The Effect of Events on Australian Parliamentary Discussion (1851–2017) *

Monica Alexander *University of Toronto*
Rohan Alexander *Australian National University*

We systematically analyse how parliamentary discussion changes in response to different types of events in Australian history. We first create a dataset of what was said in Australian state and federal parliaments from the mid-1800s through to 2017 based on available public records. To reduce the dimensionality of this dataset we then use a topic model that allows correlation and a steady change over time. Finally, we consider the effect of various events using a Bayesian hierarchical Dirichlet model. We find that: 1) elections are associated with topic changes only when the party in power changes; 2) a change in government can be associated with topic changes even if the party in power does not change; 3) economic events, such as financial crises, have less significant and persistent effects than other events such as terrorist attacks. Our findings have implications for how we think about the longer-term trajectory of government policymaking as the media cycle becomes increasingly focused on short-term events.

Keywords: text analysis, politics, Australia

1 Introduction

Government policy is partly driven by parliamentary discussion. Conversely, that same discussion can indicate a government's priorities. But major events–both expected, such as an election, and unexpected, such as a recession or terrorist attack–can affect the course of parliamentary discussion. For instance, think of how a new government often goes to some trouble to appear different to the one they replace, or how events such as the 9/11 attacks altered government priorities.

In this paper we examine text records of what was said in Australian state and federal parliaments. Our earliest record is from 1851 and we consider the time period through to 2017. We use a topic model for dimensionality reduction and to allow for correlation between topics and steady changes over time. We then analyse the topics using a Bayesian hierarchical Dirichlet model to examine changes at various types of events. These events include: changes in government or elections; changes in economic conditions; and other significant events such as the 9/11 attacks.

We find: 1) elections are associated with topic changes only when the party in power changes; 2) a change in government can be associated with topic changes even if the party in power does not change; 3) economic events, such as financial crises, have less

*Thank you to John Tang, Zach Ward, Tim Hatton, Martine Mariotti, Tianyi Wang, Matt Jacob, Leslie Root, Jill Sheppard, Matthew Kerby, Chris Cochrane and Dan Simpson for their helpful suggestions; and to the UC Berkeley Demography Department for the use of their computing resources. Comments on the 06 October 2018 version of this paper welcome at: rohan.alexander@anu.edu.au.

significant and persistent effects than other events such as terrorist attacks. **[More here on the nature of the differences.]**

In this paper we apply a topic model to a dataset of larger-scale parliamentary text records from multiple Australian parliaments and then use a Bayesian hierarchical Dirichlet model to examine events. From a data perspective we contribute **[?????]**. From a methods perspective, we contribute **[?????]**. Our model can be used in many settings where the results of topic models need to be further analysed.

Our work is cross-disciplinary and in the traditions of political economy, economic history, political science and statistics. It contributes to a growing modern quantitative social sciences literature that considers text as an input to more traditional methods, rather than requiring separate analysis. As the digitisation of historical sources continues and computational power becomes cheaper, we expect interest in this approach to only increase.

2 Data

2.1 Parliamentary text data

Following the example of the UK a daily text record called Hansard of what was said in Australian parliaments has been made available since their establishment.[1] Earlier work on the influence of parliaments, such as Van Zanden, Buringh and Bosker (2012), often examined broader activity measures such as counts of sitting days. This allowed for long time frames and wide comparisons. But analysing Hansard records and their equivalents directly are an increasingly viable and popular source of data as new methods and reduced computational costs make larger-scale analysis easier, complementing other measures.

The recent digitisation of the Canadian Hansard, Beelen et al. (2017), allowed Rheault and Cochrane (2018) to examine ideology and party polarisation in Britain and Canada. In the UK, Duthie, Budzynska and Reed (2016) analysed Hansard records to examine which politicians made supportive or aggressive statements toward other politicians between 1979 and 1990 and Peterson and Spirling (2018) examined polarisation. And as digitisation methods improve older UK records can be analysed, for instance Dimitruk (2018) considers the effect of estate bills on prorogations in seventeenth century England. In New Zealand, Curran et al. (2017) modelled the topics discussed between 2003 and 2016, and Graham (2016) examined unparliamentary language between 1890 and 1950. And in the US Gentzkow, Shapiro and Taddy (2018) examine congressional speech records from 1873 to 2016 to find that partisanship has risen in the past few decades.

Australian Hansard records have been analysed for various purposes. For instance, Rasiah (2010) examined Hansard records for the Australian House of Representatives to examine whether politicians attempted to evade questions about Iraq during February and March 2003. And Gans and Leigh (2012) examined Australian Hansard records to

[1] While Hansard is not necessarily verbatim, it is considered close enough for text-as-data purposes. For instance, Mollin (2008) found that in the case of the UK Hansard the differences would only affect specialised linguistic analysis. Edwards (2016) examined Australia, New Zealand and the UK, and found that changes were usually made by those responsible for creating the Hansard record, instead of the parliamentarians.

Table 1: Hansard records used

Parliament	House	Years used	Number of records
Commonwealth	House of Representatives	1901 - 2017	7873
	Senate	1901 - 2017	TBD
Queensland	Legislative Assembly	1860 - 2017	9699
	Legislative Council	1860 - 1922	4156
New South Wales	Legislative Assembly	1879 - 2017	8903
	Legislative Council	1879 - 2017	6463
Victoria	Legislative Assembly	1856 - 2017	-
	Legislative Council	1851 - 2017	-
Tasmania	House of Assembly	TBD - 2017	TBD
	Legislative Council	TBD - 2017	TBD
South Australia	House of Assembly	TBD - 2017	TBD
	Legislative Council	TBD - 2017	TBD
Western Australia	Legislative Assembly	1890 - 2017	5977
	Legislative Council	1870 - 2017	5725

associate mentions by politicians of certain public intellectuals with neutral or positive sentiment.

Australian parliaments generally make their daily Hansard records available online as PDFs and these are considered the official release. Additionally, XML records are available in some cases.[2] We detail the sources of our Hansard PDFs in A.1 and Appendix A.2 provides an example of what a Hansard PDF looks like. There are roughly 54,331 **[UPDATE]** days worth of publicly available Hansard records across the state and federal parliaments (Table 1). Our data cleaning process indicates concerns with a small number of PDFs and these are detailed in Appendix A.3.

The formatting of the Hansard records changes between the different parliaments and over time. We use scripts written in R (R Core Team (2018)) to convert the PDFs into daily text records. An example of the workflow and some reduced-detail scripts are provided in Appendix A.4. Some error is introduced at this stage because many of the records are in a two-column format that need to be separated, and the PDF parsing is not always accurate especially for older records. An example of the latter issue is that 'the' is often parsed as 'thc'. These errors are corrected when they occur at scale and can be identified.

The percentage of stop-words in each record is reasonably consistent over time. This suggests that there is no significant difference in the quality of the parsing over time. Details of this process are provided in Appendix A.5. We mainly use Hansard records on a daily basis in this paper. Text is usually pre-processed before topic models are used.

[2]Tim Sherratt makes these XML records available as a single download and also presents them in a website (http://historichansard.net/) that can be used to explore Commonwealth Hansard records from 1901 to 1980. Commonwealth XML records from 1998 to 2014 are available from Andrew Turpin's website, and from 2006 through to today from Open Australia's website. The records can also be downloaded from the Australian Hansard website.

The specific steps that we take are to remove numbers and punctuation and to change the words to lower case. **[Need to bundle_ngrams i.e. New Zealand is new_zealand.]** Then the sentences are deconstructed and each word considered individually.

3 Model

The models that we use in this paper are the Structural Topic Model (STM) as implemented by the `stm` R package of Roberts, Stewart and Tingley (2018*a*), and Bayesian hierarchical Dirichlet analysis model. In a similar way to Mueller and Rauh (2018), we use topics as an input to another model, in our case to analyse the effect of various events. In Appendix B, we include an alternative approach that follows Taddy (2015) by using word2vec, which more closely uses words, rather than topics, as an input.

The basis of the STM is the Latent Dirichlet Allocation (LDA) model of Blei, Ng and Jordan (2003). In this section we provide a brief overview of both the LDA model and the STM. We consider the outputs of the topic model as reduced dimension inputs that can be analysed within another model. We then discuss the Bayesian hierarchical Dirichlet analysis model that we use for this purpose.

3.1 Latent Dirichlet Allocation

Although more- or less-fine levels of analysis are possible, but here we are primarily interested in considering a day's topics. This means that each day's Hansard record needs to be classified by its topics. Sometimes Hansard records includes titles that make the topic clear. But not every statement has a title and the titles do not always define topics in a well-defined and consistent way, especially over longer time periods. One way to get consistent estimates of the topics discussed in Hansard is to use the LDA method of Blei, Ng and Jordan (2003), for instance as implemented by the R `topicmodels` package of Grün and Hornik (2011).

