EXPLORING AUSTRALIA'S HANSARD (1901-2017)
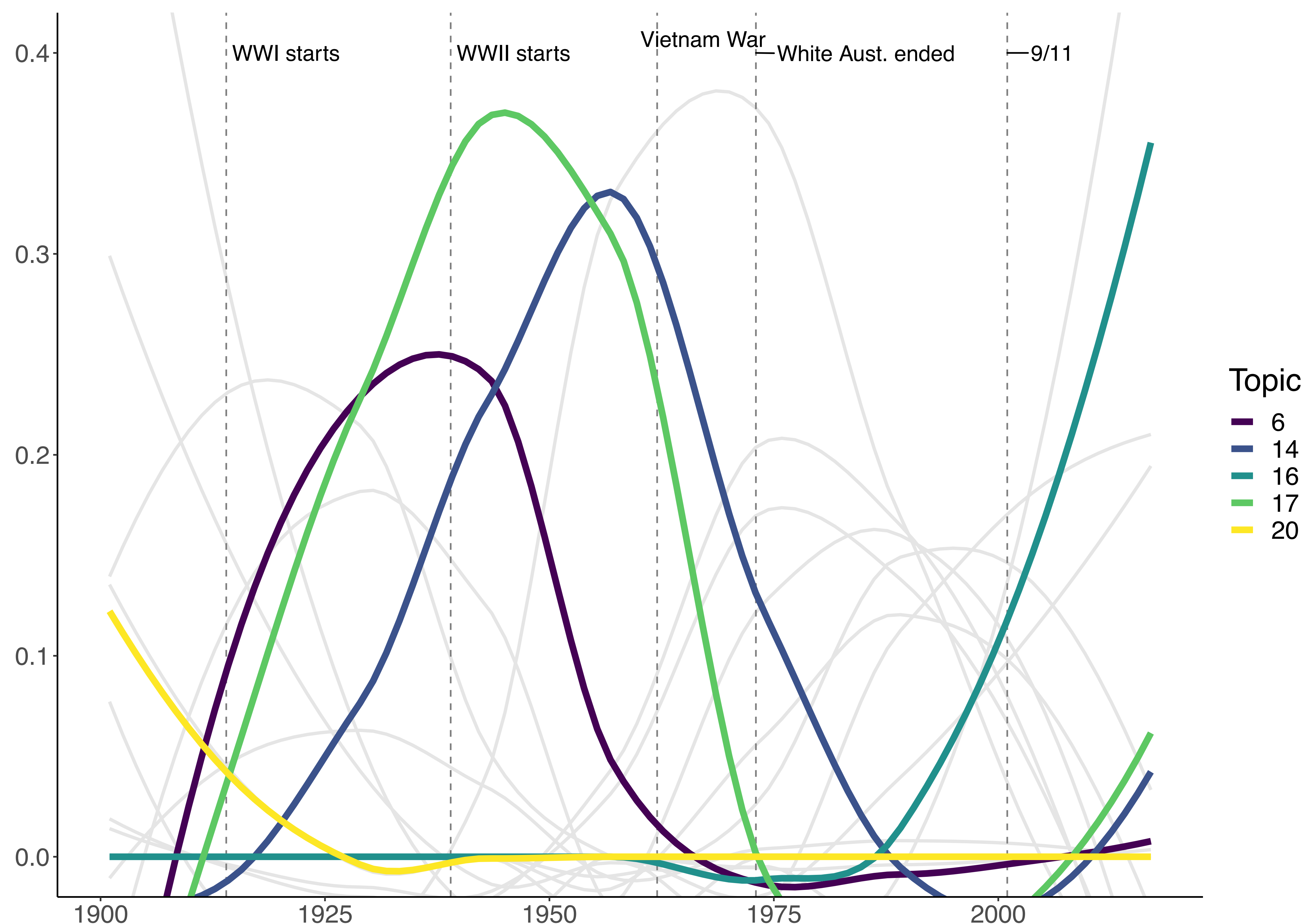
Rohan Alexander

Overview

I look at what was said in Australia's House of Representatives and Senate between 1901 and 2017. Using natural language processing I attribute a topic to each Hansard statement. This creates a dataset of the topics discussed over time. I then associate topics with various economic and social factors. If a parliament in a liberal democracy is essentially just a 'talk shop' with little actual impact, then research should focus on other institutions. Or if parliament plays an important role, then further analysis of how these decisions are made is needed. Either way, there are implications for our understanding of the institutions that directly control about a quarter of GDP and more broadly have a unique and extensive role in all aspects of life.

Data

The Australian Parliament House makes each day's Hansard available online as a PDF. I parsed these PDFs to convert them to text-based CSV files. The data are messy, but there are about 3 million statements across about 15,000 house-days. Numbers, punctuation and common but less-informative words (such as 'Commonwealth') were removed. and each word was stemmed, which allows, say, words and their plurals to be made indistinct. Topic modelling requires counts of each word by the analysis unit, in this case, speaker-days.

Topic modelling

The latent Dirichlet allocation method (Blei, et al., 2003), as implemented by the `stm` R package (Roberts, et al., 2018), provides consistent estimates of the topics of each statement. The key assumption behind LDA is that each document is made by a speaker who decides the topics they would like to talk about in that document, and then chooses 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. The Hansard words group themselves to define topics.

Next steps

Each day tends to be dominated by a specific topic and these change in response to events. It has been difficult to find factors that are led by Hansard.

Acknowledgements

Sincere thanks to my supervisor, John Tang, and to my panel: Zach Ward, Tim Hatton and Martine Mariotti. Thank you to Monica Alexander for her help in all aspects of this research.

✉ rohan.alexander@anu.edu.au
🐦 @RohanAlexander
github.com/RohanAlexander