjects (child, wife, husband), legal concepts (property) and legal technology (trust, estate). Intuitively, this topic seems to involve questions of the ordering and conflicts surrounding estates and inheritance, whilst also perhaps questions of ‘family law’. (It is interesting that the weighting of the legal subjects listed in the topic inversely mirrors the order of the legal agency typically ascribed to such subjects: ‘child’, here representing the highest weight in the distribution, is also the legal agent with the least legal agency generally ascribed to it.) The single case that embodied the highest classification according to this topic (97.6 per cent of the case related to the topic) was one decided by the Court in 1948, *Marks v Trustees Executors & Agency Co Ltd*.⁸⁸ Manually reviewing the full text of the case shows that the editorial catchwords in the headnote of the CLR version are ‘Will – Construction – “Vested” – Vested in interest – Vested in possession’.⁸⁹ In this case, the Court dismisses an appeal relating to the vesting of real property held on trust, in favour of the testator’s children and their subsequent children. This topic’s four next most relevant cases each concern the same matters.⁹⁰ Each pertained to the interpretation of a will, with the catchwords, where present, representing the subject matter as principally that of ‘will’ and its ‘construction’. The prominence of the word ‘children’ is clear, as the central beneficiaries of inheritances in most cases.

The topic’s occurrence as a portion of the Court’s *total annual legal subject matter* can similarly be represented by its occurrence over time as a proportion of the total subject matter of the Court:

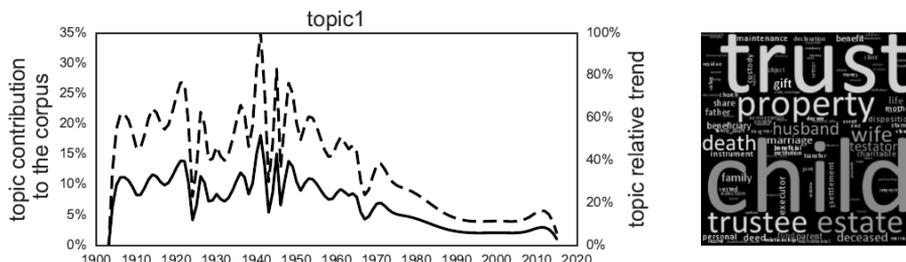
87 We appreciate that these labels are themselves a central area of contest and debate, and in some senses would prefer to refrain from providing a label, which of its nature must do some sort of ‘injustice’ to the underlying topic itself. See our commentary in Part VII, in relation to the work of Rachel Buurma in particular, for further information on the difficult tension between – and need to maintain such tension – standard ‘given’ legal taxonomies and those generated and labelled here.

88 (1948) 77 CLR 497.

89 Ibid 497.

90 *Duncan v Equity Trustees Executors & Agency Co Ltd* (1958) 99 CLR 513; *Kenna v Conolly* (1938) 60 CLR 583; *Sumpton v Downing* (1947) 75 CLR 76; *Russell v Perpetual Trustee Co (Ltd)* (1956) 95 CLR 389.

Figure 4: Timeline and Token, Topic 1 (10 Topic Model); solid line indicates a topic's contribution to the corpus in that year, dotted line is the topic's relative contribution normalised to the highest single year contribution



In this chart, we have plotted the occurrence of the topic over time, smoothing the output.⁹¹ The topic contribution or ‘presence’ is calculated as total weight of a particular topic across all cases in the model.⁹² More specifically, the solid line indicates that the topic accounts for approximately 10 per cent of the Court’s total subject matter until approximately 1940, reducing to approximately 2.5 per cent from about 1990 until 2015. The dashed trend line shows the same trend normalised to its highest value, so as to enhance the variation in the less important topics.

We have identified a single ‘representative’ case for each topic in Table 2, below. Again, by this we mean that the cases listed below contain the highest proportion or coverage of their nominated topic. This does not mean they stand as central or leading cases in the usual sense of the word, nor as landmark cases in their topic area. Rather, they represent a case essentially dedicated to a single topic to the exclusion of others. For example, approximately 86 per cent of *Minister for Immigration and Border Protection v WZAPN* (2015) 254 CLR 610 is classified as belonging to Topic 0, ‘Government action in relation to immigration’. Recalling the words or tokens associated in descending weight with that topic,⁹³ it is clear here that this topic describes administrative decision-making, with a focus or concentration on questions of ‘review’ and immigration-related terminology such as ‘protection’ and ‘convention’. This analysis is borne out in the classification of this case as representative, concerning as it does the Court’s confirmation that a period of temporary administrative detention of a person in relation to a reason specified in the 1951 *Convention Relating to the*

91 We apply the UnivariateSpline function from the SciPy library. The parameter is chosen so that the highest frequencies are smoothed out on the main topics, whilst preserving the trend of weaker topics. The parameter must be adjusted depending on the number of topics, which was undertaken according to ‘taste’.

92 For instance, were there only two cases in the model, where Topic 1 constituted 20 per cent of the first and 40 per cent of the second, Topic 1 would then represent a contribution of 30 per cent of the overall subject matter of the modelled cases.

93 ‘minister’, ‘tribunal’, ‘australian’, ‘review’, ‘protection’, ‘conduct’, ‘submission’, ‘convention’.

*Status of Refugees*⁹⁴ is not, in and of itself, a threat to liberty within the meaning of section 91R(2)(a) of the *Migration Act 1958* (Cth).

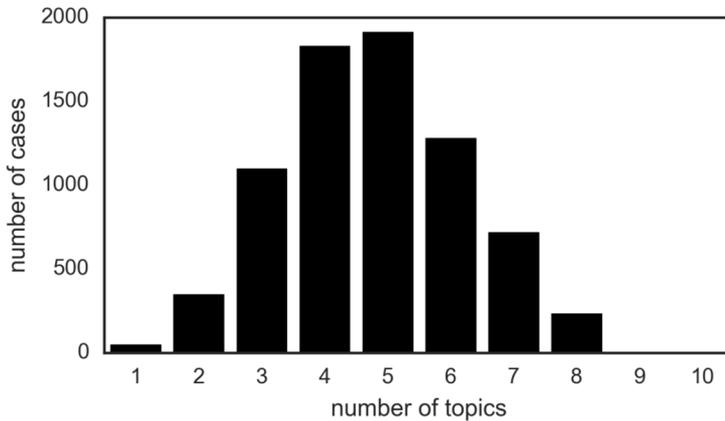
Table 2: 10 Topic Model – Representative Cases

Topic	Most Representative Case	Topic Label
0	<i>Minister for Immigration and Border Protection v WZAPN</i> (2015) 254 CLR 610	Government action in relation to immigration
1	<i>Marks v Trustees Executors & Agency Co Ltd</i> (1948) 77 CLR 497	Beneficiaries' rights to property in an estate
2	<i>Michaelides v The Queen</i> (2013) 296 ALR 1	Trial process for criminal matters
3	<i>Russell v Walters</i> (1957) 96 CLR 177	Trade licensing, regulation and IP
4	<i>Alldridge v Mulcahey</i> (1950) 81 CLR 337	Damage to persons through injury
5	<i>Uther v Federal Commissioner of Taxation</i> (1964) 111 CLR 318	Company financial flows
6	<i>R v Owens; Ex parte Seaton</i> (1933) 49 CLR 20	Jurisdictional divisions and actions
7	<i>Scott v Sun Alliance Australia Ltd</i> (1993) 178 CLR 1	Employment entitlements and disputes
8	<i>Everingham v Minister for Lands (NSW)</i> (1916) 21 CLR 269	Land contracts and agreements
9	<i>Polites v Commonwealth</i> (1945) 70 CLR 60	Constitutional actors and relationships

Despite the usefulness of 'representative' cases, individual cases are very rarely classified as belonging to a single topic. Rather, as shown in the figure below, cases are classified as 'belonging' to up to eight topics, with five the most common number of topics a single case is classified as representing.

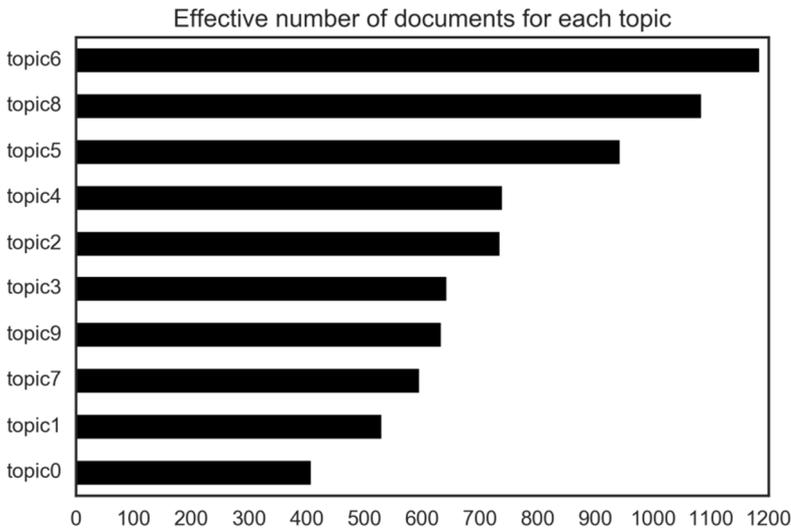
94 *Convention Relating to the Status of Refugees*, opened for signature 28 July 1951, 189 UNTS 137 (entered into force 22 April 1954).

