Mixture Models in Text Mining— Tools in R

Bettina Grün
Institut für Angewandte Statistik
Johannes Kepler Universität Linz

The nowadays ready availability of large electronic document collections requires automatic statistical tools to learn meaningful information from natural language text. Often these statistical methods are bag-of-words models where the information how often terms occur in a document is assumed to be sufficient and the order in which words occur is not taken into account.

Among these bag-of-words models two different mixture models have been proposed: finite mixtures of von Mises-Fisher distributions and the latent Dirichlet allocation topic model. Finite mixtures of von Mises-Fisher distributions are fitted based on the assumptions that each document belongs to only one cluster and that only the directional information in the data is of importance. In the latent Dirichlet allocation topic model the words in each document are assumed to be from a mixture of topics and each topic is characterized by its own term distribution. Each document belongs to several topics simultaneously and the topic distribution is allowed to vary over documents.

In this talk both models are introduced and their estimation outlined. In addition an overview on the R packages **movMF** and **topicmodels** is given which allow to fit these models. Both packages build on and extend functionality from the text mining package **tm**. The available functionality in the packages is presented and the application illustrated.