Computational Linguistics

Assignment 5 (2021-01-15)

Winter Semester 2020/21 – Prof. Dr. Alexander Koller

Latent Dirichlet Allocation

In this final assignment, you will implement LDA and try it out on a corpus. Be sure to start work on this assignment early, because running your code will take some time. A full run of my implementation took about four hours. Consider running your code on smaller subcorpora during debugging.

Implement a Gibbs sampler which resamples a topic for each word in the corpus according to the probability distribution in formula [5] of Griffiths & Steyvers (2004). Initialize the topic assignments by choosing a topic uniformly at random for each token in the corpus.

Try your Gibbs sampler on the corpus of 2000 movie reviews from Pang & Lee (2004), available on Moodle. The first line of the file specifies the number of documents. Then each subsequent line is one document, with the tokens separated by whitespace. I generated the file on Moodle from the original movie reviews by tokenizing them and then removing punctuation and stopwords, so you get nicer topics.

Try out different numbers of topics and iterations and different values for the hyperparameters. You should get good results with 20 topics, 500 iterations over the corpus, $\alpha = 0.02$, and $\beta = 0.1$. Print, for each topic, the most frequent words for that topic in the final sample, and discuss to which extent the topics actually represent thematically coherent semantic fields.

Extra credit: Here are some ideas: Use Numpy to speed up the calculations; try your system on other corpora; experiment with different strategies for initializing the topic assignments (e.g., give the same topic to all words of the same document); use a bit of supervised information (e.g., fix the topic assignments for specific words that you know should belong to that topic, and do not allow the sampler to change them).

Turn in before class on 2021-01-29 on Moodle.