# Lexically Healthy: Using Text to Predict Health Bill Outcomes

Michelle Manning

UMass Amherst
Data Analytics and Computational Social Science Program

## Context & Goal

The 116th Congress introduced 909 health bills in the House, but only a fraction of a percent became law. Understanding the lexical topics found within passing bills and the those that don't will enable lawmakers to more effectively create legislation that will be successful & improve healthcare.
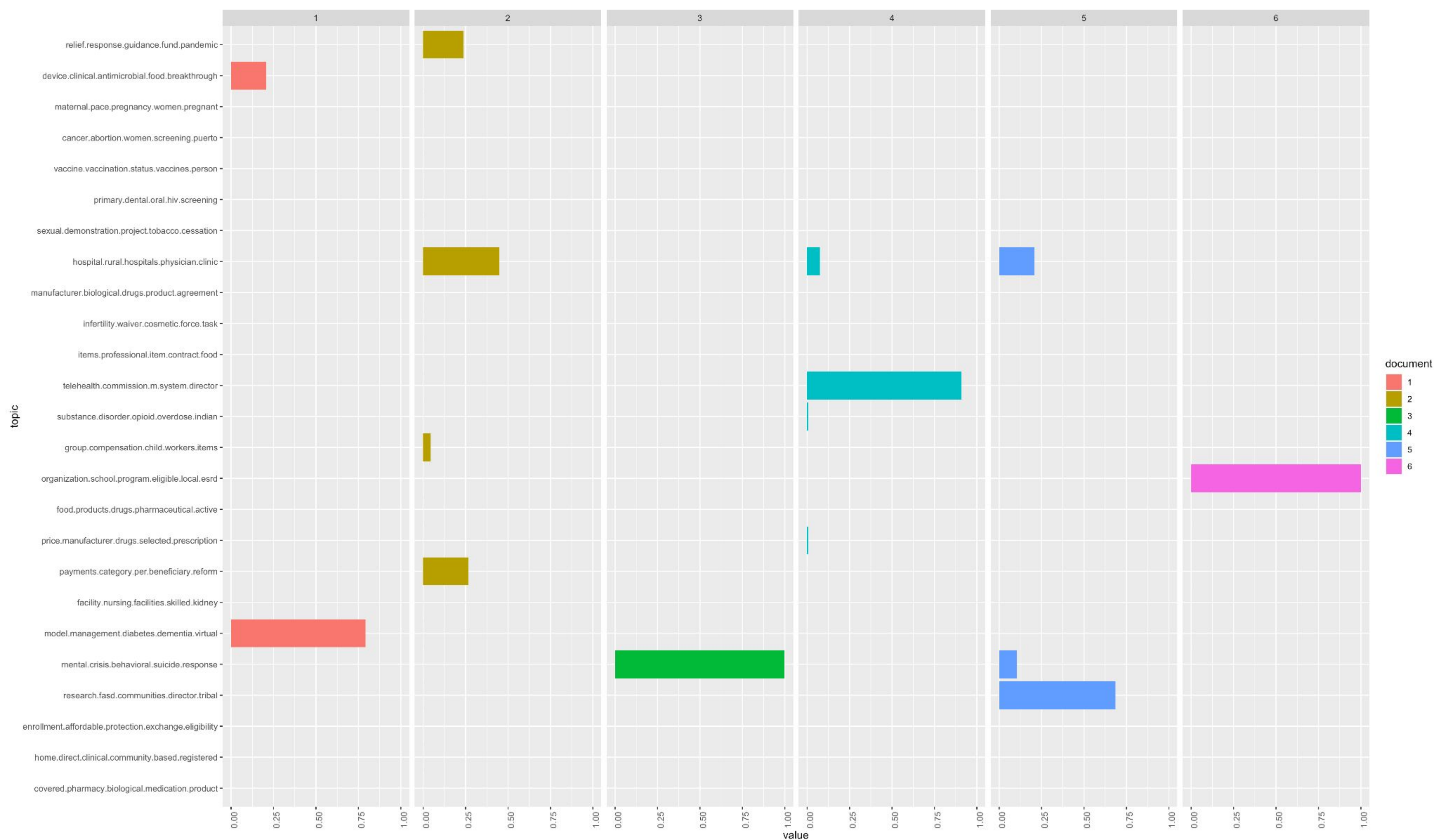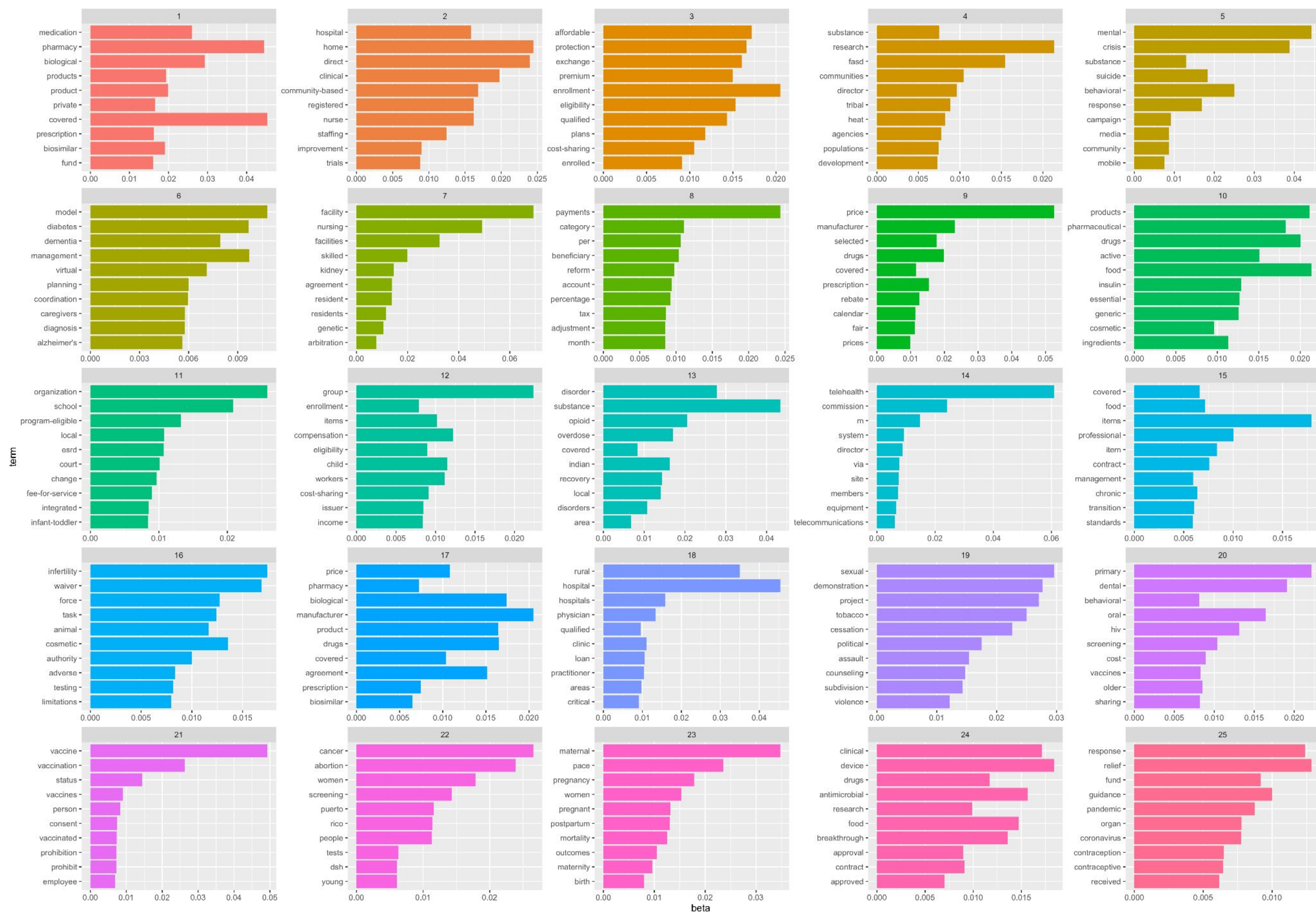
## Preprocessing & Model

