**Latent Dirichlet Allocation**

Description

Latent Dirichlet Allocation is a generative model. This means the aim of Latent Dirichlet Allocation is to find the model, which is assumed to have the generated the documents in the corpus. The aim of the algorithm is to find the model, which has the highest probability to have generated the corpus, i. e. the training data set. This is comparable to the maximum likelihood method. The generation works like in a Bayesian network, step after step, e.g. first generate the topic(s) of a document and from this topic(s) generate the words in the document. The generative process can be described as follows:

1. Choose
2. Choose θ
3. For each of the N words :
   1. Choose a topic
   2. Choose a word from , a multinomial probability conditioned on the topic .

Here θ is a k dimensional vector, where k is the number of topics. Θ is the probability for a certain topic in the multinomial distribution and is the resulting topic of this multinomial distribution. N is the number of words within a document and is the nth word. is a parameter for how likely a word n is generated if the topic is .

The joint distribution of a topic mixture θ, a set of N topics z, and a set of N words w is given by:

From this formula it is clear, how the total probability is generated. This can be seen as a chain of information: Given , we have probability for the multinomial parameter . Given this parameter, we get the probability for a topic . Given the topic and the probability for a word, if a topic is given , we get the probability for a word.

The formula above can be rewritten to the probability of a corpus, where just a fixed number of words shall be generated:

Since in (Blei et al 2003) it is stated that the calculation of the posterior as formulated above is computationally infeasible and that a feasible algorithm can just calculate an approximation, it is very likely that this is also true for the implemented algorithm of LDA in Gensim. This approximation is done via variational inference, where in short variational parameters are used instead of the proper ones. These parameters follow a simpler distribution assumption.

The parameter estimation is done via maximizing the log likelihood. Since the computation is not tractable, a lower bound for log likelihood is aimed at instead. The method used is called alternating variational EM procedure:

* E - Step: Find the optimizing values of the variational parameters.
* M - Step: Maximize the resulting lower bound on the log likelihood with respect to the model parameters.

These two steps are repeated until convergence is reached.

The goodness of fit used for the LDA model is the perplexity:

The perplexity is high if the word vector likelihood of a the document is high. Keep in mind that the perplexity is also an approximation, since the method LDA does not determine the optimal likelihood, but rather an approximation or lower bound.

Most relevant modifiable Parameters of training method in Gensim

* num\_topics (int): number of topics
* passes (int): number of passes over the corpus for the initial LdaModel
* random\_state (int) – use a seed integer for reproducability
* per\_word\_topics (Bool) – If true, a probability is calculated by which topic a word is generated

Expected Results

In (Chang et al. 2009) the Latent Dirichlet Allocation is found to be the most similar method to the way as humans determine the topic associations.

Literature

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288-296).