A Topic analysis

A.1 Overview

Each statement in Hansard needs to be classified by its topic. Sometimes Hansard 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. One way to get consistent estimates of the topics of each statement in Hansard is to use the latent Dirichlet allocation (LDA) method of Blei, Ng, and Jordan (2003), as implemented by the R package 'topicmodels' by Grün and Hornik (2011).

The key assumption behind the LDA method is that each statement, 'a document', in Hansard is made by a speaker who decides the topics they would like to talk about in that document, and then chooses 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. The topics are not specified *ex ante*; they are an outcome of the method. 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 Hansard group themselves to define topics.

A.2 Document generation process

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. 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 1).

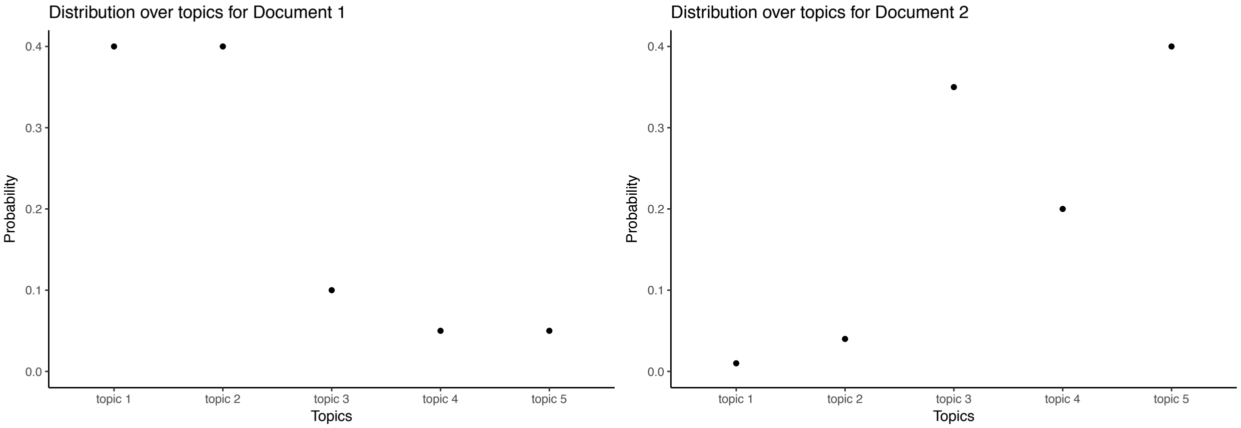

Figure 1: Probability distributions over topics

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. 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 2).

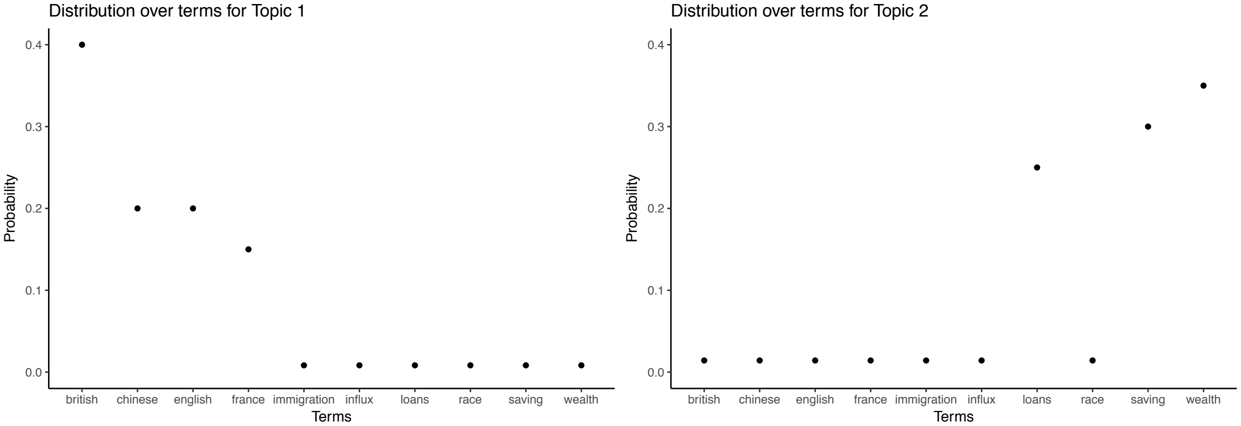

Figure 2: Probability distributions over terms

Following Blei and Lafferty (2009), Blei (2012) and T. 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: $\beta_k \sim \text{Dirichlet}(\eta)$.[1] 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, $z_{d,n} \sim \text{Multinomial}(\theta_d)$.

[1] 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.

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 in the next section, 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 (T. Griffiths and Steyvers (2004) and Blei (2012)).

