POLI SCI 490: Machine Learning & Text-as-Data

HW4

Sarah B. Bouchat

Due 28 February

Note: Topic models can take some time to fit, so do not be alarmed if your computer is "thinking" for a while as you work on these problems.

1 LDA

For this problem, use the repository of NSF abstracts from 2000–2003, located here.

1. Import the data (*hint*: use DirSource), pre-process, and set up a DTM.
2. Use LDA to assess topics in these abstracts, first with 5 topics, then with 10.
3. Report these topics in both table and visual formats.
4. Compare your results: how does the 10-topic model differ from the 5-topic model?

2 Structural Topic Models

For this problem, use the data on TED talks, which are posted on Canvas.

1. Import the data, pre-process the transcripts, and create a DTM.
2. Set up the DTM to be correctly formatted for use with the `stm` package. Use the documentation for the package to assist with this.
3. Use `stm` to fit a structural topic model with 9 topics, conditioning on `ted_type`, the venue of each of these TED talks.
4. Label the topics with `labelTopics`.
5. Using the originally cleaned and pre-processed data, fit a standard "vanilla" LDA model with 9 topics.
6. Compare your results. How do the topics you find when conditioning on venue differ from those you found using standard LDA?