Figure 5: Number of Topics per Case (Only Topics with a Minimum Weight of One Per Cent Are Taken into Account).



Yet reviewing how the entire corpus is classified according to a case’s dominant topic begins to explain how the overall judicial workload of the Court is constructed. The figure below presents the effective number of cases assigned to each topic in the model.

Figure 6: Effective Number of Documents for Each Topic



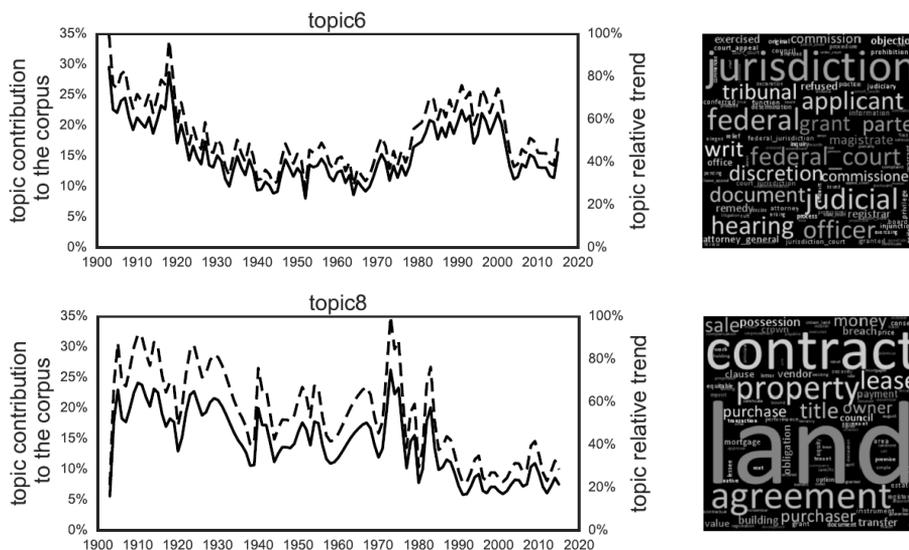
It is clear that Topic 6 (jurisdictional divisions and actions)⁹⁵ holds the greatest weight across the corpus. In other words, it dominates the legal subject

95 'jurisdiction', 'federal', 'judicial', 'applicant', 'hearing', 'officer', 'federal_court', 'document'.

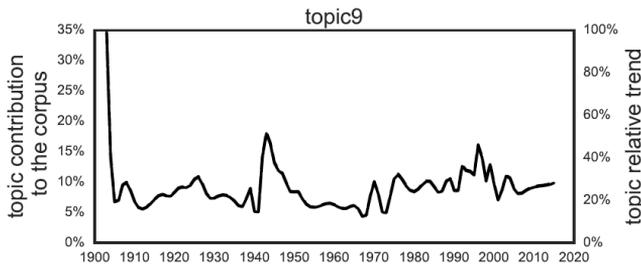
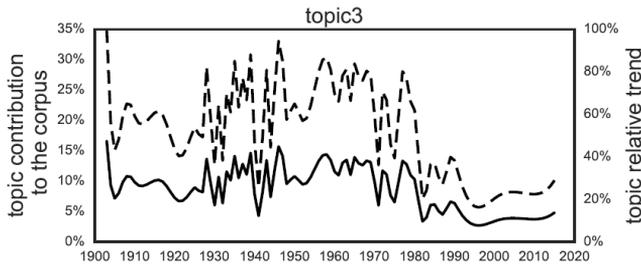
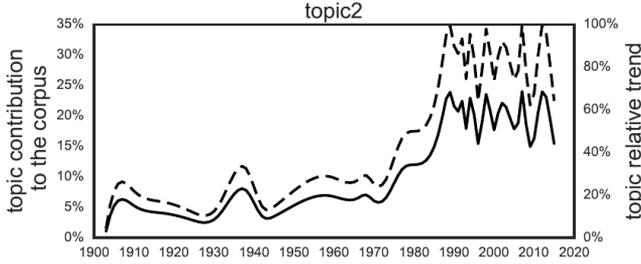
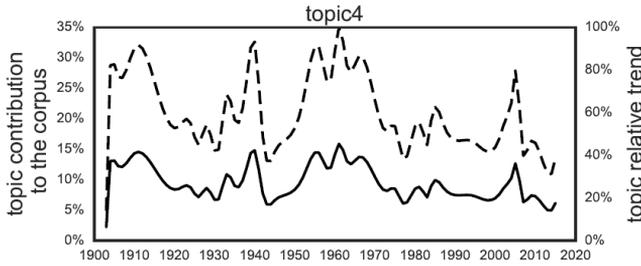
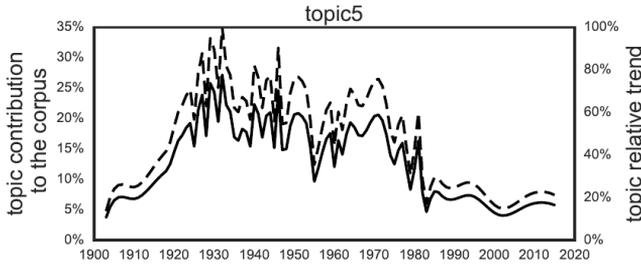
matter of the Court from 1903 to 2015. This is followed relatively closely by Topic 8 (land contracts and agreements) and Topic 5 (company financial flows). The dominance of a topic that relates to questions of judicial power and process is not surprising. The dominance of the real property discourse is perhaps more surprising, as it is not immediately clear why such a topic would so dominate the subject matter of the Court's judicial workload.

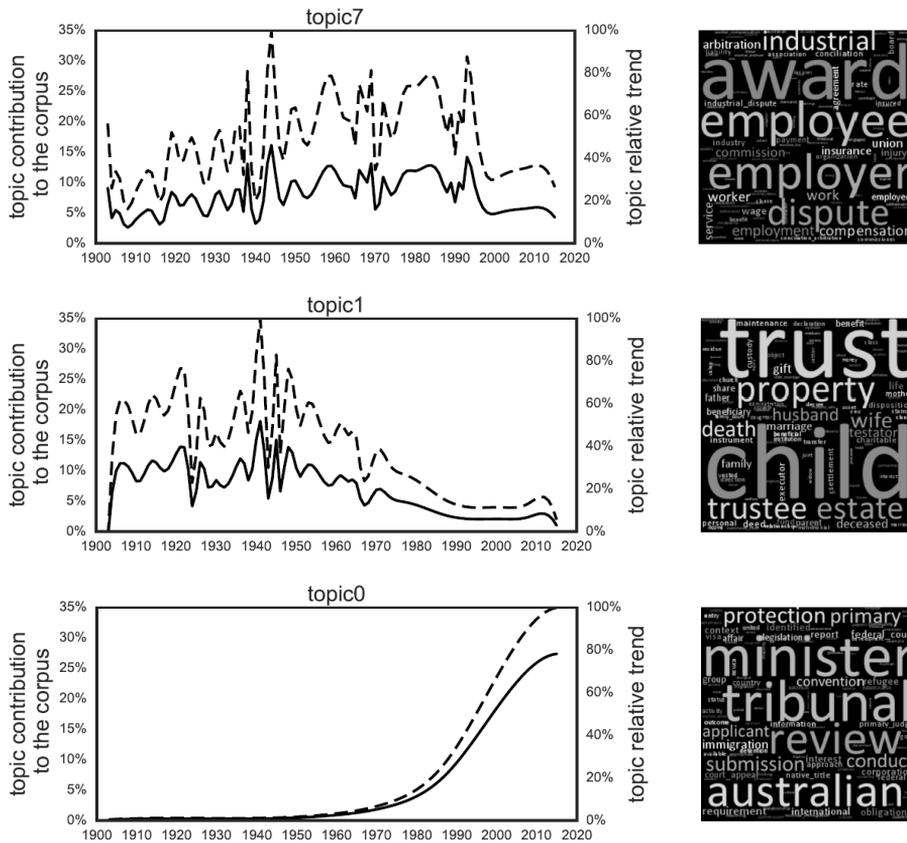
Understanding the dominance of topics is assisted by reviewing the *shifting* make-up of the Court's legal subject. Reviewing the figures below, there have been downward trends for Topic 3 (trade licencing, regulation and IP), Topic 5 (company financial flows) and Topic 8 (land contracts and agreements), whilst Topic 2 (trial process for criminal matters) and Topic 0 (government action in relation to immigration) have seen a marked rise in the dominance of that topic as a part of the Court's overall legal subject matter. What has remained relatively static, although with some volatility, are Topic 6 (jurisdictional divisions and actions), Topic 9 (constitutional actors and relationships) and Topic 4 (damage to persons through injury). These shifts throughout time are illustrated in the following figure, which presents topics in descending order of presence for the period 1903–2015:⁹⁶

Figure 7: 10 Topic Model: Timelines and Tokens Order of Dominance 1903–2015; solid line indicates a topic's contribution to the corpus in that year, dotted line is the topic's relative contribution normalised to the highest single year contribution



96 The word clouds visible at the right of each chart display the tokens for each topic, with the font size of each token driven by the weight of that word or token for that topic. The solid line series charts the topic's contribution to the corpus by year, with the dashed line series the topic relative trend.