The key assumption behind the LDA method is that each day's text, 'a document', in Hansard is made by speakers who decide the topics they would like to talk about in that document, and then choose words, 'terms', that are appropriate to those topics. A topic could be thought of as a collection of terms, and a document as a collection of topics, where these collections are defined by probability distributions. The topics are not specified *ex ante*; they are an outcome of the method. In this sense, this approach can be considered unsupervised machine learning. Terms are not necessarily unique to a particular topic, and a document could be about more than one topic. This provides more flexibility than other approaches such as a strict word count method. The goal is to have the words found in each day's Hansard group themselves to define topics.

As applied to Hansard, the LDA method considers each statement to be a result of a process where a politician first chooses the topics they want to speak about. After choosing the topics, the speaker then chooses appropriate words to use for each of those topics. More generally, the LDA topic model works by considering each document as having been generated by some probability distribution over topics. Similarly, each topic could be considered a probability distribution over terms. To choose the terms used in each document the speaker picks terms from each topic in the appropriate proportion. Figures

7 and 8 in Appendix C.1 illustrates an example with five topics, two documents, and ten terms.

Following Blei and Lafferty (2009), Blei (2012) and Griffiths and Steyvers (2004), the process by which a document is generated is more formally considered to be:

1. There are $1, 2, \ldots, k, \ldots, K$ topics and the vocabulary consists of $1, 2, \ldots, V$ terms. For each topic, decide the terms that the topic uses by randomly drawing distributions over the terms. The distribution over the terms for the kth topic is β_k. Typically a topic would be a small number of terms and so the Dirichlet distribution with hyperparameter $0 < \eta < 1$ is used: $\beta_k \sim \text{Dirichlet}(\eta)$.[3] Strictly, η is actually a vector of hyperparameters, one for each K, but in practice they all tend to be the same value.
2. Decide the topics that each document will cover by randomly drawing distributions over the K topics for each of the $1, 2, \ldots, d, \ldots, D$ documents. The topic distributions for the dth document are θ_d, and $\theta_{d,k}$ is the topic distribution for topic k in document d. Again, the Dirichlet distribution with the hyperparameter $0 < \alpha < 1$ is used here because usually a document would only cover a handful of topics: $\theta_d \sim \text{Dirichlet}(\alpha)$. Again, strictly α is vector of length K of hyperparameters, but in practice each is usually the same value.
3. If there are $1, 2, \ldots, n, \ldots, N$ terms in the dth document, then to choose the nth term, $w_{d,n}$:
    a. Randomly choose a topic for that term n, in that document d, $z_{d,n}$, from the multinomial distribution over topics in that document, $z_{d,n} \sim \text{Multinomial}(\theta_d)$.
    b. Randomly choose a term from the relevant multinomial distribution over the terms for that topic, $w_{d,n} \sim \text{Multinomial}(\beta_{z_{d,n}})$.

Given this set-up, the joint distribution for the variables is (Blei (2012), p.6):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p\left(w_{d,n}|\beta_{1:K}, z_{d,n}\right) \right).$$

Based on this document generation process the analysis problem, discussed next, is to compute a posterior over $\beta_{1:K}$ and $\theta_{1:D}$, given $w_{1:D,1:N}$. This is intractable directly, but can be approximated (Griffiths and Steyvers (2004) and Blei (2012)).

After the documents are created, they are all that we have to analyse. The term usage in each document, $w_{1:D,1:N}$, is observed, but the topics are hidden, or 'latent'. We do not know the topics of each document, nor how terms defined the topics. That is, we do not know the probability distributions of Figures 7 or 8. In a sense we are trying to reverse the document generation process – we have the terms and we would like to discover the topics.

[3] The Dirichlet distribution is a variation of the beta distribution that is commonly used as a prior for categorical and multinomial variables. If there are just two categories, then the Dirichlet and the beta distributions are the same. In the special case of a symmetric Dirichlet distribution, $\eta = 1$, it is equivalent to a uniform distribution. If $\eta < 1$, then the distribution is sparse and concentrated on a smaller number of the values, and this number decreases as η decreases. A hyperparameter is a parameter of a prior distribution.

If the earlier process around how the documents were generated is assumed and we observe the terms in each document, then we can obtain estimates of the topics (Steyvers and Griffiths (2006)). The outcomes of the LDA process are probability distributions and these define the topics. Each term will be given a probability of being a member of a particular topic, and each document will be given a probability of being about a particular topic. That is, we are trying to calculate the posterior distribution of the topics given the terms observed in each document (Blei (2012), p. 7):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N}) = \frac{p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}\right)}{p(w_{1:D,1:N})}.$$

Gibbs sampling or the variational expectation-maximization algorithm can be used to approximate the posterior. A summary of these approaches is provided in Appendix C.2. The choice of the number of topics, k, drives the results and must be specified *a priori*. If there is a strong reason for a particular number, then this can be used. Otherwise, one way to choose an appropriate number is to use cross validation.

One weakness of the LDA method is that it considers a 'bag of words' where the order of those words does not matter (Blei (2012)). It is possible to extend the model to reduce the impact of the bag-of-words assumption and add conditionality to word order. Additionally, alternatives to the Dirichlet distribution can be used to extend the model to allow for correlation. For instance, in Hansard topics related the army may be expected to be more commonly found with topics related to the navy, but less commonly with topics related to banking. This motivates the use of the Structural Topic Model, described in the next section.

3.2 Structural Topic Model

The distinguishing aspect of the Structural Topic Model (STM) of Roberts, Stewart and Airoldi (2016) is that it considers more than just a document's content when constructing topics. For instance, we generally have some information about the author and the date that a document was created. In the case of Hansard, we know who was speaking and the date they spoke. The STM allows this additional information to affect the construction of topics, though influencing either topical prevalence or topical content. That said, the assumption that there is some document generation process is the same as the LDA method, it is just that this process now includes metadata.

The STM is set-up to most easily include metadata to do with prevalence and content. Prevalence relates to the topic proportions in each document. For instance, we expect that topics related to the reasons for Federation, such as tariffs and trade, should be more prevalent in those earlier years than later. Similarly, we may expect topics to do with terrorism to be more prevalent in recent years. Content relates to the words that make up each topic. For instance, there are changes in the use of language over the period for which we have data, and it would be better for these to not be responsible for defining different topics rather than being part of the same topic. The prevalence meta-data for the dth document are in X_d, which has one column for each aspect. For instance, if there were 10 documents and each had a date and an author, then X would be 10×2. Similarly, the content meta-data are

As with LDA, the process assumed to generate the documents is the key aspect as this will be reversed to estimate the topics. The document generation process of Blei, Ng and Jordan (2003) discussed earlier, is slightly modified by Roberts, Stewart and Airoldi (2016) for the STM:

1. As with LDA, the topic distributions, that is, the proportion of a document dedicated to a topic, for the dth document are θ_d, and θ is a vector with length D. In contrast to LDA, this is a drawn from a logistic-normal distribution, parameterised such that the mean of that distribution, μ, is affected by a vector of document covariates, X_d (following Roberts, Stewart and Tingley (2018*b*), p.3):

$$\theta_d | X_d \gamma \Sigma \sim \text{Logistic Normal}(\mu = X_d \gamma, \Sigma)$$

2. To decide the distribution over terms for each topic, $\beta_{d,k}$, start with some baseline distribution over the terms, m. Topic-k-specific deviations from this are controlled by $\kappa_k^{(t)}$, deviations due to the document meta-data are controlled by $\kappa_{y_d}^{(c)}$, and the interaction between these two deviations is controlled by $\kappa_{y_d,k}^{(i)}$:

$$\beta_{d,k} \propto \exp\left(m + \kappa_k^{(t)} + \kappa_{y_d}^{(c)} + \kappa_{y_d k}^{(i)}\right)$$

3. Then if there are n terms in the dth document, then to choose the nth term, $w_{d,n}$:

    a. Randomly choose a topic for that term from the document-specific multinomial distribution over topics.
    b. Randomly choose a term from the topic-specific multinomial distribution over terms.

We primarily implement the STM on the daily-level parliamentary text data described earlier using the `stm` R package of Roberts, Stewart and Tingley (2018*a*). We consider both topic prevalence and content to be functions of time. The choice of the number of topics to use in the model is a situation-specific compromise. We use a standard diagnostic approach to decide on 80 **[UPDATE]** topics. More detail on this selection process is available in Appendix C.3.

Figure 1 illustrates an output of the STM. This illustrative model only considers five per cent of data from the House of Representatives between 1901 and 2017. It shows how each day's parliamentary discussion can be allocated to a topic and highlights how the prominence of these topics changes over time.

