Machine Learning 2
Exercise Sheet 11

Prof. Dr. Dr. Lars Schmidt-Thieme, Nicolas Schilling
Information Systems and Machine Learning Lab
University of Hildesheim

June 28th, 2016

Submission until July 5th, 18.00 to schilling@ismll.de

Exercise 18: Latent Dirichlet Allocation (10 + 5* Points)

a) Check out the data set given in *bbc_small.zip*. This data set consists of documents which are mapped to four different topics, namely *business*, *football*, *politics* and *tech*. Convert the whole data set to a dictionary where you map each word to a unique id. Use this to then create word count data of each document.

b) Run LDA on the word count data with $K = 4$. Implement it yourself using Gibbs Sampling for the bonus points! For each topic, what are the ten most likely words to appear?