STCS 6701: Foundations of Graphical Models: Reading 13

Maxime TCHIBOZO (MT3390)

December 2020

1 stm: An R Package for Structural Topic Models - Roberts et al. (2019)

The Strutural Topic Model extends established topic models such as Correlated Topic Models (CTM) and LDA by modeling dependence on exogenous variables associated to the metadata. This allows researchers to measure the effect of a variable on the topics within the document. For example, questions such as: "do Democrats use a different vocabulary than Republicans" could be easily answered in an STM framework.

This paper details the inner workings of the R stm package, which offers STM functionalities in R. The authors of this paper detail each step of how one might use this package for a personal project – while respecting best topic modeling practices. They use the CMU Poliblog dataset to evaluate how political ideology and other metadata covariates affect topic prevalence and topic distribution.

The key functions of this package are:

- stm(): trains an stm model using variational inference or collapsed Gibbs sampling.

  Topic quality (semantic coherence, topic exclusivity, FREX) and prevalence diagnostics for this model can be displayed with the plotModels function.

- selectModel(): runs several identical models with different initializations and retains the best one.

- searchK(): Proposes different optimization methods such as grid-searching and low-dimensional projections to set the optimal number of topics parameter k.

- interpretation plots: labelTopics(), sageLabels(), plot.STM(), findThoughts(), plotQuote(), topicCorr().
- estimateEffect(): STM-specific function to estimate relationships between the metadata and topics in the data.

The estimateEffect functionality is particularaly useful to the analysis section of my final project, and the variety of model checking graphs this package offers will help facilitate iterating through Box's loop.

Along with information about these functions, the authors share advice on how to best perform data exploration with the package. For example, users should first look at the collections of words contained in each topic, and then view which documents best represent each topic.

Given all the powerful functionalities it offers, this package is an underrated gem in my opinion.