First step in preprocessing, I created a corpus and tokenized.
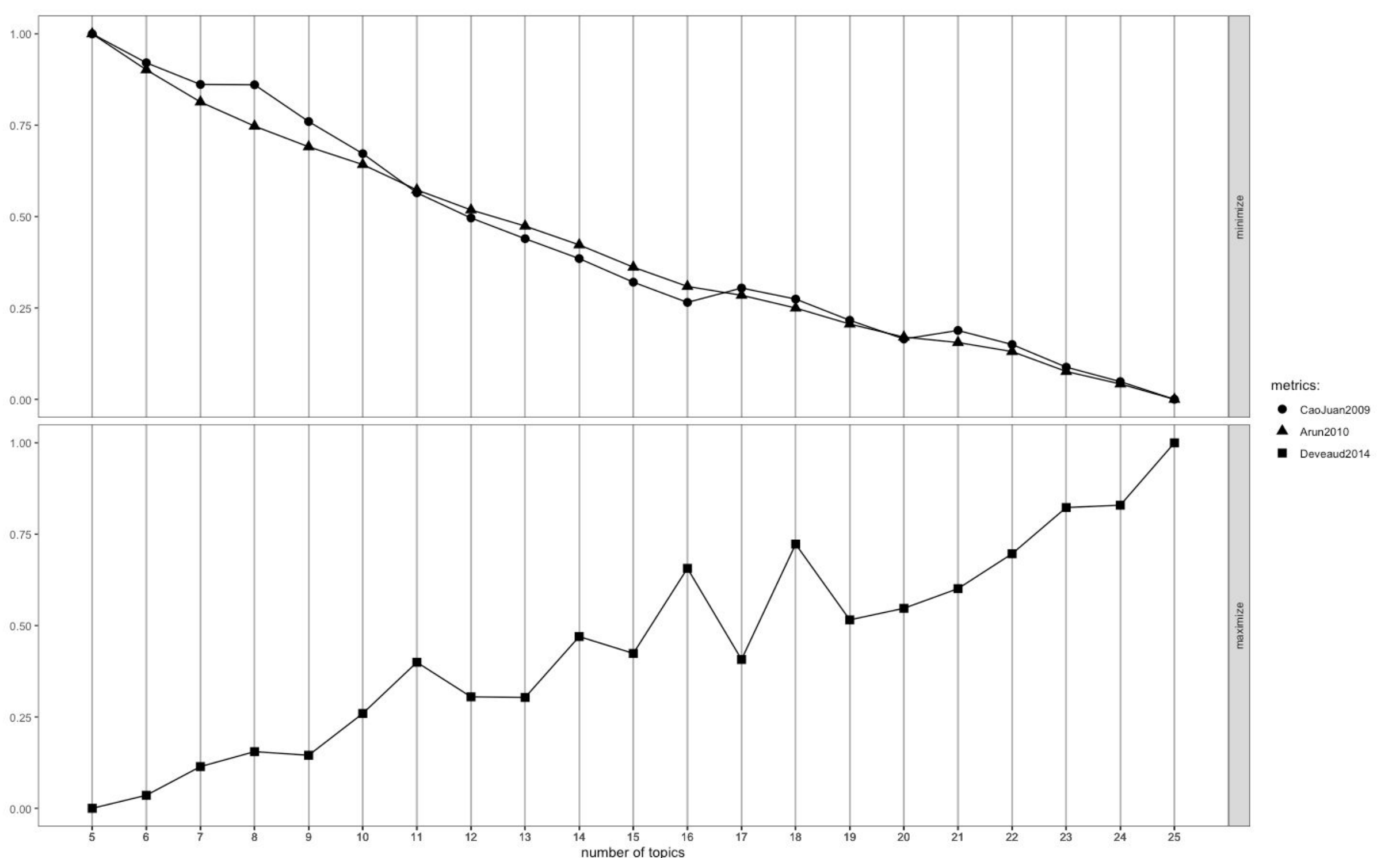
Tokenization:

