Assignment 2

Load and Prepare the data set

1. Load the `corpus_us_debate_speaker` from the data folder on the course's github repository main site, reshape the corpus into paragraphs and select only those paragraphs with at least 8 words in them. Create a tokens object in which words are converted to lower case and remove numbers, punctuation, common stop words, and tokens with less than two characters. Convert the tokens object to a dfm and trim it to include only tokens that appear at least in 7.5% of the documents and at most in 90% of the documents.

2. Use the `convert()` function from `quanteda` to convert the dfm to an stm.

Identify optimal number of topics

3. Fit five different structural topic models with different numbers of topics for K = 10, 20, 30, 40, 50 respectively. Note that this might take a while to run and converge.

4. Create a diagnostic plot showing the held-out likelihood, the residuals, the semantic coherence of the topics, and the lower bound. Also create a plot to contrast the semantic coherence with the exclusivity of the topics, i.e. how much each topic has its own vocabulary not used by other topics. Explain your results and decide on an optimal number of topics to continue with.

Create topic labels and explore topic prevalence

5. After you selected your preferred topic model, explore the word probabilities (β) and create meaningful labels for each topic. For some topics, identify documents that are very representative for a those particular topics. Discuss what some of your topics are about in a bit more detail.

6. Extract the topic proportions (θ) from the model and plot the prevalence of each topic across the overall corpus. Also create a perspective plot visualizing the combination of two topics and discuss the results.

Fit an stm with covariates

7. Refit your favorite stm model but this time include a covariant for party affiliation into the model (i.e. saying that party affiliation impacts topic prevalence). Once the model converged, estimate the effect that party affiliation has on the prevalence of two topics, i.e. are Democrats or Republicans more likely to speak about either of the topics. Create a plot of your estimated effects and discuss your results.

Summarize

8. Summarize your findings across all tasks in a paragraph or two.

Submission form and deadline

- Deadline: 30 November
- Submission form: submit your code, plots, and discussion of the results in a single RMarkdown file.