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

$$P(z_i|w) = \frac{(N_{dz_i}+\alpha)*(N_{wz_i}+\beta)}{N_{z_i}+VocabularySize*\beta}$$

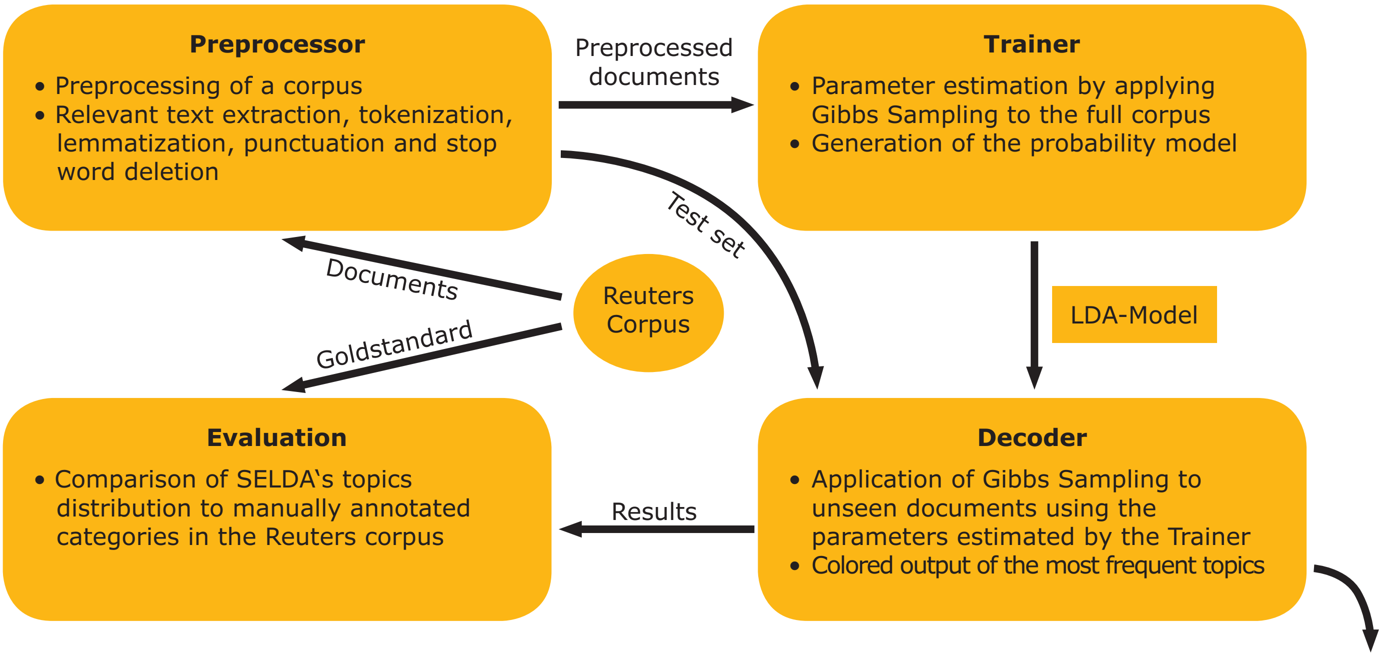
Gibbs Sampling

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)
```

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



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

Topic 5	Topic 12	Topic 19	Topic 20
0.46503	0.07343	0.05245	0.05245
government	service	european	rate
minister	company	council	bank
state	sprint	September	federal
plan	online	commission	inflation
service	internet	brussels	expect

cuban [5] party [5] crisis [5] weaken [1] support [12] ruling [5] party [5] Tuesday [5] economic [5] crisis [5] generate [1] increase [20] crime [5] society [19] weaken [12] popular [3] support [5] rule [5] public [5] support [5] party [5] previously [15] island [5] million [15] party [5] central [5] committee [19] political [5] analysis [5] economic [5] reform [5] introduce [5] counter [11] recession [15] trade [15] aid [19] create [5] social [5] turn [3] lead [5] support [5] social [19] group [1] person [5] crisis [5] party [5] effect [5] list [3] analysis [5] publish [9] official [9] newspaper [5] include [12] increase [20] crime [5] state [5] personal [5] property [13] commit [5] state [5] government [5] action [19] emerge [5] group [1] party [5] note [14] cuban [5] search [5] society [17] solution [5] leave [5] country [5] addition [16] economic [5] crisis [5] create [5] fear [20] search [12] personal [6] document [5] majority [5] cuban [5] worker [12] return [17] solution [5] problem [5] cuban [5] fully [14] understand [6] economic [5] reform [5] introduce [5] beat [20] crisis [5] manage [11] significant [16] number [12] cuban [5] provide [12] goods [15] service [5] reform [5] economic [20] add [20] view [5] leader [5] make [12] clear [2]

References

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research 3, 2003.
- Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. PNAS, 101(Suppl. 1), page 5228–5235, 2004.
- David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, pages 5:361–397, 2004.
- Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. 2002.
- Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. Plda: Parallel latent dirichlet allocation for large-scale applications. Proc. of 5th International Conference on Algorithmic Aspects in Information and Management, pages 301–314, 2009.