TIME DYNAMIC TOPIC MODELS

PATRICK JÄHNICHEN

The unsupervised extraction of information from large text corpora has become the basis for research in a wide range of humanities and social sciences. For example, media and political communication science can benefit greatly from insights into online media or digitized text collections as a quantitative method to complement classical qualitative analysis. To support this approach, we give a linguistically motivated interpretation of topic modeling, a state-of-the-art analysis method for extracting latent semantic clusters of words, i.e. words that capture themes that pervade a corpus. We further extend this interpretation to cover issues and issue-cycles as theoretical constructs coming from political communication science. In particular, we build on a dynamic extension of topic modeling, an approach that allows latent word clusters to evolve in time, allowing the identification of themes in the corpus and at the same time observe their change through time. Initially, this change is modeled by a simple stochastic process, Brownian motion. We provide a novel method for analyzing a topic's evolution that is based on the notion of volatility as in the rate of change of stock prices or derivatives, a common model in econometrics. We claim that the rate of change of sets of semantically related words can be interpreted as issue-cycles, the word sets as a description for underlying issue. Further, we propose to extend the existing dynamic topic model by generalizing it to be driven by general Gaussian processes, of which Brownian motion is one special case. Gaussian processes are a family of stochastic processes that are sufficiently defined by functions that model their covariance structure. By applying a certain class of covariance functions, we allow the semantic word sets to rapidly change when necessary but to preserve semantic relatedness among words in between these changes. Applying our findings to a large newspaper data set, the New York Times Annotated corpus (all articles between 1987 and 2007), we are able to identify sub-topics in confined time frames, *time-localized topics*, and find patterns in their behavior over time. Our approach includes a new perspective on topics as generated by dynamic models. We drop the assumption of semantic relatedness over all available time for any one topic. Time-localized topics are consistent in themselves but do not necessarily share semantic meaning between each other. They can, however, be interpreted to capture the notion of issues and their behavior that of issue-cycles.