# Generative Model For Latent Reasons For Modifications

**David Lassner**
davidlassner@mailbox.tu-berlin.de
TU Berlin, Germany

## Problem

The idea that writing makes its way from the authors first draft manuscript to the intended reader without any detours or modifications is often inaccurate and oversimplified. In general, the author or a close person performs corrections and stylistic modifications in subsequent iterations. Additionally, there may be an editor or even an official censor who perform censorship of too private or too extreme parts of the document. The different versions of a document generated by these correction layers often become intransparent in printed versions of the document, while manuscripts are more likely to display traces of how the document has been modified to its current state. The digital scholarly edition "Letters and texts. Intellectual Berlin around 1800" (Baillot, 2016, IB in the following) combines genetic edition and entity annotation. The corpus encompasses literary and scholarly testimonies by a group of people, who influenced the intellectual Berlin between Enlightenment and Romanticism. The genetic encoding gives precise information regarding deletions and additions in the manuscript text. However, the reason for these modifications is not encoded. Three main domains for reasons why to modify such a document as a letter in the intellectual context of the time around 1800 have been identified:

1. Correction of mistakes
2. Stylistic modification
3. Moral censorship based on the topic

This paper proposes an unsupervised machine learning approach, which assigns the according reason to every modification. The proposed method focuses on dealing with stylistic modifications and moral censorships. I am aiming to increase the accessibility to manuscripts, by providing a structure for the modifications and to assist in evaluation of certain modifications. Furthermore the proposed method may be applied on different editorial problems, which I will discuss in the Outlook section.

## Method

As brought up in the Problem section, the proposed method focuses on stylistic and moral censorship reasons, based on the assumption that these two types of reasons relate to the topic of the modification. I convey a generative topic model, that is based on Latent Dirichlet Allocation (D. Blei, Ng, & Jordan, 2003) and is able to take into account the structural information of modifications. There exists a wide range of topic models that customize LDA and many of these take into account additional structural information. To replace the Bag-of-words approach by introducing structural information about the word order is a major field of LDA research (Gruber, Rosen-Zvi, & Weiss, 2007; Wallach, 2006). Moreover there exists a lot of research on topic hierarchies (D. M. Blei, Griffiths, & Jordan, 2010; Paisley, Wang, Blei, & Jordan, 2015). LDA has also been modified to work with graph-structured documents (Xuan, Lu, Zhang, & Luo, 2015). However I am not aware of any literature that shows how to model modification reasons in a corpus of natural language.

Figure 1 illustrates the conceptual functioning of the method from left to right. As input on the left, a collection of documents is given. The documents have parts marked as modified. The generative model in the center infers reasons by taking into account all text, inside and outside the modifications. Every reason may stand for a stylistic, or a certain moral censorship reason (e.g. political, religious). On the right side, the model outputs a reason-modification assignment.

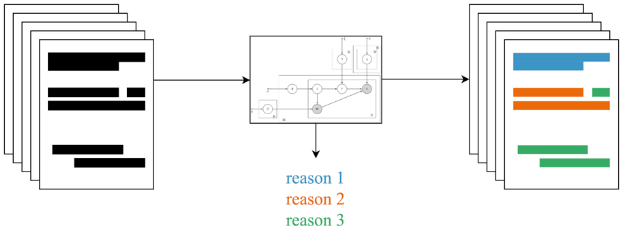

Figure 1: The generative model in the center receives input documents with modifications. It outputs reasons for modification and a reason assignment to each modification

In addition to the LDA latent variables, I introduce a topic-reason variable $\gamma$, a word-reason-modification tendency $\lambda$ and a token-reason assignment r. The complete model in plate notation is shown in Figure 2. c (observed) models whether a token has been modified. For every topic, $\gamma$ holds a distribution over reasons, which may cause a modification. For most modifications this distribution should be sparse, for example if a censor crosses out a sentence that discusses the financial situation of the author, the the topic and the reason for censorship would be identical. On the contrary a stylistic modification wouldn't always have one or two clear corresponding topics.