3.3 Analysis model

The main output of the STM that is of interest to us is the share of each topic in each document. We would like to know how those shares change given events. But with around 80 **[UPDATE]** topics for each of the roughly 54,331 **[UPDATE]** sitting days across the parliaments and houses the data are still too noisy to easily visualise changes around

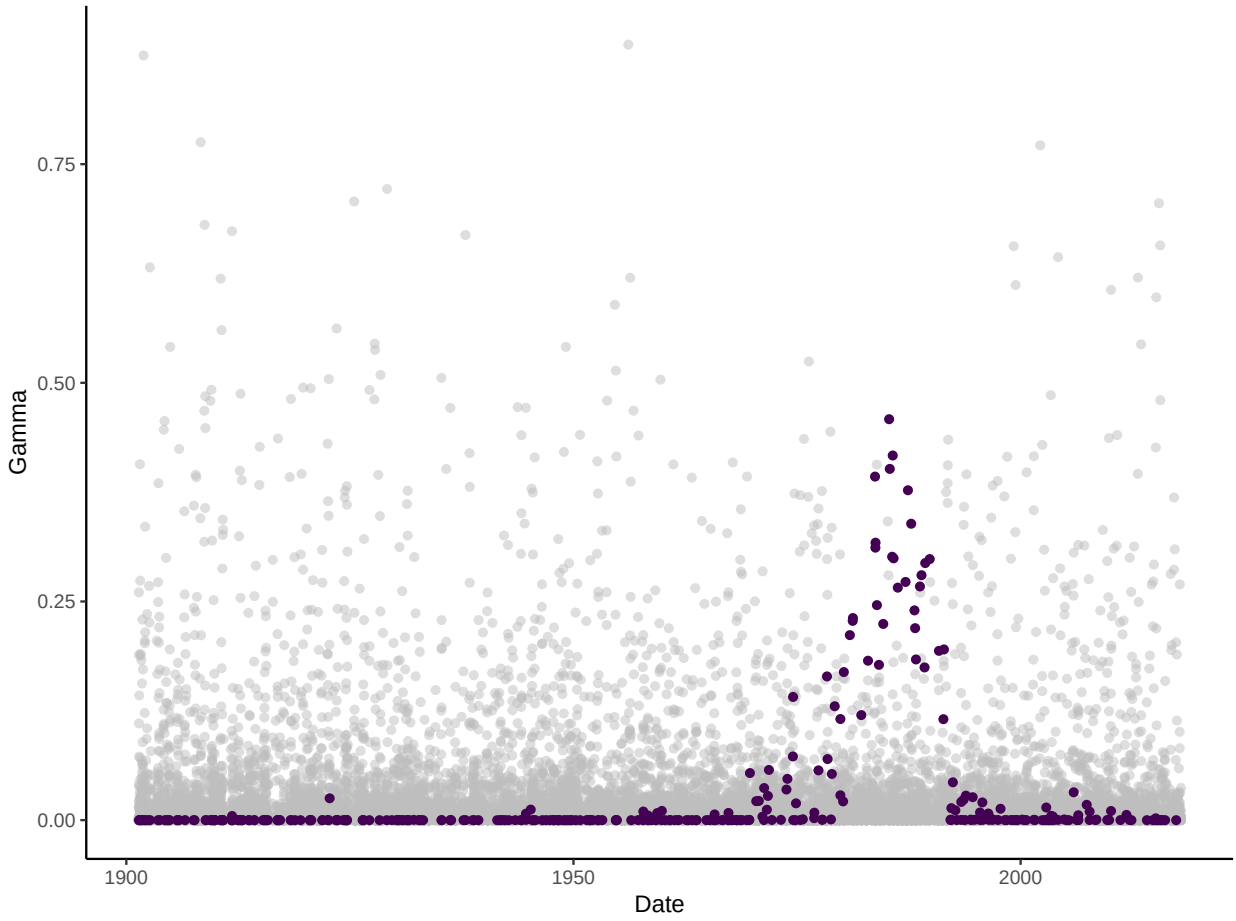

Figure 1: Example topic model output, with Topic 17 highlighted

events. Instead, we put together a Bayesian hierarchical Dirichlet model that takes those shares as an input.

We index each proportion, γ, by sitting period s, time t and topic p, that is γ_{stp}. Note that time t refers to the day index within a sitting period which is any collection of days separated from other days by at least a week. Our analysis model is:

$$\gamma_{stp} \sim \text{Dirichlet}(\mu_{stp}) \tag{1}$$
$$\log \mu_{stp} = \alpha_{sp} + \delta_{stp} \tag{2}$$
$$\alpha_{s1} = 0 \tag{3}$$
$$\alpha_{sp} \sim \text{Normal}\left(\eta_{g[s]}, \sigma_g^2\right) \text{ for } p > 1 \tag{4}$$
$$\delta_{s,t,p} \sim \text{Normal}\left(\delta_{s,t-1,p}, \sigma_\delta^2\right) \tag{5}$$
$$\sigma_\delta^2 \sim \text{Uniform}(0, 40). \tag{6}$$

[Does the variance parameter used in Equation 4 need to be specified too?]

Equation (1) describes the proportion of each day within a sitting period given to a particular topic as a draw from a Dirichlet distribution parameterised by μ_{stp}. In a sense, μ_{stp} will be mean topic distributions over whatever time period is of interest. The Dirichlet distribution is the multivariate extension of the Beta distribution, and the parameter controls where the 'weight' of the distribution lies. Equation (2) describes that distribution parameter, μ_{stp}, as the sum of α_{sp}, which are sitting period specific topic distributions and mean the model is linear in time for each topic and sitting period, and δ_{stp}, which is a simple time series that allows for a little variation within a sitting period. These are themselves parameters. The primary aim of the modelling task is to find an appropriate and meaningful functional form for μ.

Equation (3) constrains the first intercept to zero for identification purposes. Equation (4) shows how we consider α_{sp} as a draw from a Normal distribution with a mean that depends on **[SOMETHING]** and a variance that depends on **[SOMETHING ELSE]**. Similarly, Equation (5) show how we model δ_{stp} as a draw from a Normal distribution with a mean that depends on the preceding δ and some variance which is drawn from the Uniform distribution in Equation (6), that is, a random walk around each intercept.

We run the model in Stan using the `RStan` package of Stan Development Team (2018). An illustration of the validity of our analysis model using simulated data is in Appendix E.

[Probably need to add: 1) where does each covariate come into it and how do they interact; 2) how does correlation work; 3) what are the random effects and fixed effects?]

4 Results

We are interested in considering the effect of various events on what is talked about in Australia's parliaments. Each of the parliaments and their houses are treated independently here. Future work could expand the model to better understand, and allow, for correlation between them.

Political events are those related to a change of government or an election. Economic events are defined by substantial changes in various economic measures, such as the onset of the Great Depression or floating the currency. Other events are those such as the 9/11 attacks. The full list of events that we consider are detailed in Appendix D.

We consider four types of events and group the day's topic proportions by these events. One unfortunate effect of this is that we only allow an event to have a long-run effect and ignore short-run effects. Another is that we ignore coincidence of events.

Figure 2 shows the grouping by election. If elections had significant effects on the discussion in parliament then there would be considerable change between groups. However changes generally appear to be longer-term rather than election to election.

Figure 3 shows the grouping by government. This can be different to election groupings, for instance, the change from the Rudd to Gillard governments happened without an election. Similarly, John Howard's three terms are all considered the one government for our purposes.

Here we see much more difference between adjacent events. This suggests that new governments tend to talk about different topics than the government they replace.

Figure 4 shows the grouping by major economic changes. There are considerable changes between adjacent events.

Figure 5 shows the grouping by major other events, such as the 9/11 attacks.

Another way to consider the effect of events is by including incrementing variables in the prevelance metadata. This then gets included in the topic modelling and the significance to each topic can be computed. Table 2 summarises these by the type of dummy and the topic. At the five per cent level the events that we consider were significant for eight topics. This significance tended to be correlated, especially for economic and other events.

For instance, all of the incrementing variables have a significant effect on the prevalence of Topic 1. Table 3 shows that this has to do with tariffs and trade. On the other hand, Topic 2 has to do with aspects of daily life such as community and children, and Topic 3 has to do with legal issues and neither is significant. This is not surprising given the centrality of these concerns at all times.

5 Summary and conclusions

In this paper we examined what was said in Australia's parliaments. We downloaded and parsed PDFs for Australian states and federal parliaments. We then used a text model to group the discussions into topics and analysed the effect of various events on the distribution of the discussion.

In general we found that changes in government changed the distribution of topics discussed in parliament, but that elections did not. We found that significant events such as 9/11 had substantial and lasting changes, but that with certain exceptions, economic events did not.

Text analysis has well-known biases and weaknesses and is a complement to more detailed analysis such as qualitative methods and case studies. We consider the results presented in this paper, as well as many of those results of the larger text-as-data research program, as fitting within findings based on other methods.