With this data, we can begin to further analyse the Court's varying level of published output throughout the 20th century.

In discussion earlier in this article we identified the period 1939–1945 during the Latham Court as a period marked lower annual aggregate number of judgments handed down by the Court. We hypothesised that this was due to the Second World War indirectly influencing the number of cases the Court was asked to adjudicate or allowed itself to adjudicate.⁹⁷ Here we apply the topic model account of the subject matter the Court was adjudicating as a helpful data point in aid of that analysis. In short, this data shows that the period was defined by significant *polarisation* in the Court's subject matter. During no other period has the Court demonstrated a more varied attention to different subject matter areas/topics, with some topics contributing the least to the Court's overall workload in the 113 years of our data, contrasted with others contributing the most to the Court's overall workload in their history. More interestingly, others

97 The matter of case selection and the formation of the Court's own agenda is a fascinating topic for which work by Livermore and colleagues is continuing, see Livermore, Riddell and Rockmore, 'A Topic Model Approach', above n 17.

still contributed historically low levels and then later historically high levels, all within the same period of the Second World War. In order of descending presence of each topic in the Court's overall subject matter:

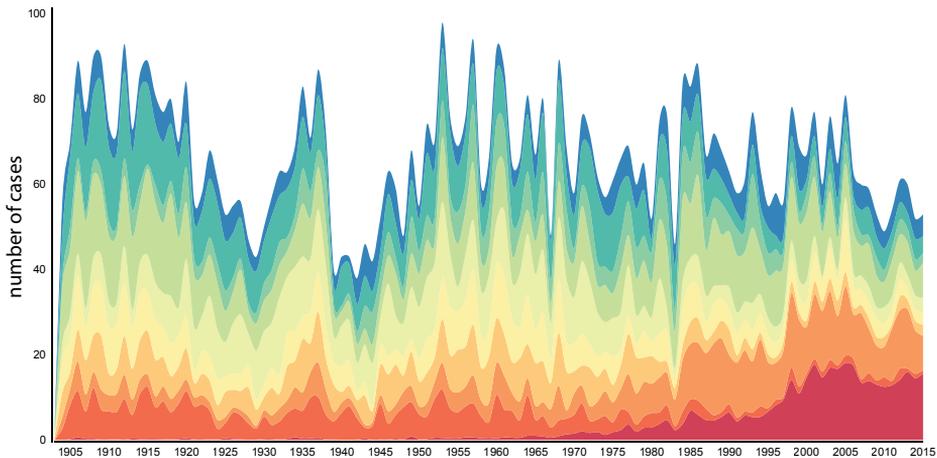
- Topic 6 (jurisdictional divisions and actions) accounts for less of the Court's attention than in any other period the Court's history;
- Topic 5 (company financial flows) dominates, at a level almost as high as during the Great Depression, and not seen since;
- Topic 4 (damage to persons through injury) contributes less to the corpus than in any other period;
- Topic 2 (trial process for criminal matters) falls away from a 1930s peak, punctuating a trend which continues from the 1930s onwards with a short-lived drop to levels almost the lowest in the history of the Court;
- Topic 3 (trade licencing, regulation and IP) experienced both its lowest and highest ever contributions to the annual workload of the Court, save for the recent dramatic fall from the 1980s onwards;
- Topic 9 (constitutional actors and relationships) also experienced a period of dramatic polarisation, reaching its equal-lowest contribution to the Court's workload in the early part of the decade, then peaking sharply in the mid-decade to reach its most dominant level since the outlier period of the first two years of the Court;
- Topic 7 (employment entitlements and disputes) continues this trend of dropping-off and return, with a sharp drop in the early years to its lowest contribution in the Court's history, followed by an equally dramatic surge peaking in 1944 with the greatest contribution from that topic in the lifetime of the Court;
- Topic 1 (beneficiaries' rights to property in an estate) shows clear polarisation, growing from the later years of the 1930s to make the greatest contribution in its history to the Court's subject matter in 1941, followed by a period of volatility in which it contributes the least in its history to that time and the lowest until the late 1960s, by which time the topic has notably fallen away in importance for the Court.

Never has the Court's judicial workload been less dominated by some topics and more dominated by others. This polarisation exists not only *between* topics but also *within* some topics. The pattern of polarisation requires further analysis. However, there is a distinctive, if temporary, decrease in matters related to the actions of natural persons and interpersonal/personal harms, whilst also a dropping away of appellate disputes. This is contrasted with a parallel initial sharp drop followed by a pronounced rise in matters related to constitutional (powers) and 'employment entitlements and disputes'. The theme of the state's relationship to natural persons could function as the heuristic device with which to read this pattern, with the shifting polarities pivoting about this division.

Such polarisation can also be seen at a whole-of-topic view. Note how in the 10 topic model those topics most directly related to the relationship or interface between natural persons and state power (Topics 2 and 0) see the most

pronounced rise in attention by the Court over time. Conversely, those topics which relate to private relationships between natural persons (Topic 1 and also perhaps Topic 8) and perhaps most notably, the regulation of commercial and trade relationships (Topics 8, 5 and 3) are those which have most reduced their dominance of the Court's subject matter. The use of a streamgraph (Figure 8), presents one method of visualising these relationships through time. Here, each topic of the 10 topic model is represented by a stacked single plot and colour, visualising the contribution of each topic compared with others to the corpus as a whole. The streamgraph allows us to visualise at once both the overall workload of the Court (the envelope of the plot), and the relative importance of the topics (the streams of colours) across time.

Figure 8: Streamgraph (10 Topic Model). Topics Ordered 0–9 When Reading from Bottom to Top of Figure.



This form of analysis is a unique contribution of a topic modelling approach. Similarly, it bears fruit in relation to the lower case numbers handed down during the 1920s by the Knox Court. Whilst the Court handed down fewer cases, the aggregate annual character length of those cases increased during that period. We find that the dominant topics during that period were Topic 5 (company financial flows) and Topic 6 (jurisdictional divisions and actions). Topic 5 increased steeply in weight as a proportion of the total topic load considered by the Court, peaking in the early 1920s. Topic 6, on the other hand, began a sharp downward trend through the decade. Attention to matters of real property showed little change. Tortious matters were stable, increasing in importance during the 1930s to reverse a slight downward trend in the 1920s (with a weight of eight per cent at the opening of the 1930s extending to approximately 15 per cent by the end of the decade). The decreasing dominance of jurisdictional/appellate questions during the period may be linked to the Knox Court's 'watershed' movement

away from previous forms of constitutional interpretation,⁹⁸ thus driving the progressive lowering of cases handed down during that period. On the other hand, the extended rise of Topic 5, with a peak in the early 1930s, presents an interesting spur to further historical and doctrinal analysis. It seems to show that there was a notable increase in judicial consideration of commercial and financial matters commencing prior to the generally accepted period of the Great Depression.

B Fifty Topic Model

In modelling a corpus of legal text, the selection of the topic number is a critical issue. Unlike text found in newspaper articles, or romance literature, legal corpora exist within a well-defined, pre-existing and widely accepted taxonomy which *itself* plays a significant role in the legal texts; the language and taxonomical categories are themselves present in those texts. For example, the taxonomical terminology of ‘administrative law’ or ‘administrative matters’ appears in the legal texts that form part of that area of law, whereas ‘epistolary novel’ or ‘epistolary form’ will likely not appear in the text of novels belonging to that genre.

The preferred approach is to fit a topic model based upon its usefulness and accuracy in relation to the underlying text. For example, in an innovative modelling of a collection of 4500 poems from the ekphrastic tradition⁹⁹ – poems written to, for, or about the visual arts – Lisa M Rhody modelled 60 topics.¹⁰⁰ These 60 topics represented the highly figurative subject matter of those texts in a way that captured the subject matter at a level of granularity felt to be potentially productive for scholarship on the genre.¹⁰¹ In contrast, Macey and Mitts modelled just three topics, which yielded support for their theory-building regarding corporate veil-piercing.¹⁰² We report on the 50 topic model here, highlighting the effect of modelling a much larger number of topics on the representation of the underlying textual material compared to the 10 topic model above.¹⁰³

98 See Twomey, above n 76.

99 Rhody cites ‘Ode on a Grecian Urn’ by John Keats, ‘My Last Duchess’ by Robert Browning and ‘For the Union Dead’ by Robert Lowell as examples from the canon: Lisa M Rhody, ‘Topic Modeling and Figurative Language’ (2012) 2(1) *Journal of Digital Humanities* <<http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/>>.