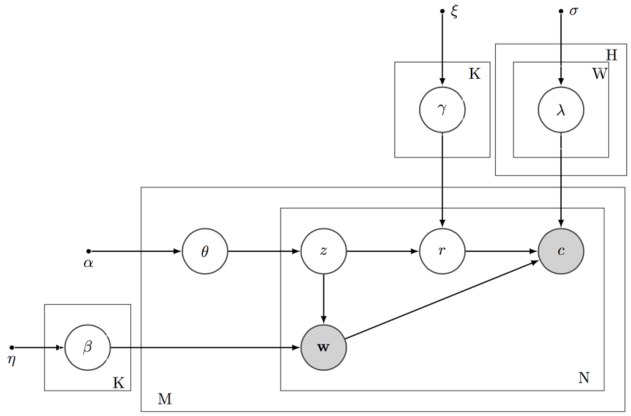

Figure 2: Plate notation of the model. The left four circled variables represent LDA, the right ones the modification part

For every token and every reason, λ holds a distribution over two states, namely whether the token tends to be modified for this reason. There may be token, that are representative for a topic, but they nonetheless do not tend to be modified. The categorical variable *r* represents the reason assignment at that position.

The latent variables can be iteratively approximated using Variational Inference (Bishop, 2006; Zhao, 2013).

## Intermediate results

In this section, evaluation methods on toy data are discussed and characteristics of the IB data set, as well as preparation steps and first intermediate results are presented.

### Toy data

To evaluate the characteristics of this method, experiments with artificial toy data can be performed. The generative model described above can be employed for inference as well as for generating artificial documents with modifications. A typical experiment to evaluate a generative model is conceived as follows:

1. Initialize the latent variables of the model randomly
2. Generate documents with modifications
3. Re-initialize the latent variables
4. Try to infer the latent variables from the generated documents

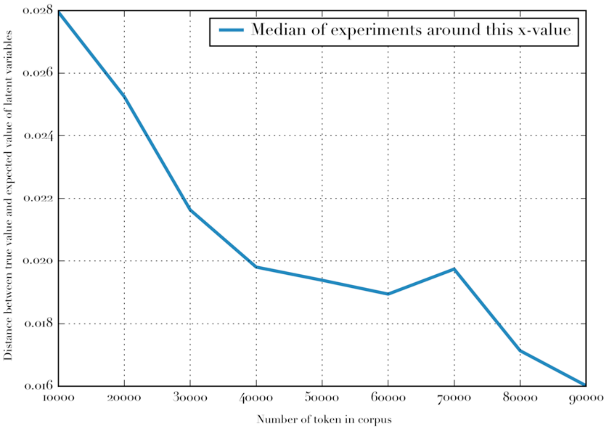

Figure 3: Experiment with 585 generated corpora varying the size from 10.000 to 90.000 token and performing inference for z, r, θ und γ, leaving the remaining model parameters fixed

I have performed a series of experiments to investigate the sensitivity of the model to corpus size. As expected, the accuracy of the model increases with an increasing amount of data. Figure 3 shows a decreasing distance between the true value and the expected one, when increasing the size of the data set. The accuracy does also largely depend on the sparsity of the concentration factors, which means that in order to predict the minimal size a real data set should have, one has to come up with according prior concentration factors for the Dirichlet variables.

### IB data set

To apply this method on the IB data set, some preprocessing steps are necessary. Apart from standard natural language preprocessing, one has to filter out all corrections of mistakes.

Table 1 shows the change of data set characteristics that are caused by the pre-processing. A lot of modifications have been considered to be corrections of mistakes and thus have been filtered out.

The visualization in Figure 4 reveals a great variety in the structure of the modifications. The figure shows the state of all tokens of two letters from the IB data set. The upper letter contains a lot of small changes, where often a green (added part) and red (deleted part) occur as a combination. The letter below contains a lot of longer deleted parts, concluding, that the letter above contains corrections of mistakes, whereby the lower contains modifications related to the topic.

