The Effect of Events on Discussion in the Australian Federal Parliament (1901–2017)

Monica Alexander, University of Toronto

Rohan Alexander, Australian National University

24 October 2018

Presentation at SPIR, ANU.

Summary

Approach

- Create a dataset of what was said in the Australian Federal Parliament from 1901 through to 2017 based on available public records.
- Use a correlated topic model to reduce dimemsionality.
- Analyse the effect of various events using a Bayesian hierarchical Dirichlet model.

Findings

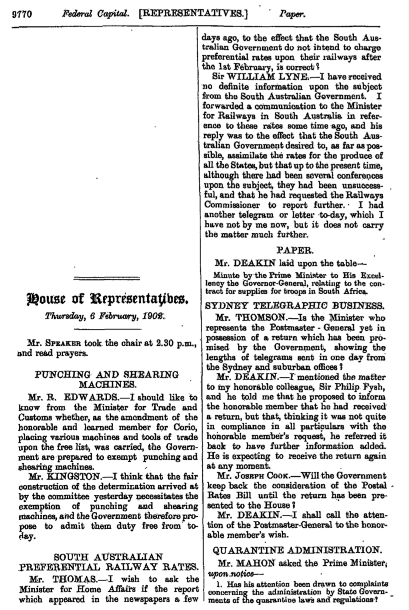
- Changes in government tend be associated with topic changes even when the party in power does not change;
- Elections that do not result in a change in government are rarely associated with topic changes;
- Economic events, such as financial crises, have less significant effects than other events such as terrorist attacks;
- and the effect of events is much more pronounced in the second half of our sample, and especially in the past two decades.

Data

Hansard

- A daily text record called Hansard of what was said in the Australian Federal Parliament has been made available since it was established in 1901. It's not vertatim, but it's pretty close.
- Hansard records have been used in other counties but not really at scale in Australia.
- Daily PDFs are available Hansard records available online as PDFs and these are considered the official release.
- There are 14,551 days worth of publicly available Hansard records across both chambers of the Australian Federal Parliament.

Example PDF page



9770 *Federal Capital.* [REPRESENTATIVES.] *Paper.*

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?

PDF parsing

- We use scripts written in R to convert the PDFs into daily text records.
- 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 e.g. ‘the’ is often parsed as ‘thc’.
- We: remove numbers and punctuation; change the words to lower case; and concatenate multi-word names titles and phrases, such as new zealand to new_zealand. Then the sentences are de-constructed and each word considered individually.

Model

Latent Dirichlet Allocation - Overview

- The key assumption behind LDA is that each 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 - this is an unsupervised machine learning method.

Latent Dirichlet Allocation - Example

Add figure Statistically, LDA considers each document as having been generated by some probability distribution over topics. Similarly, each topic is considered a probability distribution over terms. To choose the terms used in each document, terms are picked from each topic in the appropriate proportion.

Latent Dirichlet Allocation - Data generation process

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 hyper-parameter $\boldsymbol{\eta}$ is used: $\beta_k \sim \text{Dirichlet}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$. In practice, a symmetric Dirichlet distribution is usually used, where all elements of $\boldsymbol{\eta}$ are equal.
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

Latent Dirichlet Allocation - Data generation process

- 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. In a sense we are trying to reverse the document generation process – we have the terms and we would like to discover the topics.

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. The outcomes of the LDA process are probability distributions and these define the topics. Each term will be given a probability

Correlated Topic Model

Slight modification of LDA - rather than assuming that the distribution of topics in a document, θ_d, are a draw from a Dirichlet distribution, as in step 2 in LDA above, CTM assumes