100 Ibid.

101 Ibid.

102 Macey and Mitts, above n 39, 150–1.

103 Whilst we have modelled a 100 topic model, the number of topics in that model is felt to be too great to facilitate straightforward human ‘reading’ of them. They will, however, be the subject of forthcoming work on the topic.

Table 3: 50 Topic Model

Topic	Words	Topic Labels
0	'judicial', 'function', 'judicial_power', 'criminal', 'attorney', 'attorney_general', 'executive', 'constitutional'	Judicial Function and Power – In Relation to Executive Government and Constitution
1	'medical', 'hospital', 'school', 'charitable', 'church', 'practitioner', 'treatment', 'institution'	Entities – Income Tax Exempt – Charitable
2	'finding', 'trial', 'court_appeal', 'primary', 'trial_judge', 'conduct', 'primary_judge', 'error'	Primary Trial Finding – Appeal – Judicial Conduct
3	'ship', 'owner', 'custom', 'port', 'vessel', 'master', 'shipping', 'carrier'	Shipping
4	'jury', 'trial', 'accused', 'verdict', 'trial_judge', 'witness', 'applicant', 'direction'	Jury Trial – Process
5	'offence', 'criminal', 'crime', 'prosecution', 'charge', 'accused', 'conviction', 'charged'	Criminal Charge
6	'damage', 'loss', 'care', 'liability', 'economic', 'tort', 'suffered', 'duty_care'	Damages – Tortious
7	'corporation', 'conduct', 'market', 'supply', 'club', 'trade', 'trading', 'competition'	Corporation, Conduct of
8	'child', 'wife', 'husband', 'marriage', 'family', 'parent', 'mother', 'father'	Family Relationships
9	'trust', 'trustee', 'deed', 'equity', 'fiduciary', 'beneficiary', 'joint', 'equitable'	Trust – Equity – Fiduciary and Beneficiary
10	'lease', 'premise', 'possession', 'rent', 'lessee', 'tenant', 'covenant', 'landlord'	Leases
11	'payment', 'rate', 'payable', 'pound', 'money', 'value', 'cent', 'charge'	Payments
12	'magistrate', 'special_leave', 'hearing', 'applicant', 'process', 'leave_appeal', 'jurisdiction', 'summons'	Leave to Appeal
13	'learned_judge', 'mere', 'bound', 'great', 'chief', 'english', 'house', 'chief_justice'	Precedent – Judicial Deference – Court Hierarchy
14	'trade', 'commerce', 'inter', 'inter_state', 'trade_commerce', 'sale', 'licence', 'excise'	Trade and Commerce – Interstate Trade
15	'election', 'constitution', 'political', 'house', 'representative', 'vote', 'freedom', 'electoral'	Elections
16	'compensation', 'injury', 'worker', 'employer', 'worker_compensation', 'accident', 'disease', 'employment'	Compensation – Worker/Industrial
17	'business', 'profit', 'sale', 'value', 'pound', 'partnership', 'stock', 'asset'	Business Transactions

Topic	Words	Topic Labels
18	'negligence', 'care', 'injury', 'risk', 'accident', 'reasonable', 'danger', 'caused'	Negligence – Assessment
19	'contract', 'agreement', 'purchaser', 'clause', 'vendor', 'obligation', 'sale', 'breach'	Contract Interpretation – Enforcement
20	'australian', 'international', 'country', 'convention', 'united', 'foreign', 'territory', 'resident'	Australia – Nation – Convention and Agreements – Foreign Affairs
21	'board', 'price', 'wheat', 'wool', 'coal', 'delivery', 'grower', 'export'	Export and Industrial Regulation And Controls
22	'tribunal', 'minister', 'immigration', 'visa', 'applicant', 'refugee', 'review', 'protection'	Administrative Decision-Making – Immigration
23	'regulation', 'legislature', 'prescribed', 'defence', 'parliament', 'interpretation', 'specified', 'amendment'	Regulation and Delegated Legislation – Interpretation and Powers Granted
24	'jurisdiction', 'federal', 'federal_court', 'judicial', 'constitution', 'judiciary', 'federal_jurisdiction', 'jurisdiction_court'	Jurisdiction – Judicial/Courts – Constitutional Definition – Judiciary
25	'company', 'share', 'director', 'shareholder', 'dividend', 'capital', 'article', 'meeting'	Company – Corporate Actors/Office Holders – Relationships and Conduct
26	'service', 'officer', 'crown', 'commissioner', 'office', 'government', 'department', 'appointment'	Office – Office Holder
27	'award', 'dispute', 'industrial', 'arbitration', 'union', 'employee', 'employer', 'conciliation'	Awards – Industrial Conflict
28	'damage', 'liability', 'cause_action', 'insurer', 'limitation', 'liable', 'indemnity', 'statement_claim'	Damages – Liability and Limitations
29	'debt', 'creditor', 'bankruptcy', 'payment', 'debtor', 'charge', 'bankrupt', 'company'	Debt and Credit – Bankruptcy
30	'licence', 'title', 'native', 'territory', 'native_title', 'aboriginal', 'northern', 'northern_territory'	Licence/Authorisation – Native Title
31	'police', 'publication', 'officer', 'defence', 'conduct', 'contempt', 'report', 'defamation'	Publication – Risks of/to Police Effectiveness – Contempt of Court and Defamation
32	'commission', 'applicant', 'discretion', 'grant', 'injunction', 'minister', 'inquiry', 'hearing'	Commissions – Government Action/Decision Making – Review
33	'constitution', 'parliament', 'government', 'federal', 'legislative', 'constitutional', 'legislation', 'territory'	Constitutional Powers – Federal – Parliamentary and Legislative Power
34	'bank', 'money', 'mortgage', 'security', 'transaction', 'loan', 'mortgagee', 'cheque'	Banking Technologies – Regulated Instruments

Topic	Words	Topic Labels
35	'mark', 'registration', 'trade', 'registered', 'trade_mark', 'murder', 'death', 'register'	Trade Marks
36	'estate', 'death', 'testator', 'deceased', 'gift', 'share', 'executor', 'life'	Estates – Inheritance
37	'vehicle', 'motor', 'road', 'motor_vehicle', 'transport', 'driver', 'driving', 'carriage'	Motor Vehicles
38	'letter', 'certificate', 'document', 'march', 'december', 'february', 'signed', 'april'	Execution of Documents – Date
39	'document', 'information', 'privilege', 'disclosure', 'warrant', 'advice', 'professional', 'client'	Legal Privilege in Documents, Advice and Information
40	'insurance', 'insured', 'society', 'business', 'incurred', 'company', 'expenditure', 'loss'	Insurance
41	'land', 'crown', 'title', 'owner', 'value', 'acquisition', 'crown_land', 'grant'	Crown Land – Grants and Acquisition
42	'council', 'water', 'building', 'road', 'local', 'area', 'development', 'street'	Local Government – Civil Works and Planning
43	'patent', 'invention', 'process', 'specification', 'product', 'mining', 'infringement', 'manufacture'	Patent – Subject Matter – Application – Infringement
44	'sentence', 'criminal', 'sentencing', 'offence', 'criminal_appeal', 'court_criminal', 'imprisonment', 'offender'	Criminal Sentence
45	'property', 'transfer', 'value', 'stamp', 'instrument', 'estate', 'stamp_duty', 'asset'	Duties and Taxes on Estates
46	'court_appeal', 'submission', 'legislation', 'australian', 'approach', 'context', 'joint', 'requirement'	Appeal – Submissions and Process
47	'income', 'commissioner', 'assessment', 'taxpayer', 'taxation', 'commissioner_taxation', 'federal', 'federal_commissioner'	Federal Taxation – Collection and Administration
48	'work', 'employee', 'employment', 'employer', 'contractor', 'employed', 'hour', 'working'	Employment – Status
49	'fund', 'benefit', 'scheme', 'money', 'contribution', 'appropriation', 'payment', 'superannuation'	Non-Corporate Legal Persons – Funds and Societies – Taxation

The results in this 50 topic model bring with them a noticeably greater granularity, yet retain a consistency with the 10 topic model. The 50 topic model includes some familiar examples: Topic 33¹⁰⁴ (constitutional powers – federal – parliamentary and legislative power) reflects the tokens used to describe Topic 9 in the 10 topic model.¹⁰⁵ Topic 33 in the 50 topic model is joined by further

104 'constitution', 'parliament', 'government', 'federal', 'legislative', 'constitutional', 'legislation', 'territory'.