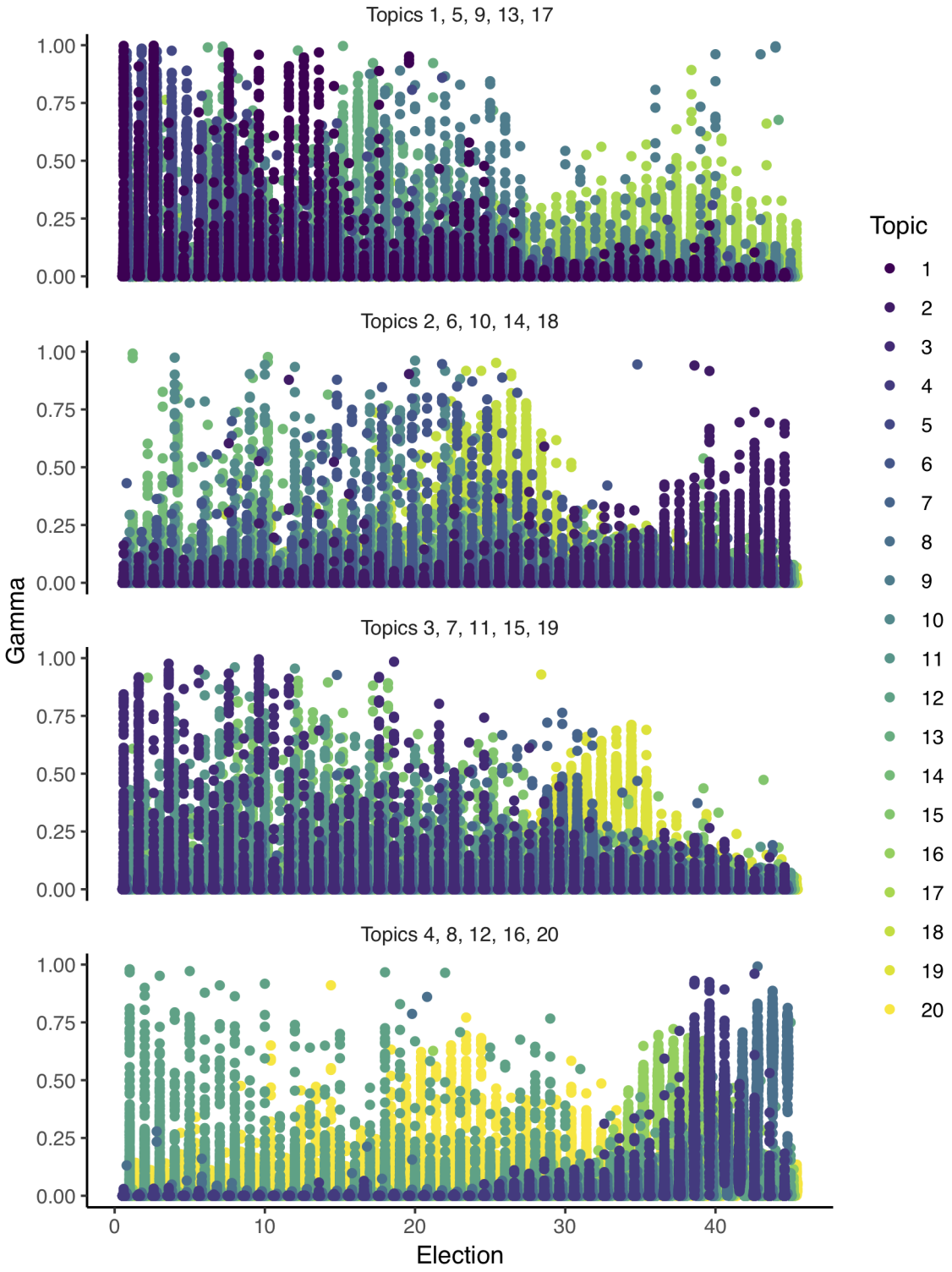

Figure 2: Topic proportions grouped by elections

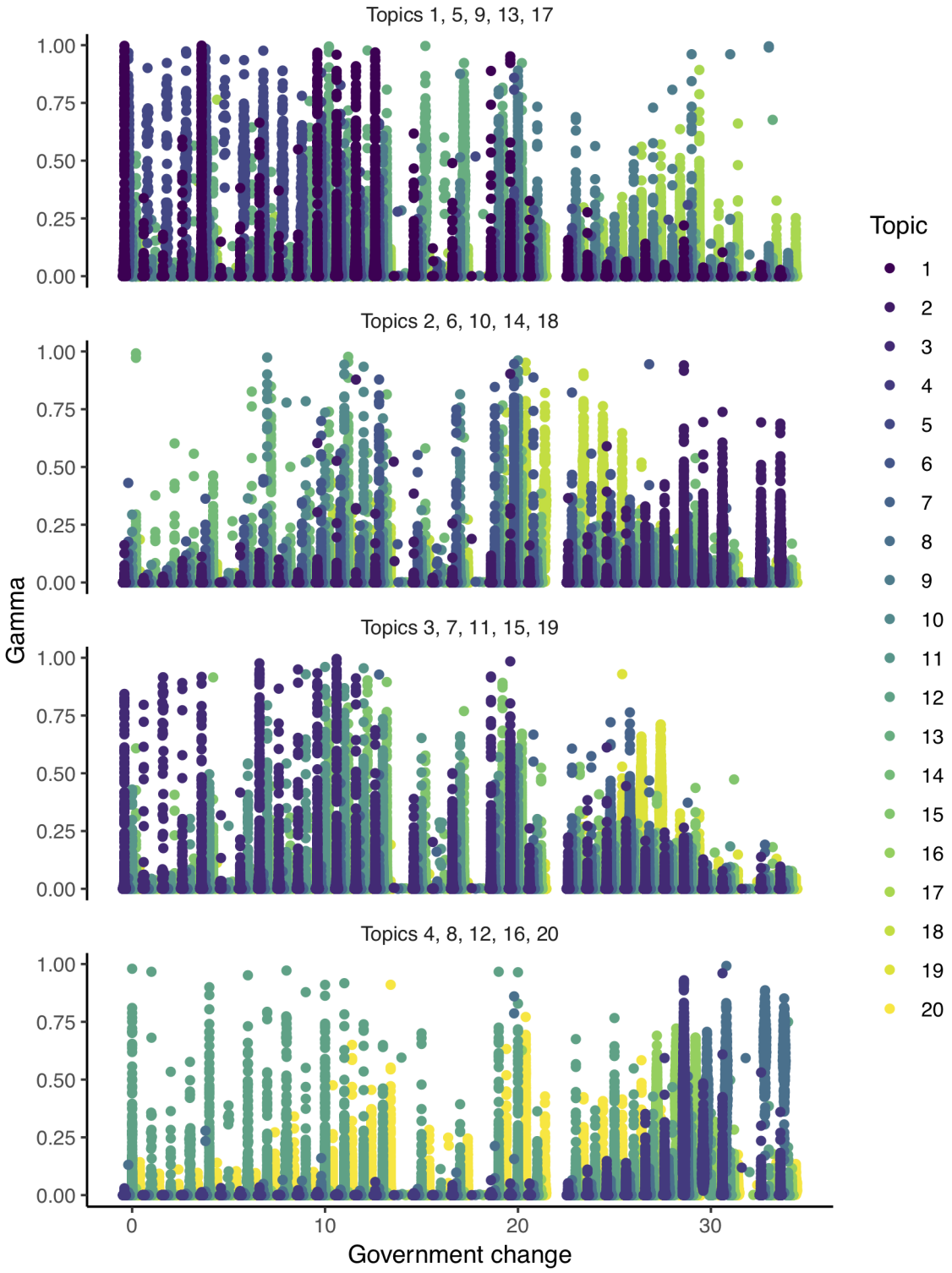

Figure 3: Topic proportions grouped by government

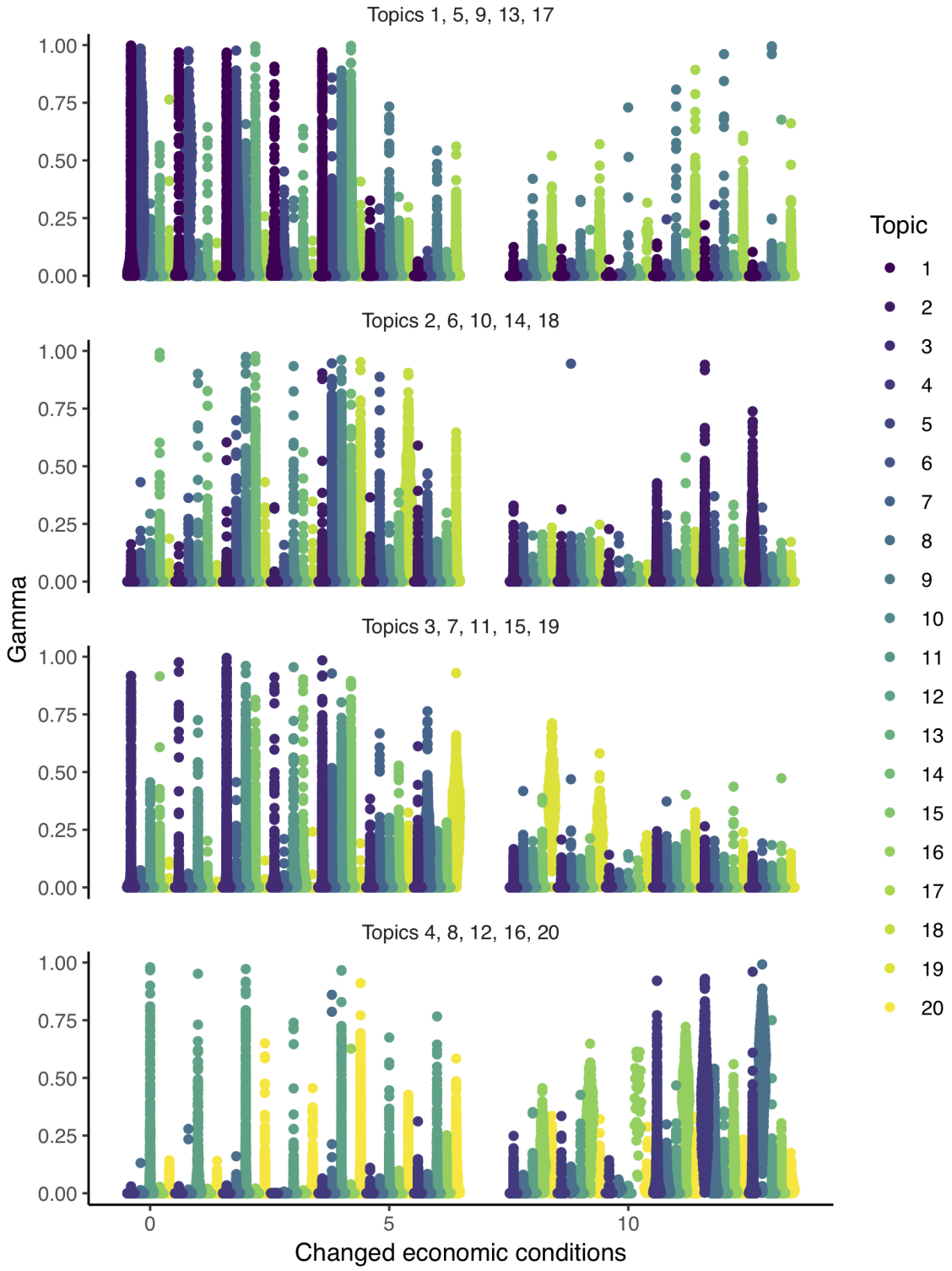

Figure 4: Topic proportions grouped by economic events

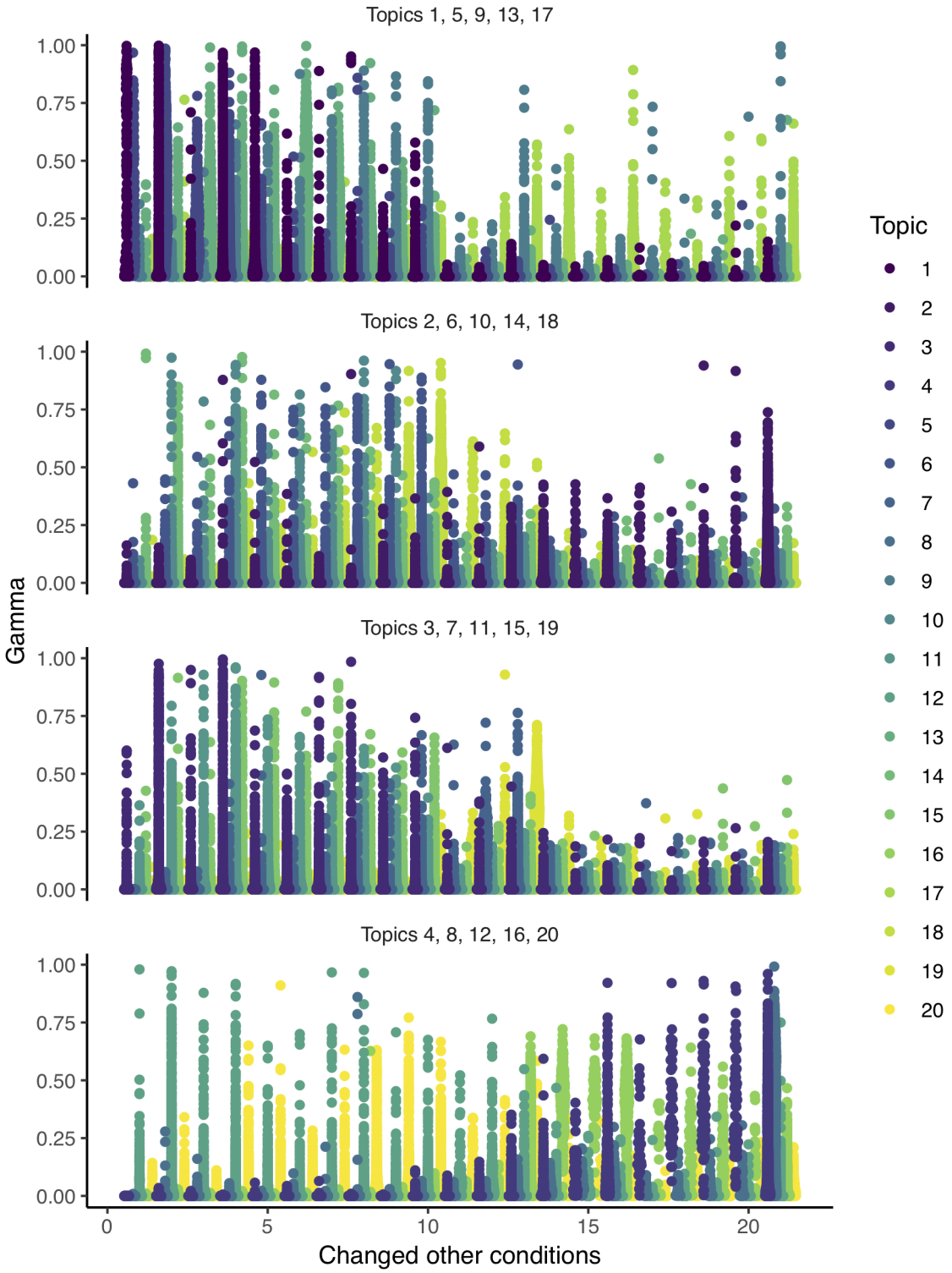

Figure 5: Topic proportions grouped by other events

Table 2: Significance of dummy

Topic	Elections	Governments	Economic	Other
1	0.00	0.00	0.00	0.04
2	0.50	0.69	0.15	0.64
3	0.18	0.36	0.28	0.58
4	0.90	0.20	0.00	0.00
5	0.88	0.15	0.15	0.70
6	0.41	0.00	0.21	0.20
7	0.03	0.00	0.00	0.00
8	0.61	0.39	0.00	0.00
9	0.18	0.24	0.83	0.62
10	0.99	0.34	0.96	0.41
11	0.00	0.13	0.18	0.22
12	0.03	0.02	0.06	0.27
13	0.00	0.00	0.00	0.00
14	0.01	0.37	0.22	0.28
15	0.62	0.50	0.12	0.64
16	0.74	0.00	0.00	0.00
17	0.01	0.40	0.80	0.16
18	0.23	0.00	0.00	0.00
19	0.31	0.00	0.00	0.00
20	0.00	0.54	0.09	0.35

Table 3: Top words for each topic

Topic	Top words
1	tariff, duties, trade, customs, protection, item, industries
2	community, life, children, family, australians, women, rights
3	court, industrial, union, law, trade, arbitration, workers
4	services, community, care, education, funding, health, school
5	joseph, connection, money, federal, sydney, officers, service
6	pension, health, medical, scheme, service, week, social
7	total, roads, wul, expenditure, road, budget, assistance
8	health, billion, business, electorate, senator, community, budget
9	united, world, nations, international, war, defence, foreign
10	bank, royal, private, control, broadcasting, banking, money
11	taxation, income, treasurer, revenue, money, land, financial
12	speaker, vote, motion, senate, election, standing, electoral
13	war, defence, service, british, ill, britain, soldiers
14	territory, oil, northern, line, shipping, ships, company
15	wheat, wool, primary, production, farmers, export, zealand
16	tion, business, ing, governments, issue, speaker, health
17	amendments, section, information, service, person, provisions, law
18	education, service, development, housing, capital, territory, services
19	speaker, deputy, petitioners, petition, ing, pray, citizens
20	budget, economic, hear, employment, unemployment, governments, economy

While using text as data has well-known shortcomings, it allows larger-scale analysis that would not be viable using less-automated approaches and so it can identify patterns that may otherwise be overlooked.

