Exercise Sheet 5

Due Date: February 15, 8 pm

Note on Submission

All solutions have to be uploaded together as a single zip file to LernraumPlus. Solve the exercises by implementing the functions in the file `exercise_sheet5.py`.

Exercise 1 – Preprocessing [5 points]

The provided corpus contains documents of 20 different topics (one topic per folder). Preprocess the corpus by doing the following steps:

a) Remove all meta information from each document in the corpus

b) Install NLTK (www.nltk.org)

c) Choose an appropriate tokenizer from NLTK and convert each document into a list of tokens

d) Remove all stopwords

e) Randomly assign a topic to each word for all documents (20 different topics in total)

f) Add some instance variables to the class `LdaModel` for the results of the preprocessing steps

Exercise 2 – LDA Gibbs sampling [15 points]

Implement the LDA Gibbs sampling algorithm (`gibbs_sampling`). Use $\alpha_i = 0.25$ and $\beta_i = 0.1$ for all words.