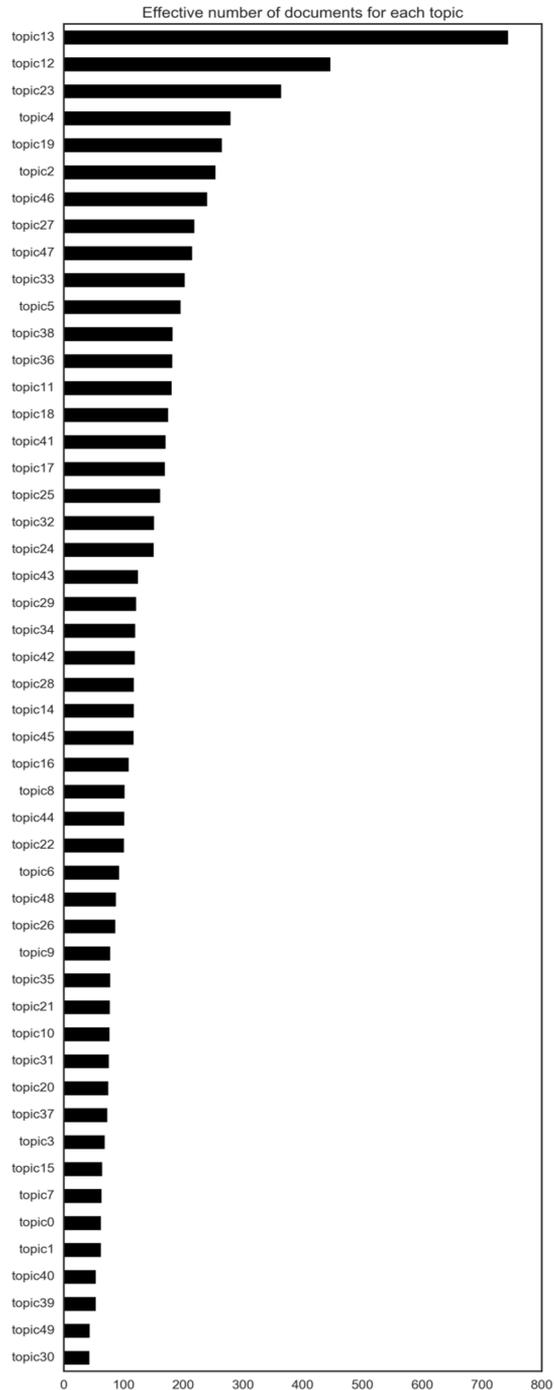
105 'constitution', 'parliament', 'government', 'regulation', 'territory', 'legislative', 'federal', 'constitutional'.

‘constitutional’ topics (Topics 0, 15 and 24), whereas Topic 9 in the 10 topic model is alone in that respect.

Reviewing the effective number of cases per topic,¹⁰⁶ Topic 13 (precedent – judicial deference – court hierarchy), Topic 12 (leave to appeal) and Topic 23 (regulation and delegated legislation – interpretation and powers granted) are the dominant topics of the Court’s case load. Topic 13 has an effective weight of about 750 cases (approximately 10 per cent of all cases), and Topic 12 of approximately 450 cases. Here, in line with the 10 topic model, a topic related to questions of judicial power and process dominates the content of the Court, whilst a topic in relation to payments and value seems to add a possibly more nuanced description of the Court’s second most dominant topic, when compared with the dominance of land purchase in the 10 topic model.

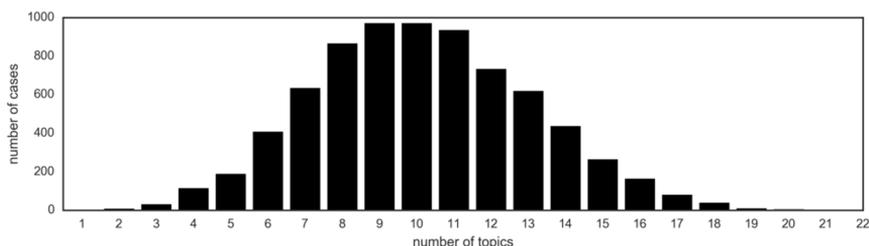
106 Here ‘effective’ indicates the effective number of cases. This means that if we add the weights (fraction of a document) associated with a particular topic across all cases in the corpus, the total weight of that topic is effectively the same size as x number of cases. For example, in relation to Topic 13, if we add all of the fractions of documents which are assigned to Topic 13, added together this equates to approximately 750 individual cases. This is not to say that Topic 13 is the *most important* or most dominant case in 750 cases, but rather that its presence equates to almost as much.

Figure 9: Effective Number of Documents for Each Topic



In this 50 topic model, the number of topics that classify a single case is markedly different from that of the 10 topic model. In the 10 topic model, five is the most common number of topics by which a case is classified (see Figure 5); in the 50 topic model, most cases are classified by or assigned 9 or 10 topics.

Figure 10: Number of Topics per Case (50 Topic Model)



This shift is driven by the greater granularity of topics in this model.

VI TESTING THE TOPIC MODEL

The following section complements the descriptive account provided above of the 10 and 50 topic models with a series of selected tests focused on the validity, predictive ability and utility of the 50 topic model for legal scholarship.

A Single Cases: The *Mabo* Cases

The 50 topic model classifies *Mabo v Queensland [No 1]*¹⁰⁷ according to the following five leading topics:¹⁰⁸

Table 4: Leading Topics for *Mabo v Queensland*: 50 Topic Model

Topic	Words/Tokens	Coverage
30	'licence', 'title', 'native', 'territory', 'native_title', 'aboriginal', 'northern', 'northern_territory'	28%
20	'australian', 'international', 'country', 'convention', 'united', 'foreign', 'territory', 'resident'	27%
23	'regulation', 'legislature', 'prescribed', 'defence', 'parliament', 'interpretation', 'specified', 'amendment'	13%
33	'constitution', 'parliament', 'government', 'federal', 'legislative', 'constitutional', 'legislation', 'territory'	12%
41	'land', 'crown', 'title', 'owner', 'value', 'acquisition', 'crown_land', 'grant'	11%

¹⁰⁷ *Mabo v Queensland* (1988) 166 CLR 186 ('*Mabo v Queensland [No 1]*').

¹⁰⁸ This case was selected as the earlier of the *Mabo* cases.

These five topics represent approximately 90 per cent of the case's content according to the topic model. Applying researcher-generated labels in the form of a headnote, they would read:

Licence/Authorisation – Native Title | Australia – Nation – Convention and Agreements – Foreign Affairs | Regulation and Delegated Legislation – Interpretation and Powers Granted | Constitutional Powers – Federal – Parliamentary and Legislative Power | Crown Land – Grants and Acquisition

The concatenation of researcher-generated labels provides an accurate description of the legal subject matter of the case. The level of detail could, however, be improved. This indicates that the construction of the researcher-generated labels themselves could be refined. For example, the use of the concept of 'grants and acquisition' or 'licence' is not entirely helpful when used to describe the question of the Crown's relationship to land in this context. Human-generated editorial catchwords provide a potentially more accurate – or at least more detailed – description of the content of a case such as *Mabo v Queensland [No 1]*. To illustrate, the catchwords used in the CLR to describe the case are as follows:

Constitutional Law (Cth) – Inconsistency between Commonwealth and State laws – Commonwealth law prohibiting racial discrimination – Prohibition of acts involving a distinction based on race – Enjoyment of right by person of particular race not enjoyed by persons of another race – Ownership of property – Islands off coast of Queensland annexed to colony in 1879 – State Act declaring islands upon annexation to have become waste lands of Crown – No compensation payable – Effect of State Act – Whether inconsistent with Commonwealth law – The Constitution (63 and 64 Vic c 12), s 109 – *Racial Discrimination Act 1975* (Cth), ss 9, 10 – *Queensland Coast Islands Declaratory Act 1985* (Q) ss 3, 4, 5.

Constitutional Law (Q) – Powers of State Parliament – Waste lands – Management and control vested in legislature – Powers – Whether subject to limitation – Power to acquire property without compensation – Islands off coast of Queensland annexed to colony in 1879 – Action by persons claiming traditional ownership of islands – Subsequent Act declaring islands upon annexation to have become waste lands of Crown – No compensation payable – Effect of deeming declaration – Whether Act interferes with judicial process – *Constitution Act 1867* (Q), s 40 – *Queensland Coast Islands Declaratory Act 1985* (Q), ss 3, 4, 5.¹⁰⁹

There is a clear difference in the level of detail expressed in the human generated version and the concatenated topic model of the same.

This difference may relate to the granularity and specificity of the topics themselves – driven largely by the number of topics the modeller requests be generated. In order to test this hypothesis, we apply results from the 100 topic model.¹¹⁰ Whilst we do not report on the entirety of the 100 topic model (the subject of forthcoming work) in this article, we include here the leading five topics with which that model classifies *Mabo v Queensland [No 1]* in order to test the hypothesis:

109 *Mabo v Queensland [No 1]* (1988) 166 CLR 186.

110 We alter the format of these tokens in the 100 topic model in order to visually distinguish them from those present in the 10 and 50 topic models.