A Hansard details

A.1 Hansard sources

Where from?

Which years are being used (not non-OCRd)

[IS THIS BEING REFERENCED IN THE TEXT????]

A.2 Example Hansard page

Figure 6

A.3 Known Hansard issues

The NSW Legislative Council was established earlier than 1856, however the earlier Hansard records have not been through an independent OCR process and were not used in this paper. However, the Google Tesseract OCR engine as implemented by Ooms (2018*b*) provided useful data and these could be used in the future.

Which PDFs are missing or have no content, etc.

Notes: Positive means I am missing some, Negative means I have too many. Source: https://www.aph.gov.au/Parliamentary_Business/Statistics/Senate_StatsNet/General/sittingdaysyear

A.4 Example Hansard PDf to text record workflow

Example of the workflow from PDF to text

These scripts are primarily based on: the `PDFtools` package of Ooms (2018*a*); the `tidyverse` package of Wickham (2017); the `tm` package of Feinerer and Hornik (2018); the `lubridate` package of Grolemund and Wickham (2011); and the `stringi` package of Gagolewski (2018). The functions of those packages are supported by: the `furrr` package of Vaughan and Dancho (2018); and the `tictoc` package of Izrailev (2014). The `hunspell` package of Ooms (2017) is also used to help find spelling issues. In addition to the packages already mentioned, in this step the R scripts to do this use the `tidytext` R package of Silge and Robinson (2016).

A.5 Stopwords over time

Insert graph of stop words over time.

House of Representatives.

Thursday, 6 February, 1902.

Mr. SPEAKER took the chair at 2.30 p.m., and read prayers.

PUNCHING AND SHEARING MACHINES.

Mr. R. EDWARDS.—I should like to know from the Minister for Trade and Customs whether, as the amendment of the honorable and learned member for Corio, placing various machines and tools of trade upon the free list, was carried, the Government are prepared to exempt punching and shearing machines.

Mr. KINGSTON.—I think that the fair construction of the determination arrived at by the committee yesterday necessitates the exemption of punching and shearing machines, and the Government therefore propose to admit them duty free from to-day.

SOUTH AUSTRALIAN PREFERENTIAL RAILWAY RATES.

Mr. THOMAS.—I wish to ask the Minister for Home Affairs if the report which appeared in the newspapers a few days ago, to the effect that the South Australian Government do not intend to charge preferential rates upon their railways after the 1st February, is correct?

Sir WILLIAM LYNE.—I have received no definite information upon the subject from the South Australian Government. I forwarded a communication to the Minister for Railways in South Australia in reference to these rates some time ago, and his reply was to the effect that the South Australian Government desired to, as far as possible, assimilate the rates for the produce of all the States, but that up to the present time, although there had been several conferences upon the subject, they had been unsuccessful, and that he had requested the Railways Commissioner to report further. I had another telegram or letter to-day, which I have not by me now, but it does not carry the matter much further.

PAPER.

Mr. DEAKIN laid upon the table—

Minute by the Prime Minister to His Excellency the Governor-General, relating to the contract for supplies for troops in South Africa.

SYDNEY TELEGRAPHIC BUSINESS.

Mr. THOMSON.—Is the Minister who represents the Postmaster-General yet in possession of a return which has been promised by the Government, showing the lengths of telegrams sent in one day from the Sydney and suburban offices?

Mr. DEAKIN.—I mentioned the matter to my honorable colleague, Sir Philip Fysh, and he told me that he proposed to inform the honorable member that he had received a return, but that, thinking it was not quite in compliance in all particulars with the honorable member's request, he referred it back to have further information added. He is expecting to receive the return again at any moment.

Mr. JOSEPH COOK.—Will the Government keep back the consideration of the Postal Rates Bill until the return has been presented to the House?

Mr. DEAKIN.—I shall call the attention of the Postmaster-General to the honorable member's wish.

QUARANTINE ADMINISTRATION.

Mr. MAHON asked the Prime Minister, *upon notice*—

1. Has his attention been drawn to complaints concerning the administration by State Governments of the quarantine laws and regulations?

Figure 6: Example Hansard page - 6 February 1902

Table 4: Missing days

year	HoR PDFs	Senate sitting days	HoR sitting days	HoR difference
1902	106	93	107	1
1908	93	84	91	-2
1909	97	71	98	1
1918	87	68	86	-1
1920	113	76	114	1
1921	92	79	93	1
1934	36	22	35	-1
1935	54	37	55	1
1942	44	36	45	1
1948	89	39	90	1
1951	55	40	56	1
1955	53	36	52	-1
1974	64	64	62	-2
1985	65	74	66	1
1991	66	83	67	1
1992	60	76	44	-16
1993	47	53	46	-1
1994	69	80	68	-1
1995	71	78	70	-1
1997	79	82	76	-3
1998	56	57	54	-2
2000	71	71	73	2
2002	68	60	69	1
2003	74	64	75	1
2004	58	49	59	1
2012	63	57	67	4

B word2vec alternative

An alternative approach that follows Taddy (2015).

C Topic modelling example and details

C.1 Examples

For instance, if there were five topics and two documents, then the first document may be comprised mostly of the first few topics; the other document may be mostly about the final few topics (Figure 7).

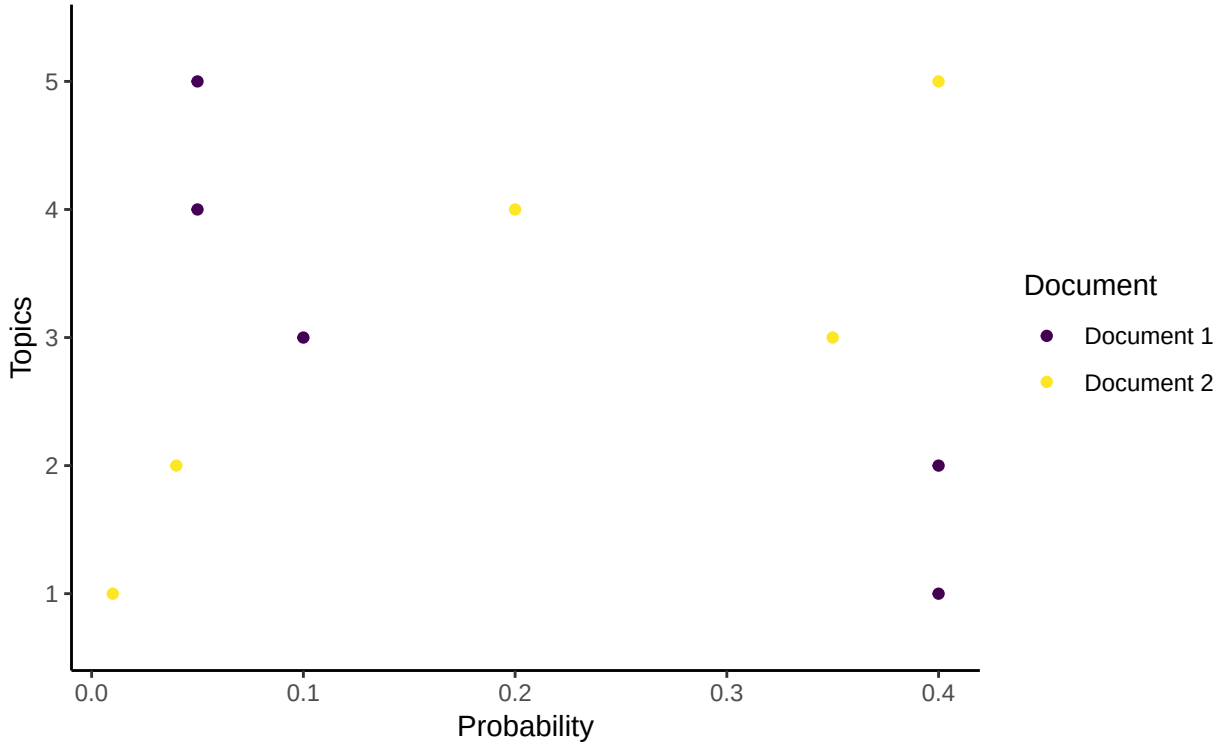

Figure 7: Probability distributions over topics for two documents

For instance, if there were ten terms, then one topic could be defined by giving more weight to terms related to immigration; and some other topic may give more weight to terms related to the economy (Figure 8).