$$\theta_d \sim \text{Logistic Normal}(\mu, \Sigma)$$

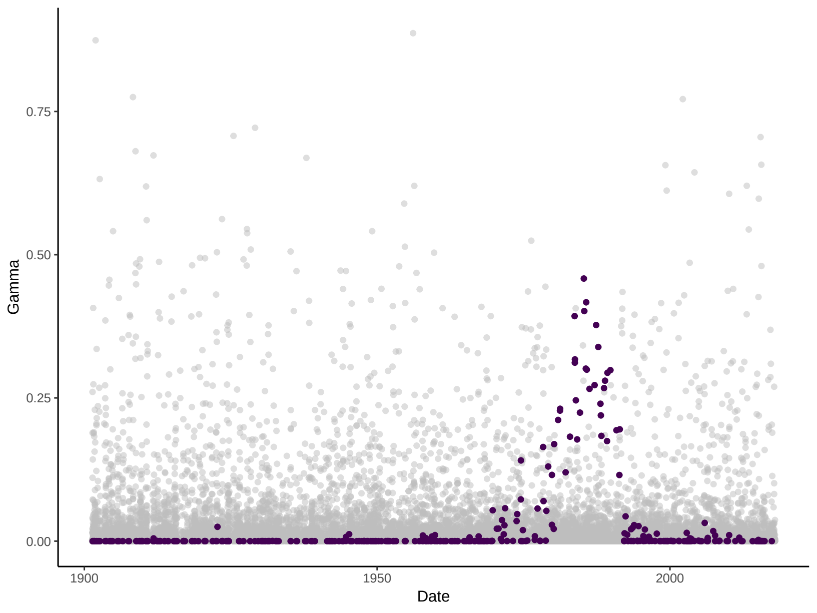
Essentially it swaps the Dirichlet for the Logistic Multivariate Normal. This sounds easy, but it's hard to implement.

The Structural Topic Model approach then adds a covariate to μ which allows consideration of additional information.

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

Again, sounds easy, but hard to implement and they implement a very nice algorithm.

Correlated Topic Model - Example output



Why not STM?

- No way to specify more complicated auto-correlated functional forms of the effects of events over time.
- There is no way to implement partial pooling across groups of similar documents.
- There is no way of identifying 'outlying' topic distributions – and therefore events that had an important effect – without pre-specifying the event of interest in the model.

Analysis model

Specifically, we use the estimated topic distributions from the CTM described in the previous section as an input into a Bayesian hierarchical Dirichlet regression framework, which relates the proportions of each topic to underlying time trends, changes in governments and elections. This set-up also allows us to identify 'outlying' topic distributions and relate these to other events.

Define θ_{dp} to be the proportion of topic of topic p on day d. Note that the $\theta_{d,1:P}$ for $p = 1, 2, \dots P = 40$ are equal to the estimated values of θ_d from the CTM. We assume that the majority of variation in topics is across sitting periods s, where a sitting period is defined as any group of days that are less than one week apart. Using this definition, there are a total of 745 sitting periods over the period 1901 to 2017 inclusive

Results

Future work

Future work

- Group identity: which is more important?
-

Questions?

- `rohan.alexander@anu.edu.au`
- @RohanAlexander
- rohanalexander.com

Typography

```
The theme provides sensible defaults to
\emph{emphasize} text, \alert{accent} parts
or show \textbf{bold} results.

In Markdown, you can also use _emphasize_ and **bold**
```

becomes

The theme provides sensible defaults to *emphasize* text, accent parts or show **bold** results.

In Markdown, you can also use *emphasize* and **bold**.

Math

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$$

R Figure Example

The following code generates the plot on the next slide (taken from `help(bxp)` and modified slightly):

```r
library(stats)
set.seed(753)
bx.p <- boxplot(split(rt(100, 4),
                      gl(5, 20)), plot=FALSE)
bxp(bx.p, notch = FALSE, boxfill = "lightblue",
    frame = FALSE, outl = TRUE,
    main = "Example from help(bxp)")
```

R Table Example

A simple `knitr::kable` example:

```r
knitr::kable(mtcars[1:5, 1:8],
             caption="(Parts of) the mtcars dataset")
```

Table 1: (Parts of) the mtcars dataset

	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02
Datsun 710	22.8	4	108	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02

Resources

For more information:

- See the Metropolis repository for more on Metropolis
- See the RMarkdown repository for more on RMarkdown
- See the binb repository for more on binb
- See the binb vignettes for more examples.