SELDA
– Scalable Efficient Latent Dirichlet Allocation –
Software Project by Annika Berger, Stephan Kienzle, Neri Kranz, Jan Pawellek
WS 2010/2011
Institut für Computerlinguistik, Uni Heidelberg

Abstract
SELDA provides a scalable (i.e. parallelized) implementation of the Latent Dirichlet Allocation (LDA). LDA uses generative probabilistic models to perform unsupervised identification of hidden topics in documents, so that each document can be seen as a mixture over these topics. **SELDA thus offers unsupervised categorization of documents.**

Background
1. LDA
   • Assertion: Documents are mixtures of topics (e.g. sports, politics etc.)
   • LDA: Words of a document are generated by a topic probability model (but the actual topic distribution is latent)
   • Various methods for estimation of the model’s parameters, we implemented Gibbs Sampling

2. Inference (Gibbs Sampling)
   Randomly assign a topic to each word
   For each iteration:
   • For each word in each document:
     Update topic assignment probability (i.e. determine the most probable topic in regard to all other word-topic-assignments)
   • Various methods for estimation of the model’s parameters, we implemented Gibbs Sampling

3. Distributed Inference (Parallel LDA)
   • Parallelization using the Hadoop MapReduce framework
   • Map and Reduce phases are executed in parallel on a number of clusters
   • Map phase: Perform Gibbs Sampling on local data subset
   • Reduce phase: Update model and topic assignments

Preprocessor
• Preprocessing of a corpus
• Relevant text extraction, tokenization, lemmatization, punctuation and stop word deletion

Decoder
• Application of Gibbs Sampling to unseen documents using the parameters estimated by the Trainer
• Colored output of the most frequent topics

Evaluation
• Idea: Supervised classification should correspond to SELDA’s topic distribution
• 55 human annotated categories provided by Reuters
• Similarity detection between topic distributions using the Kullback-Leibler-Divergence as measure of similarity
• Average similarity (of SELDA’s topic distribution) is better the more (human annotated) categories are shared by the documents

References