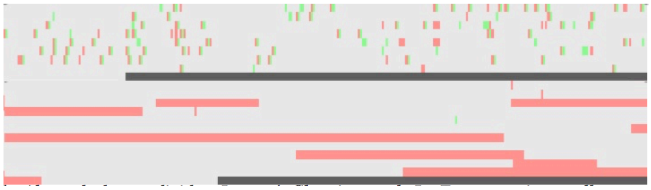

Figure 4: Above dark grey divider: Letter 4, Chamisso to de La Foye contains small corrections. Below: Letter 14, Dorothea Tieck to Uechtritz contains larger modifications. Deletions (red), additions (green)

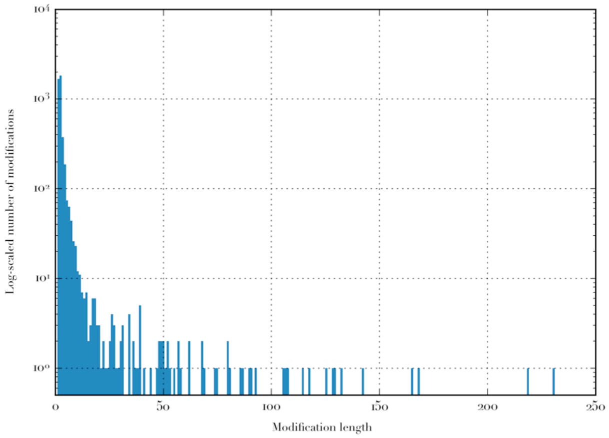

Figure 5: 93% of the modifications are shorter than 6 token. Two outliers of length 349 and 470 have not been included in the visualization

The size of the modifications seems to be a criterion to distinguish between corrections and topic related modifications. The distribution over the length of modifications however, reveals, that a lot of the modifications in the data set are small, thus likely to be corrections of mistakes (Figure 5).

The first preliminary experiments have been carried out with a binary setup: The model should distinguish between a) one particular moral censorship reason and b) everything else. To do so, prior knowledge about the topics has been introduced to the model in form of a keyword.

For example the topic sickness has been introduced to the model by the keyword “Krankheit”. The first results on this look very promising, as they reveal a precision = 1 and a recall = 0,67.

## Outlook

In the near future, I will undertake further experiments with the IB data set. To do so, I will incrementally increase the number of modification reasons. The results will be made accessible as part of the IB corpus, making the permeability between editorial and algorithmic work more visible and accessible to all interested DH communities for reuse.

In a further step, I would like to look into different applications of this method. A promising idea would be, to look into different editions of the same text and consider each difference as a modification.

## Bibliography

**Baillot, A.** (Ed.). (2016). *Letters and texts. Intellectual Berlin around 1800.* Berlin: Humboldt-Universität zu Berlin. Retrieved from http://www.berliner-intellektuelle.eu/. Please visit the web page for an up-to-date version.

**Bishop, C. M.** (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer Science+Business Media, LLC.

**Blei, D. M., Griffiths, T. L., & Jordan, M. I.** (2010). *The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. J.* ACM, 57(2), 1–30.

**Blei, D., Ng, A., & Jordan, M.** (2003). *Latent Dirichlet allocation.* JMLR.

**Gruber, A., Rosen-Zvi, M., & Weiss, Y.** (2007). Hidden Topic Markov Models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. JMLR.

**Paisley, J., Wang, C., Blei, D. M., & Jordan, M. I.** (2015). Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 37(2).

**Wallach, H.** (2006). Topic Modeling: Beyond Bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning.* New York: ACM.

**Xuan, J., Lu, J., Zhang, G., & Luo, X.** (2015). Topic model for graph mining. *IEEE Transactions on Cybernetics*, 45(12).

**Zhao, W. X.** (2013). *Varitional Methods for Latent Dirichlet Allocation.* Retrieved from http://net.pku.edu.cn/~zhaoxin/vEMLDA.pdf (accessed on 1st of November 2016)