Table 5: Leading Topics for *Mabo v Queensland [No 1]*: 100 Topic Model

Topic	Words/Tokens	Coverage (%)
96	political native government freedom people	25
41	land title crown mining grant	23
29	territory water northern area aboriginal	8
58	council queensland committee privy privy_council	6
89	paragraph clause specified requirement description	5

Expressed in the form of catchwords in a headnote:

political|native|government|freedom|people--
land|title|crown|mining|grant--
territory|water|northern|area|aboriginal--
council|queensland|committee|privy|privy_council--
paragraph|clause|specified|requirement|description

Topics 41 and 29 speak to ‘land’ and ‘territory’ respectively. Whilst both may be expressed by a single concept – namely, the legal practices associated with areas of the earth’s surface – the collection of tokens/words demonstrate how for the Court they represent quite distinct topics. Here, the topic model has described and differentiated between ‘land’, which it recognises as related to words (ie, concepts and practices) such as ‘title’, ‘crown’, ‘mining’ and ‘grant’, and ‘territory’, which it understands in relation to ‘water’, ‘northern’, ‘area’ and ‘aboriginal’. The implication is that the Court itself understands and utilises two distinct topics in the text of their judgments, marking out two distinct practices. One topic *classifies, controls* and *uses* ‘land’ by legal technologies of ownership (‘title’, ‘crown’), with control/exclusion of others (‘title’, ‘grant’) for specific purposes such as ‘mining’ or ‘grant[ing]’. The other, ‘territory’, deals largely in *spatialised terms* lacking verbs or activity, with areas inert and not characterised by use or control. This second term is associated with one natural or legal subject, ‘Aboriginal’, whilst the first, ‘land’, is associated with the ‘crown’ and, by extension, those who are engaged in the activity of ‘grant[ing]’ or ‘mining’ and who possess or claim ‘title’. Both such topics have distinct characteristics, which together make possible a variety of conceptual frames and practices. There is a distinctive treatment in the underlying text of a legal concept and practice of ‘land’ and ‘territory’. It seems a larger number of topics generates a more nuanced representation of the underlying material.

In relation to classification of individual cases, the results demonstrate that the 50 topic model provides accurate identification of similar cases within the limitations of an automated process. For example, the 50 topic model identifies that *Mabo v Queensland [No 1]* is most similar to the following High Court cases:

Table 6: *Mabo v Queensland [No 1]*: Most Similar HCA Cases

Case	Similarity (%)
<i>Mabo v Queensland [No 2]</i> (1992) 175 CLR 1	87
<i>Commonwealth v Yarmirr</i> (2001) 208 CLR 1	86
<i>Western Australia v Commonwealth</i> (1995) 183 CLR 373	86
<i>Coe v Commonwealth</i> (1979) 24 ALR 118	82
<i>Wacando v Commonwealth</i> (1981) 148 CLR 1	81
<i>Gerhardy v Brown</i> (1985) 159 CLR 70	79
<i>Commonwealth v WMC Resources Ltd</i> (1998) 194 CLR 1	79
<i>A Raptis & Son v South Australia</i> (1977) 138 CLR 346	78
<i>Pearce v Florenca</i> (1976) 135 CLR 507	77
<i>Bonser v La Macchia</i> (1969) 122 CLR 177	77

Most pleasing is its clear relationship to *Mabo v Queensland [No 2]*.¹¹¹ The common character of *Mabo v Queensland [No 1]* and the first five cases is clearly bound up in their common tracing of the establishment of native title as we understand it today, including *Coe v Commonwealth*.¹¹² Those cases which follow seem to pivot away from direct and specific questions of native title, such as the case of *Gerhardy v Brown*¹¹³ and its question of the validity of the *Pitjantjatjara Land Rights Act 1981* (SA) in relation to the Commonwealth racial discrimination law. In that case, there was a question of the granting of land by South Australia to a particular Aboriginal nation, and so the question was not one of native title but, rather, the status and ownership of land and Aboriginal persons. The link to *Mabo v Queensland [No 1]* is clear.

The remaining cases are each linked by their subject matter pertaining to the dealing with lands by the Crown, be it through acquisition or other acts. Thus, the topic model is here able to identify conceptually, factually and legally/doctrinally related cases – identifying in rough form the level of such relatedness. The results are not simply ‘accurate’ on these grounds: the model actually seems to report a pattern of relationships not immediately obvious to a human classifier. A less advanced way of classifying might be to group only native title or land use/status claims made in relation to Aboriginal use, or all Commonwealth land acquisition claims and uses of land-related powers. Here, however, the cases represent a more specific issue, the Commonwealth’s claim of jurisdiction over lands or territory with contested or overlapping jurisdictional status. That is to say, this is not a question of the Commonwealth’s use or acquisition of land being contested generally – as a conflict about ‘just terms’ might be – but instead something more

111 (1992) 175 CLR 1.

112 (1979) 24 ALR 118.

113 (1985) 159 CLR 70.

nanced. The ability of the model to construct and produce this material points to a value beyond ‘accuracy’. It seems from this reading that there is significant opportunity for analysis of this output and its application, alongside other methods, to important questions about the Australian legal system and this particular part of it.

B Constitutional Matters

Finally, we apply our model to the task of identifying constitutional matters. Lynch and Williams’ statistical account of the Court explicitly focuses on constitutional matters, as befits both the Court’s original jurisdiction in this area and their subject matter expertise. From a definitional standpoint, Lynch and Williams follow Stephen Gageler SC,¹¹⁴ now Gageler J of the High Court, in his inaugural survey of the Court’s constitutional decisions delivered in 2001.¹¹⁵

To test the classificatory process of the model, we utilised the list of cases identified by Lynch and Williams in their recent review of High Court activity (2014) and presented the classification as made by the 10 and 50 topic models.¹¹⁶ Lynch notes that classification works within a context where other legal questions have varying dominance;¹¹⁷ this is expressed in our model by the mix of topics the model fits to the cases.

The 50 topic model provides the following classification of those cases:

Attorney-General (NT) v Emmerson (2014) 307 ALR 174

49%	Topic 0:judicial function judicial_power criminal attorney
12%	Topic 33:constitution parliament government federal legislative
10%	Topic 5:offence criminal crime prosecution charge
8%	Topic 46:court_appeal submission legislation australian approach
7%	Topic 45:property transfer value stamp instrument

Plaintiff S156-2013 v Minister for Immigration and Border Protection (2014) 254 CLR 28

57%	Topic 22:tribunal minister immigration visa applicant
12%	Topic 33:constitution parliament government federal legislative
9%	Topic 23:regulation legislature prescribed defence parliament
6%	Topic 20:australian international country convention united
5%	Topic 24:jurisdiction federal federal_court judicial constitution

114 Stephen Gageler, ‘The High Court on Constitutional Law: The 2001 Term’ (2002) 25 *University of New South Wales Law Journal* 194, 195. To this, Lynch and Williams have included any matters which involve questions of purely state or territory constitutional law: Lynch and Williams, above n 5, 1081.

115 Lynch and Williams, above n 5, 1081. See also Andrew Lynch, ‘Does the High Court Disagree More Often in Constitutional Cases? A Statistical Study of Judgment Delivery 1981–2003’ (2005) 33 *Federal Law Review* 485, 490.

116 Lynch and Williams, above n 5.

117 Andrew Lynch, ‘Dissent: Towards a Methodology for Measuring Judicial Disagreement in the High Court of Australia’ (2002) 24 *Sydney Law Review* 470, 491.

Williams v Commonwealth of Australia (2014) 252 CLR 416 ('*Williams [No 2]*')

37%	Topic 33:constitution parliament government federal legislative
16%	Topic 46:court_appeal submission legislation australian approach
11%	Topic 49:fund benefit scheme money contribution
8%	Topic 1:medical hospital school charitable church
8%	Topic 11:payment rate payable pound money
7%	Topic 0:judicial function judicial_power criminal attorney

Pollentine v Bleijie (2014) 253 CLR 629

36%	Topic 0:judicial function judicial_power criminal attorney
26%	Topic 44:sentence criminal sentencing offence criminal_appeal
8%	Topic 23:regulation legislature prescribed defence parliament
7%	Topic 1:medical hospital school charitable church
5%	Topic 46:court_appeal submission legislation australian approach

Tajjour v New South Wales (2014) 254 CLR 508

38%	Topic 15:election constitution political house representative
19%	Topic 0:judicial function judicial_power criminal attorney
12%	Topic 5:offence criminal crime prosecution charge
8%	Topic 33:constitution parliament government federal legislative
7%	Topic 46:court_appeal submission legislation australian approach

Kuczborski v Queensland (2014) 254 CLR 51

60%	Topic 0:judicial function judicial_power criminal attorney
20%	Topic 5:offence criminal crime prosecution charge
6%	Topic 23:regulation legislature prescribed defence parliament

To test the process 'in reverse', Topic 0, which captures the definition of judicial power and the separation of powers doctrine, is the primary classification for the following 2014 cases:

Table 7: Cases with Topic 0 as Primary Classification (2014)

Topic 0: 'judicial', 'function', 'judicial_power', 'criminal', 'attorney', 'attorney_general', 'executive', 'constitutional': 2014 Cases	Classification (%)
<i>Kuczborski v Queensland</i> (2014) 254 CLR 51	60
<i>Attorney-General (NT) v Emmerson</i> (2014) 307 ALR 174	49
<i>Pollentine v Bleijie</i> (2014) 253 CLR 629	36
<i>Tajjour v New South Wales</i> (2014) 254 CLR 508	19
<i>Henderson v Queensland</i> (2014) 255 CLR 1	18

VII CONCLUSION

The aim of this article was to present the results of a topic model fitted to an Australian legal corpus. Analysis of the results produces two principal outcomes. The first is a new and unique view of the judicial activity and legal subject matter before the High Court throughout its history. The second is a view through its texts into the ways in which the Court constructs and uses topics; that is, how it forms and uses legal concepts and practices.