A.3 Analysis 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. That is, we do not know the probability distributions of Figures 1 or 2. 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 (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):

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

The initial practical step when implementing LDA given a corpus of documents is to remove 'stop words'. These are words that are common, but that don't typically help to define topics. There is a general list of stop words such as: "a"; "a's"; "able"; "about"; "above"… An additional list of words that are commonly found in Hansard, but likely don't help define topics is added to the general list. These additions include words such as: "act"; "amendment"; "amount"; "australia"; "australian"; "bill"… A full list can be found in Appendix B. We also remove punctuation and capitalisation. The documents need to then be transformed into a document-term-matrix. This is essentially a table with a column of the number of times each term appears in each document.

After the dataset is ready, the R package 'topicmodels' by Grün and Hornik (2011) can be used to implement LDA and approximate the posterior. It does this using Gibbs sampling or the variational expectation-maximization algorithm. 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 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.

A.4 Warnings and extensions

The choice of the number of topics, k, affects 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 a test and training set process. Essentially, this means running the process on a variety of possible values for k and then picking an appropriate value that performs well.

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.

B Hansard stop word

The following words were added to the usual list of stop words: “1”, “2”, “act”, “amendment”, “amount”, “australia”, “australian”, “bill”, “board”, “cent”, “clause”, “commission”, “committee”, “commonwealth”, “countries”, “country”, “day”, “deal”, “debate”, “department”, “desire”, “duty”, “gentleman”, “government”, “honorable”, “honourable”, “house”, “increase”, “labor”, “labour”, “leader”, “legislation”, “matter”, “minister”, “money”, “national”, “opposition”, “parliament”, “party”, “people”, “policy”, “position”, “power”, “prime”, “proposed”, “public”, “question”, “regard”, “report”, “service”, “situation”, “south”, “speaker”, “statement”, “support”, “system”, “time”, “united”, “vote”, and “wales”.

References

Beelen, Kaspar, Timothy Alberdingk Thim, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, et al. 2017. "Digitization of the Canadian Parliamentary Debates." *Canadian Journal of Political Science*, 1–16.

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

Blei, David M, and John D Lafferty. 2009. "Topic Models." In *Text Mining*, 101–24. Chapman; Hall/CRC.

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

Butlin, N. G., A Barnard, and J. J. Pincus. 1982. *Government and Capitalism: Public and Private Choice in Twentieth Century Australia.* George Allen & Unwin.

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*, July.

Darling, William M. 2011. "A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling." In.

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

Edwards, Cecilia. 2016. "The Political Consequences of Hansard Editorial Policies: The Case for Greater Transparency." *Australasian Parliamentary Review* 31 (2): 145–60.

Gans, Joshua, and Andrew Leigh. 2012. "How Partisan Is the Press? Multiple Measures of Media Slant." *The Economic Record* 88 (280): 127–47.

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

Griffiths, Thomas, and Mark Steyvers. 2004. "Finding Scientific Topics." *PNAS* 101: 5228–35.

Grijzenhout, Steven, Maarten Marx, and Valentin Jijkoun. 2014. "Sentiment Analysis in Parliamentary Proceedings." In *From Text to Political Positions: Text Analysis Across Disciplines*, edited by Bertie Kaal, Isa Maks, and Annemarie van Elfrinkhof, 117–34. John Benjamins Publishing Company.

Grün, Bettina, and Kurt Hornik. 2011. "Topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software, Articles* 40 (13): 1–30. doi:10.18637/jss.v040.i13.

Marx, Maarten, and Anne Schuth. 2010. "DutchParl: A Corpus of Parliamentary Documents in Dutch." In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (Lrec'10)*, edited by N Calzolari, K Choukri, B Maegaard, J Mariani, J Odijk, S Piperidis, M Rosner, and D Tapias, 3670–7. European Language Resources Association (ELRA).

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

Rasiah, Parameswary. 2010. "A Framework for the Systematic Analysis of Evasion in Parliamentary Discourse." *Journal of Pragmatics* 42: 664–80.

Sealey, Alison, and Stephen Bates. 2016. "Prime Ministerial Self-Reported Actions in Prime Minister's Questions 1979–2010: A Corpus-Assisted Analysis." *Journal of Pragmatics* 104: 18–31.

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

Whyte, Tanya. 2014. "Some Honourable Members: A Quantitative Analysis of Parliamentary Decorum in Canada and the United Kingdom." *Paper Presented to the 23rd World Congress of Political Science, 23 July 2014, Montreal Quebec, Canada.*

———. 2017. "Oh, Oh! Modeling Parliamentary Interruptions in Canada, 1926-2015." *Paper Presented at Canadian Political Science Association Annual Conference, Ryerson University, Toronto, 27 May – 2 June, 2017.*

Wilkerson, John, David Smith, and Nick Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." *American Journal of Political Science* 59 (4): 943–56.

Willis, Rebecca. 2017. "Taming the Climate? Corpus Analysis of Politicians' Speech on Climate Change." *Environmental Politics* 26 (2): 212–31.

Wilson, John K. 2015. "Government and the Evolution of Public Policy." In *The Cambridge Economic History of Australia*, edited by Simon Wille and Glenn Withers, 330–50. Cambridge University Press.