C.2 Posterior estimation

Following Steyvers and Griffiths (2006) and Darling (2011), the Gibbs sampling process attempts to find a topic for a particular term in a particular document, given the topics of all other terms for all other documents. Broadly, it does this by first assigning every term in every document to a random topic, specified by Dirichlet priors with $\alpha = \frac{50}{K}$ and $\eta = 0.1$ (Steyvers and Griffiths (2006) recommends $\eta = 0.01$), where α refers to the distribution over topics and η refers to the distribution over terms (Grün and Hornik (2011), p. 7). It then selects a particular term in a particular document and assigns it to a new topic based on the conditional distribution where the topics for all other terms in all

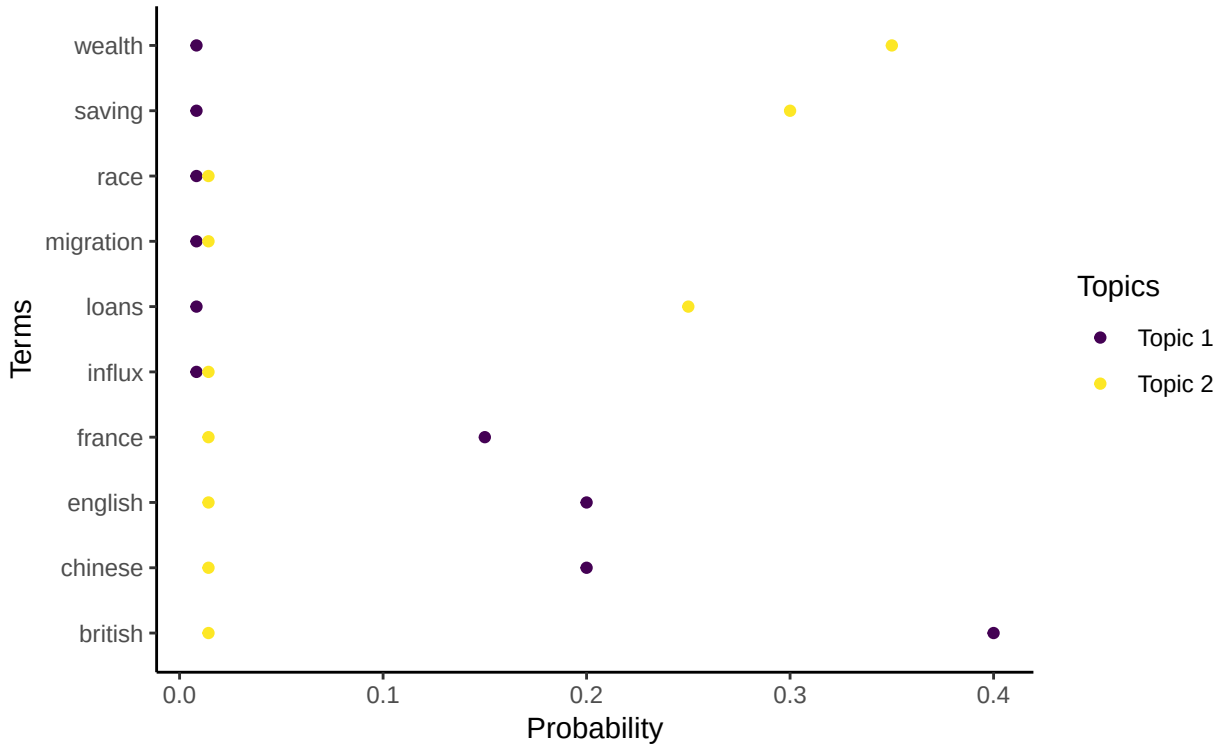

Figure 8: Probability distributions over terms

documents are taken as given (Grün and Hornik (2011), p. 6):

$$p(z_{d,n} = k | w_{1:D,1:N}, z'_{d,n}) \propto \frac{\lambda'_{n\to k} + \eta}{\lambda'_{\cdot\to k} + V\eta} \frac{\lambda'^{(d)}_{n\to k} + \alpha}{\lambda'^{(d)}_{-i} + K\alpha}$$

where $z'_{d,n}$ refers to all other topic assignments; $\lambda'_{n\to k}$ is a count of how many other times that term has been assigned to topic k; $\lambda'_{\cdot\to k}$ is a count of how many other times that any term has been assigned to topic k; $\lambda'^{(d)}_{n\to k}$ is a count of how many other times that term has been assigned to topic k in that particular document; and $\lambda'^{(d)}_{-i}$ is a count of how many other times that term has been assigned in that document. Once $z_{d,n}$ has been estimated, then estimates for the distribution of words into topics and topics into documents can be backed out.

This conditional distribution assigns topics depending on how often a term has been assigned to that topic previously, and how common the topic is in that document (Steyvers and Griffiths (2006)). The initial random allocation of topics means that the results of early passes through the corpus of document are poor, but given enough time the algorithm converges to an appropriate estimate.

C.3 Selection of number of topics

[TBD]

C.4 Robustness of results

[IS THIS BEING USED?] Here we change the number of sitting days considered either side of an event. The results in the main section of the paper are for the nearest ten days either side of an event. Here are show that the results are essentially the same if the nearest one, two, five, and twenty days either side of an event.

D Events

Add the graphs and procedures.

Table 5: Change in governments

government	primeMinister	party	start	end	diedInOffice
Barton	Edmund Barton	Protectionist	1901-01-01	1903-09-24	No
Deakin 1	Alfred Deakin	Protectionist	1903-09-24	1904-04-27	No
Watson	Chris Watson	Labour	1904-04-27	1904-08-18	No
Reid	George Reid	Free Trade	1904-08-18	1905-07-05	No
Deakin 2	Alfred Deakin	Protectionist	1905-07-05	1908-11-13	No
Fisher 1	Andrew Fisher	Labour	1908-11-13	1909-06-02	No
Deakin 3	Alfred Deakin	Commonwealth Liberal	1909-06-02	1910-04-29	No
Fisher 2	Andrew Fisher	Labor	1910-04-29	1913-06-24	No
Cook	Joseph Cook	Commonwealth Liberal	1913-06-24	1914-09-17	No
Fisher 3	Andrew Fisher	Labor	1914-09-17	1915-10-27	No
Hughes	Billy Hughes	Labor, National Labor and Nationalist	1915-10-27	1923-02-09	No
Bruce	Stanley Bruce	Nationalist (Coalition)	1923-02-09	1929-10-22	No
Scullin	James Scullin	Labor	1929-10-22	1932-01-06	No
Lyons	Joseph Lyons	United Australia (Coalition)	1932-01-06	1939-04-07	Yes
Page	Earle Page	Country (Coalition)	1939-04-07	1939-04-26	No
Menzies 1	Robert Menzies	United Australia (Coalition)	1939-04-26	1941-08-28	No
Fadden	Arthur Fadden	Country (Coalition)	1941-08-28	1941-10-07	No
Curtin	John Curtin	Labor	1941-10-07	1945-07-05	Yes
Forde	Frank Forde	Labor	1945-07-06	1945-07-13	No
Chifley	Ben Chifley	Labor	1945-07-13	1949-12-19	No
Menzies 2	Robert Menzies	Liberal (Coalition)	1949-12-19	1966-01-26	No
Holt	Harold Holt	Liberal (Coalition)	1966-01-26	1967-12-19	Yes
McEwen	John McEwen	Country (Coalition)	1967-12-19	1968-01-10	No
Gorton	John Gorton	Liberal (Coalition)	1968-01-10	1971-03-10	No
McMahon	William McMahon	Liberal (Coalition)	1971-03-10	1972-12-05	No
Whitlam	Gough Whitlam	Labor	1972-12-05	1975-11-11	No
Fraser	Malcolm Fraser	Liberal (Coalition)	1975-11-11	1983-03-11	No
Hawke	Bob Hawke	Labor	1983-03-11	1991-12-20	No
Keating	Paul Keating	Labor	1991-12-20	1996-03-11	No
Howard	John Howard	Liberal (Coalition)	1996-03-11	2007-12-03	No
Rudd 1	Kevin Rudd	Labor	2007-12-03	2010-06-24	No
Gillard	Julia Gillard	Labor	2010-06-24	2013-06-27	No
Rudd 2	Kevin Rudd	Labor	2013-06-27	2013-09-18	No
Abbott	Tony Abbott	Liberal (Coalition)	2013-09-18	2015-09-15	No
Turnbull	Malcolm Turnbull	Liberal (Coalition)	2015-09-15	2018-08-24	No
Morrison	Scott Morrison	Liberal (Coalition)	2018-08-24	NA	NA