The model's 'thematic' representation of the subject matter of the Court's judicial workload outside of pre-given categories is that which presents significant opportunities for scholarship of a very wide variety. The topics formed by the topic model process are themselves thought provoking. They form (in a sense we are developing in forthcoming work), a 'taxonomy of practice',¹¹⁸ contrasting with the taxonomies – formal, procedural, function- or reason-based – commonly used to describe and classify legal subject matter and materials.¹¹⁹ The 10 topic model may, at first glance, seem to misrepresent or represent in a less satisfactory way the totality of the legal subject matter of the Court. For example, whilst it is true that immigration matters fall largely within the traditional subject matter of 'administrative law', the representation of Topic 0 – by tokens reflecting review of ministerial and executive decision-making and applicability of international agreements in relation to immigration – seems to miss other possibilities contained within 'administrative law', and indeed seems to borrow from taxonomical categories such as 'international law'.¹²⁰ However, as a taxonomy of practice, those same words or tokens reveal that *for the High Court*, the area of 'administrative law', or control of government action, is indelibly related to immigration matters. The office of the minister, the practices of ministerial decision-making and tribunal process are understood in the textual corpus to be associated with a single dominant concept: 'immigration'. In the taxonomy of practice, it is this, rather than the theoretical (and traditional) understandings of what constitutes 'administrative law' that actually *is* what 'administrative law' is for the High Court. Put more strongly, from the perspective of a 10 topic model, 'administrative law' does not 'exist' for the Court: what does exist is a practice of writing about 'government action in relation to immigration' that includes (and not as borrowed from a related taxonomical field) questions of government powers, administrative decision-makers and process, and international agreements. As such, probabilistically, for the Court 'administrative law' in fact *is* judicial consideration of ministerial and tribunal decision-making related to immigration.

Part of the problem, and promise, of traditional taxon is their artificiality. We believe the topic modelling approach reduces the force of this critique: the taxonomical structure, and the taxa themselves, are constructed from an 'immanent plane', that is, directly from the underlying textual material itself. For

118 With thanks to Rachel Young for discussion of this aspect in particular.

119 See Sherwin, above n 54.

120 'minister', 'tribunal', 'australian', 'review', 'protection', 'conduct', 'submission', 'convention'.

this reason, the models represent the actual textual material of the Court over this period of time, providing both a predictive model of future text and a latent topic representation of the corpus.¹²¹ Rachel Buurma puts this point well, arguing that topic modelling produces a ‘counter-factual and retrospective reconstruction’¹²² of the subject matter out of which a text corpus has been developed. For her, this creates the potential for a ‘denaturalizing and unfamiliar (though crucially not “objective” or “unbiased”) view’¹²³ of the corpus itself and, in our case, the judicial workload and ‘law’ of the High Court of Australia. The process does so by offering an opportunity for this interchange between reading scale and text selection.

Beyond this question of the appropriate ‘depth’ for legal scholarship, what is at stake in the reading of legal text remains the same. Mireille Hildebrandt writes that ‘positive law, inscribed in legal texts, entails an authority not inherent in literary texts, generating legal consequences that can have real effects on a person’s life and liberty’.¹²⁴ We agree that the interpretation of legal text is an undeniably normative undertaking¹²⁵ that resists the mechanical application of rules, requiring expert human interpretation.¹²⁶ But we resist the straightforwardness of her assertion that the authority found in legal text ‘is not inherent’ in their literary counterparts. Rather, as described above, we prefer to think of that authority as being perhaps *less immediate* in those texts, as regards the consequences for the human person that is more immediately at stake in the reading of legal texts.¹²⁷

What is truly at stake in reading legal text should continue to animate debate about the appropriate scale and depth at which to ‘read’, in addition to the truly exciting possibilities the bringing together of digital text and computing power represents for scholarly practices. This is not lost upon Stephen Ramsay, one of the authors of the *Criminal Intent* project, who referred to his team’s work in purposive terms:

The Old Bailey ... has eight million stories. Accessing those stories involves understanding trial length, numbers of instances of poisoning, and rates of bigamy. ... But being stories, they find their more salient expression in the weightier motifs of the human condition: justice, revenge, dishonour, loss, trial. This is what

121 Chang et al, above n 71.

122 Rachel Sagner Buurma, ‘The Fictionality of Topic Modeling: Machine Reading Anthony Trollope’s *Barsetshire Series*’ (2015) 2(2) *Big Data & Society* <<http://bds.sagepub.com/content/spbds/2/2/2053951715610591.full.pdf>>.

123 Ibid.

124 Mireille Hildebrandt, ‘The Meaning and Mining of Legal Texts’ in David M Berry (ed), *Understanding Digital Humanities* (Palgrave Macmillan, 2012) 145, 145.

125 Ibid.

126 On this, Hildebrandt’s discussion of ‘the meaning of law, highlighting its embodiments in the technologies of the script, and the hermeneutic implications this has for legal expertise and for legal certainty’ is essential reading: *ibid* 146.

127 See Robin L West, ‘Adjudication Is Not Interpretation: Some Reservations about the Law-as-Literature Movement’ (1987) 54 *Tennessee Law Review* 203.

the humanities are about. This is the only reason for an historian to fire up Mathematica or for a student trained in French literature to get into Java.¹²⁸

The advent of more options for the study of legal text raises questions for the use and analysis of Australian legal corpora. There has been very little digital humanities scholarship referencing Australian legal materials, and none that utilises topic modelling as we do here. Based on the discussion above, we offer this first account of reading the High Court at a distance as a way to continue the conversation about these important questions.

128 Stephen Ramsay quoted in Jennifer Howard, 'Digging into Data, Day 2: Making Tools and Using Them' on Chronicle of Higher Education, *Wired Campus* (12 June 2011) <<http://chronicle.com/blogs/wiredcampus/digging-into-data-day-2-making-tools-and-using-them/31704>>. See also *With Criminal Intent* (7 August 2012) <<http://criminalintent.org/>>; Dan Cohen et al, 'Data Mining with Criminal Intent: Final White Paper' (White Paper, 31 August 2011) <<http://criminalintent.org/wp-content/uploads/2011/09/Data-Mining-with-Criminal-Intent-Final1.pdf>>.

VIII APPENDICES

A Fifty Topic Model Representative Cases

Topic	Words	Most Representative Case	Score (%)
0	'judicial', 'function', 'judicial_power', 'criminal', 'attorney', 'attorney_general', 'executive', 'constitutional'	<i>South Australia v Totani</i> (2010) 242 CLR 1	69.65
1	'medical', 'hospital', 'school', 'charitable', 'church', 'practitioner', 'treatment', 'institution'	<i>Royal Australasian College of Surgeons v Federal Commissioner of Taxation</i> (1943) 68 CLR 436	58.01
2	'finding', 'trial', 'court_appeal', 'primary', 'trial_judge', 'conduct', 'primary_judge', 'error'	<i>Louth v Diprose</i> (1992) 175 CLR 621	54.07
3	'ship', 'owner', 'custom', 'port', 'vessel', 'master', 'shipping', 'carrier'	<i>Commonwealth v Huon Channel & Peninsula Steamship Co Ltd</i> (1918) 24 CLR 385	55.75
4	'jury', 'trial', 'accused', 'verdict', 'trial_judge', 'witness', 'applicant', 'direction'	<i>Bulejck v The Queen</i> (1996) 185 CLR 375	82.76
5	'offence', 'criminal', 'crime', 'prosecution', 'charge', 'accused', 'conviction', 'charged'	<i>Tabé v The Queen</i> (2005) 225 CLR 418	72.47
6	'damage', 'loss', 'care', 'liability', 'economic', 'tort', 'suffered', 'duty_care'	<i>Bryan v Maloney</i> (1995) 182 CLR 609	71.05
7	'corporation', 'conduct', 'market', 'supply', 'club', 'trade', 'trading', 'competition'	<i>Boral Besser Masonry Ltd v Australian Competition and Consumer Commission</i> (2003) 215 CLR 374	68.19
8	'child', 'wife', 'husband', 'marriage', 'family', 'parent', 'mother', 'father'	<i>R v Cook; Ex parte C</i> (1985) 156 CLR 249	73.85
9	'child', 'wife', 'husband', 'marriage', 'family', 'parent', 'mother', 'father'	<i>Chief Commissioner of Stamp Duties (NSW) v Buckle</i> (1998) 192 CLR 226	44.15
10	'trust', 'trustee', 'deed', 'equity', 'fiduciary', 'beneficiary', 'joint', 'equitable'	<i>Hall v Hoyts Theatres Ltd</i> (1934) 51 CLR 387	52.83
11	'lease', 'premise', 'possession', 'rent', 'lessee', 'tenant', 'covenant', 'landlord'	<i>Australian Broadcasting Commission v Australasian Performing Right Association Ltd</i> (1973) 129 CLR 99	60.84
12	'payment', 'rate', 'payable', 'pound', 'money', 'value', 'cent', 'charge'	<i>Mathews v Burns</i> (1918) 25 CLR 322	80.40
13	'magistrate', 'special_leave', 'hearing', 'applicant', 'process', 'eave_appeal', 'jurisdiction', 'summons'	<i>Northway v Coulthard</i> (1913) 16 CLR 352	76.29