1. Removed punctuation, numbers, symbols, separators, stopwords = "en", & urls
2. Lowered the case
3. Removed homebrewed stopwords = "typesnum" focused on tokens that were numbers & letters.
4. Vectorizor: Vocabulary

Next, I split the data and created training & test DTMs. Finally, I found $k$ by comparing standard measures from previous studies.

After this step, I fitted the LDA model to the training set and ran it on the test data set. I validated the results with a perplexity test.

## Results



## Finding $k$



## Conclusions

This LDA model was able to parse out topics present in House health bill texts along with the percentages of each topic in all of the bills.

## Future Research

This model focuses on the lexical patterns that affect a bill passing. Other factors could influence this are the party in power, the bill sponsor's party, or the year it was introduced. Incorporating these factors into the model could make it more accurate predict bill outcomes.

### References

- Murzintcev, Nikita. "Select Number of Topics for LDA Model." *CRAN R-Project*, R, 20 Apr. 2020, https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html.
- *Text Mining of Social Media Data Using R.* https://www.cspoerlein.com/files/textanalyse.html#topic_models_and_pos_tagging. Accessed 3 May 2022.
- *Topic Modeling with R.* https://slcladal.github.io/topicmodels.html#Topic_distributions. Accessed 3 May 2022.

Learn more about me