Table 6: Elections

year	electionDate	electionWinner
1901	1901-03-29	Non-labor
1903	1903-12-16	Non-labor
1906	1906-12-12	Non-labor
1910	1910-04-13	Labor
1913	1913-05-31	Non-labor
1914	1914-09-05	Labor
1917	1917-05-05	Non-labor
1919	1919-12-13	Non-labor
1922	1922-12-16	Non-labor
1925	1925-11-14	Non-labor
1928	1928-11-17	Non-labor
1929	1929-10-12	Labor
1931	1931-12-19	Non-labor
1934	1934-09-15	Non-labor
1937	1937-10-23	Non-labor
1940	1940-09-21	Non-labor
1943	1943-08-21	Labor
1946	1946-09-28	Labor
1949	1949-12-10	Non-labor
1951	1951-08-28	Non-labor
1954	1954-05-29	Non-labor
1955	1955-12-10	Non-labor
1958	1958-11-22	Non-labor
1961	1961-12-09	Non-labor
1963	1963-11-30	Non-labor
1966	1966-11-26	Non-labor
1969	1969-10-25	Non-labor
1972	1972-12-02	Labor
1974	1974-05-18	Labor
1975	1975-12-13	Non-labor
1977	1977-12-10	Non-labor
1980	1980-10-18	Non-labor
1983	1983-03-05	Labor
1984	1984-12-01	Labor
1987	1987-07-11	Labor
1990	1990-03-24	Labor
1993	1993-03-13	Labor
1996	1996-03-02	Non-labor
1998	1998-10-03	Non-labor
2001	2001-11-10	Non-labor
2004	2004-10-09	Non-labor
2007	2007-11-24	Labor
2010	2010-08-21	Labor
2013	2013-09-07	Non-labor
2016	2016-07-02	Non-labor

Table 7: Key economic events

theDate	event
1907-11-08	Harvester case
1910-09-01	Australian pound introduced (CHECK DATE)
1929-10-29	Black Tuesday Stock Market Crash
1931-06-01	Premiers' Plan (CHECK DATE)
1966-02-14	Decimalisation
1973-01-01	Fred Gruen 25% tariff cut (DATE IS WRONG)
1983-12-12	Australian dollar is floated
1984-02-01	Medicare establised
1987-10-19	Black Monday Stock Market Crash
1991-02-10	State Bank of South Australia collapse
1990-08-27	State Bank of Victoria collapse (CHECK DATE)
2000-07-01	GST introduced
2008-09-15	Lehman Brothers bankruptcy

Table 8: Key other events

theDate	event
1899-10-11	Second Boer War starts
1901-01-01	Australia federated
1902-05-31	Second Boer War ends
1914-07-28	World War I starts
1918-11-11	World War I ends
1932-05-13	Jack Lang dismissed as NSW Premier
1939-09-01	World War II starts
1945-09-02	World War II ends
1949-10-17	Snowy Hydro construction begins
1956-11-22	Melbourne Olympics
1962-08-03	Australia enters Vietnam War
1972-12-02	Australia exits Vietnam War
1973-10-20	White Australian Policy ended
1975-11-11	The Dismissal
1992-06-03	Mabo
1996-12-23	Wik decision
1996-03-28	Port Arthur massacre
1999-09-20	INTERFET deployment begins
2000-02-28	INTERFET deployment ends
2000-09-15	Sydney Olympics
2001-09-11	9/11 attack
2002-10-12	Bali bombings

E Analysis model

Here we illustrate the validity of our analysis model using simulated data.

References

Beelen, Kaspar, Timothy Alberdingk Thim, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, Roman Polyanovsky and Tanya Whyte. 2017. "Digitization of the Canadian Parliamentary Debates." *Canadian Journal of Political Science* pp. 1–16.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan):993–1022.

Blei, David M and John D Lafferty. 2009. Topic Models. In *Text Mining*. Chapman and Hall/CRC pp. 101–124.

Curran, B., K. Higham, E. Ortiz and D. Vasques Filho. 2017. "Look Who's Talking: Bipartite Networks as Representations of a Topic Model of New Zealand Parliamentary Speeches." *ArXiv e-prints* .

Darling, William M. 2011. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. pp. 642–647.

Dimitruk, Kara. 2018. "I Intend Therefore to Prorogue: The Effects of Political Conflict and the Glorious Revolution in Parliament, 1660–1702." *European Review of Economic History* 22(3):261–297.

Duthie, Rory, Katarzyna Budzynska and Chris Reed. 2016. Mining Ethos in Political Debate. In *Computational Models of Argument*, ed. P Baroni, TF Gordon, T Scheffler and M Stede. Vol. 287 pp. 299–310.

Edwards, Cecilia. 2016. "The Political Consequences of Hansard Editorial Policies: The Case for Greater Transparency." *Australasian Parliamentary Review* 31(2):145–160.

Feinerer, Ingo and Kurt Hornik. 2018. *tm: Text Mining Package.* R package version 0.7-5.
URL: *https://CRAN.R-project.org/package=tm*

Gagolewski, Marek. 2018. *R Package stringi: Character String Processing Facilities*.
URL: *http://www.gagolewski.com/software/stringi/*

Gans, Joshua and Andrew Leigh. 2012. "How Partisan is the Press? Multiple Measures of Media Slant." *The Economic Record* 88(280):127–147.

Gentzkow, Matthew, Jesse M. Shapiro and Matt Taddy. 2018. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. Technical report Voldemort's University.
URL: *http://web.stanford.edu/ gentzkow/research/politext.pdf*

Graham, Ruth. 2016. Withdraw and Apologise: A Diachronic Study of Unparliamentary Language in the New Zealand Parliament, 1890–1950 PhD thesis.

Griffiths, Thomas and Mark Steyvers. 2004. "Finding Scientific Topics." *PNAS* 101:5228–5235.

Grolemund, Garrett and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40(3):1–25.
URL: *http://www.jstatsoft.org/v40/i03/*

Grün, Bettina and Kurt Hornik. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13):1–30.

Izrailev, Sergei. 2014. *tictoc: Functions for Timing R Scripts*. R package version 1.0.
URL: *https://CRAN.R-project.org/package=tictoc*

Mollin, Sandra. 2008. "The Hansard hazard: Gauging the Accuracy of British Parliamentary Transcripts." *Corpora* 2(2):187–210.

Mueller, Hannes and Christopher Rauh. 2018. "Reading Between the Lines: Prediction of Political Violence Using Newspaper Text." *American Political Science Review* 112(2):358–375.

Ooms, Jeroen. 2017. *hunspell: High-Performance Stemmer, Tokenizer, and Spell Checker*. R package version 2.9.
URL: *https://CRAN.R-project.org/package=hunspell*

Ooms, Jeroen. 2018*a*. *pdftools: Text Extraction, Rendering and Converting of PDF Documents*. R package version 1.8.
URL: *https://CRAN.R-project.org/package=pdftools*

Ooms, Jeroen. 2018*b*. *tesseract: Open Source OCR Engine*. R package version 2.3.
URL: *https://CRAN.R-project.org/package=tesseract*

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1):120–128.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
URL: *https://www.R-project.org/*

Rasiah, Parameswary. 2010. "A Framework for the Systematic Analysis of Evasion in Parliamentary Discourse." *Journal of Pragmatics* 42:664–680.

Rheault, Ludovic and Christopher Cochrane. 2018. Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora. In *PolMeth 2018*. Provo, UT: Society for Political Methodology.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2018*a*. *stm: R Package for Structural Topic Models*. R package version 1.3.3.
URL: *http://www.structuraltopicmodel.com*

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2018*b*. "stm: R Package for Structural Topic Models." *Journal of the Statistical Software* .

Roberts, Margaret E., Brandon M. Stewart and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Silge, Julia and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1(3).
URL: *http://dx.doi.org/10.21105/joss.00037*

Stan Development Team. 2018. "RStan: the R interface to Stan.". R package version 2.17.4.
URL: *http://mc-stan.org/*

Steyvers, Mark and Tom Griffiths. 2006. Probabilistic Topic Models. In *Latent Semantic Analysis: A Road to Meaning*, ed. T. Landauer, D McNamara, S. Dennis and W. Kintsch.

Taddy, Matt. 2015. "Distributed multinomial regression." *The Annals of Applied Statistics* 9(3):1394–1414.

Van Zanden, Jan Luiten, Eltjo Buringh and Maarten Bosker. 2012. "The Rise and Decline of European Parliaments, 1188–1789." *The Economic History Review* 65(3):835–861.

Vaughan, Davis and Matt Dancho. 2018. *furrr: Apply Mapping Functions in Parallel using Futures*. R package version 0.1.0.9002.
URL: *https://github.com/DavisVaughan/furrr*

Wickham, Hadley. 2017. *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.
URL: *https://CRAN.R-project.org/package=tidyverse*