Topic	Words	Most Representative Case	Score (%)
14	'learned_judge', 'mere', 'bound', 'great', 'chief', 'english', 'house', 'chief_justice'	<i>Barter's Farms Pty Ltd v Todd</i> (1978) 139 CLR 499	83.31
15	'election', 'constitution', 'political', 'house', 'representative', 'vote', 'freedom', 'electoral'	<i>Langer v Commonwealth</i> (1996) 186 CLR 302	83.4
16	'compensation', 'injury', 'worker', 'employer', 'worker_compensation', 'accident', 'disease', 'employment'	<i>Fraher v Wunderlich Ltd</i> (1963) 110 CLR 466	82.25
17	'business', 'profit', 'sale', 'value', 'pound', 'partnership', 'stock', 'asset'	<i>Federal Commissioner of Taxation v Ryan</i> (1926) 38 CLR 472	60.27
18	'negligence', 'care', 'injury', 'risk', 'accident', 'reasonable', 'danger', 'caused'	<i>Public Transport Commission (NSW) v Perry</i> (1977) 137 CLR 107	70.22
19	'contract', 'agreement', 'purchaser', 'clause', 'vendor', 'obligation', 'sale', 'breach'	<i>Perri v Coolangatta Investments Pty Ltd</i> (1982) 149 CLR 537	78.32
20	'australian', 'international', 'country', 'convention', 'united', 'foreign', 'territory', 'resident'	<i>Queensland v Commonwealth</i> (1989) 167 CLR 232	62.71
21	'board', 'price', 'wheat', 'wool', 'coal', 'delivery', 'grower', 'export'	<i>McClintock v Commonwealth</i> (1947) 75 CLR 1	57.20
22	'tribunal', 'minister', 'immigration', 'visa', 'applicant', 'refugee', 'review', 'protection'	<i>Minister for Immigration and Border Protection v SZSCA</i> (2014) 254 CLR 317	82.30
23	'regulation', 'legislature', 'prescribed', 'defence', 'parliament', 'interpretation', 'specified', 'amendment'	<i>Bird v John Sharp & Sons Pty Ltd</i> (1942) 66 CLR 233	66.24
24	'jurisdiction', 'federal', 'federal_court', 'judicial', 'constitution', 'judiciary', 'federal_jurisdiction', 'jurisdiction_court'	<i>Re Jarman; Ex parte Cook [No 1]</i> (1997) 188 CLR 595	63.05
25	'company', 'share', 'director', 'shareholder', 'dividend', 'capital', 'article', 'meeting'	<i>Beck v Weinstock</i> (2013) 251 CLR 425	71.15
26	'service', 'officer', 'crown', 'commissioner', 'office', 'government', 'department', 'appointment'	<i>Schedlich v Commonwealth</i> (1926) 38 CLR 518	60.74
27	'award', 'dispute', 'industrial', 'arbitration', 'union', 'employee', 'employer', 'conciliation'	<i>R v Graziers' Association of NSW; Ex parte Australian Workers Union</i> (1956) 96 CLR 317	90.22
28	'damage', 'liability', 'cause_action', 'insurer', 'limitation', 'liable', 'indemnity', 'statement_claim'	<i>Brambles Constructions Pty Ltd v Helmers</i> (1966) 114 CLR 213	64.25

Topic	Words	Most Representative Case	Score (%)
29	'debt', 'creditor', 'bankruptcy', 'payment', 'debtor', 'charge', 'bankrupt', 'company'	<i>Rae v Samuel Taylor Pty Ltd</i> (1963) 110 CLR 517	76.88
30	'licence', 'title', 'native', 'territory', 'native_title', 'aboriginal', 'northern', 'northern_territory'	<i>Akiba v Commonwealth</i> (2013) 250 CLR 209	75.53
31	'police', 'publication', 'officer', 'defence', 'conduct', 'contempt', 'report', 'defamation'	<i>Pervan v North Queensland Newspaper Co Ltd</i> (1993) 178 CLR 309	63.16
32	'commission', 'applicant', 'discretion', 'grant', 'injunction', 'minister', 'inquiry', 'hearing'	<i>Ainsworth v Criminal Justice Commission</i> (1992) 175 CLR 564	63.76
33	'constitution', 'parliament', 'government', 'federal', 'legislative', 'constitutional', 'legislation', 'territory'	<i>Svikart v Stewart</i> (1994) 181 CLR 548	79.71
34	'bank', 'money', 'mortgage', 'security', 'transaction', 'loan', 'mortgagee', 'cheque'	<i>Bank of NSW v Permanent Trustee Company of NSW Ltd</i> (1943) 68 CLR 1	56.94
35	'mark', 'registration', 'trade', 'registered', 'trade_mark', 'murder', 'death', 'register'	<i>Burger King Corporation v Registrar of Trade Marks</i> (1973) 128 CLR 417 ('Whopper Case')	67.96
36	'estate', 'death', 'testator', 'deceased', 'gift', 'share', 'executor', 'life'	<i>Sumpton v Downing</i> (1947) 75 CLR 76	90.83
37	'vehicle', 'motor', 'road', 'motor_vehicle', 'transport', 'driver', 'driving', 'carriage'	<i>Holloway v Pilkington</i> (1972) 127 CLR 391	63.92
38	'letter', 'certificate', 'document', 'march', 'december', 'february', 'signed', 'april'	<i>Snedden v Ng Chong Sun</i> (1969) 121 CLR 413	52.19
39	'document', 'information', 'privilege', 'disclosure', 'warrant', 'advice', 'professional', 'client'	<i>Commissioner of Australian Federal Police v Propend Finance Pty Ltd</i> (1997) 188 CLR 501	67.99
40	'insurance', 'insured', 'society', 'business', 'incurred', 'company', 'expenditure', 'loss'	<i>Producers & Citizens' Co-operative Assurance Co Ltd v Federal Commissioner of Taxation</i> (1972) 128 CLR 63	51.09
41	'land', 'crown', 'title', 'owner', 'value', 'acquisition', 'crown_land', 'grant'	<i>Everingham v Minister for Lands (NSW)</i> (1916) 21 CLR 269	74.19
42	'council', 'water', 'building', 'road', 'local', 'area', 'development', 'street'	<i>Camberwell v Waldmann</i> (1945) 72 CLR 250	61.52
43	'patent', 'invention', 'process', 'specification', 'product', 'mining', 'infringement', 'manufacture'	<i>Day v Perrott</i> (1924) 34 CLR 225	85.09

Topic	Words	Most Representative Case	Score (%)
44	'sentence', 'criminal', 'sentencing', 'offence', 'criminal_appeal', 'court_criminal', 'imprisonment', 'offender'	<i>Mill v The Queen</i> (1988) 166 CLR 59	91.26
45	'property', 'transfer', 'value', 'stamp', 'instrument', 'estate', 'stamp_duty', 'asset'	<i>Day v Commissioner of Stamp Duties (Qld)</i> (1940) 64 CLR 178	65.13
46	'court_appeal', 'submission', 'legislation', 'australian', 'approach', 'context', 'joint', 'requirement'	<i>Equuscorp Pty Ltd v Glengallan Investments Pty Ltd [No 2]</i> (2005) 213 ALR 309	50.39
47	'income', 'commissioner', 'assessment', 'taxpayer', 'taxation', 'commissioner_taxation', 'federal', 'federal_commissioner'	<i>Hughes v Federal Commissioner of Taxation</i> (1958) 98 CLR 345	76.92
48	'work', 'employee', 'employment', 'employer', 'contractor', 'employed', 'hour', 'working'	<i>Hatzimanolis v ANI Corporation Ltd</i> (1992) 173 CLR 473	45.71
49	'fund', 'benefit', 'scheme', 'money', 'contribution', 'appropriation', 'payment', 'superannuation'	<i>Independent Order of Odd Fellows of Victoria v Federal Commissioner of Taxation</i> (1991) 173 CLR 417	49.67

B Fifty Topic Model

