Latent Dirichlet Analysis (LDA) is a method used to model the generative process of creating discrete data such as text corpora. LDA can take a large amount of data and create descriptions of that data. This method reduces the size of the data to these short descriptions while still maintaining the relationships necessary to carry out various inferences. The algorithm assumes a generative process for the creation of documents in a corpus D using N words. The topic for a particular word, $z\_n$, is modeled as Multinomial($\theta$), where $\theta \sim Dir(\alpha)$. Then, each word, $w\_n$, is chosen from a multinomial probability that is conditioned on the topic $z\_n$, P($w\_n | z\_n, \beta)$. (Blei et al., pg. 996)

LDA is most well-known for its application in the analysis of text data. It is used to create topics for documents, classify documents based on these topics and to determine which documents are similar to one another. The algorithm can also be used in other problems that have a similar structure to the document generating method described above. For example, in Blei et.al, 2003, the authors used a data set where web site users provide information about movies they enjoy. In this example, the users are analogous to the documents and their preferred movies are “words”. Topics can be found by determining similar movies.

There are multiple alternative algorithms that can be used for text analysis. These include the term-frequency inverse-document-frequency matrix, latent semantic indexing (LSI) and the probabilistic latent semantic indexing (pLSI) model. In Section 7 of the paper, the authors present the results of document modeling and document classification. In this section, the authors’ results show that their implementation of LDA performs better than competing methods in terms of perplexity measure, where better generalization performance is defined by a smaller perplexity measure.

Perplexity measure is defined as:

\begin{center}

Exp(-\frac{\sum\_{d=1}^M log p(\textbf{w}\_d}{\sum\_{d=1}^M N\_d})

\end{center}

In addition, this section demonstrates that other methods are prone to overfitting. As the number of topic increases, some of the alternative algorithms induce words that have small probabilities. This occurs because the documents in the corpus are divided into more collections. This can result in the perplexity measure becoming very large for these alternative algorithms. The problem comes from the requirement that the topic proportions in a future document must be seen in at least one of the training documents. On the other hand, LDA does not have this overfitting problem. For more detail, see Section 7 of Blei et al.

A disadvantage of LDA is that exact inference of the posterior is impossible as a result of intractability. Thus, it is necessary for the user to implement an approximating technique such as variational inference, a Gibbs sampler, or another technique.

**References**

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research 3, 2003, pg. 993-1022.
2. Max Sklar, *Fast MLE Computation for the Dirichlet Multinomial*, May 2014.
3. F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872