# Topic Balancing with Additive Regularization of Topic Models

**Veselova Eugeniia**
Moscow Institute
of Physics and Technology
Moscow, Russia
veselova.er@phystech.edu

**Vorontsov Konstantin**
Moscow Institute
of Physics and Technology
Moscow, Russia
vokov@forecsys.ru

## Abstract

This article proposes a new approach for building topic models on unbalanced collections in topic modelling, based on the existing methods and our experiments with such methods. Real-world data collections contain topics in various proportions, and often documents of the relatively small theme become distributed all over the larger topics instead of being grouped into one topic. To address this issue, we design a new regularizer for $\Theta$ and $\Phi$ matrices in probabilistic Latent Semantic Analysis (pLSA) model. We make sure this regularizer increases the quality of topic models, trained on unbalanced collections. Besides, we conceptually support this regularizer by our experiments.

## 1 Introduction

Topic modelling is a widespread approach to unsupervised text analysis and clustering. Given the number of latent variables — topics — topic models extract hidden word×topic and topic×document probability distributions from text corpora. Topic models have proven to be relevant in a wide range of contexts and uni- and multilingual tasks (Uys et al., 2008; De Smet and Moens, 2009; Boyd-Graber et al., 2017).

Two fundamental topic models are probabilistic Latent Semantic Analysis — pLSA (Hofmann, 1999) and Latent Dirichlet Allocation — LDA (Blei et al., 2003). Various extensions of pLSA and LDA models have emerged over the past years, e.g. Additive Regularization of Topic Models (ARTM) (Vorontsov and Potapenko, 2015) modification of pLSA, where required solution properties are induced by the additional regularizer part in the model. Through regularizers one can take into consideration various problem-specific features of data, and this is a reason why we apply ARTM-framework in our work.

Despite almost 30 years of model development history, lots of problems and issues were raised in the topic modelling field. Problem of the "order effect" in LDA (Agrawal et al., 2018), for example. It consists in converging to the different topics set while during training on the unstructured data. Even with the structured data solution in the pLSA or LDA model is non-unique and unstable. Such unstability may be reduced by tuning the model with regularizers, as in the ARTM model. Inserting $\Phi$ and $\Theta$ prior distribution into the model, according to the (Wallach et al., 2009), promotes convergence to the better and stable solution along with regularization. However, many problems with models itself and with quality metrics remain unsolved.

In this article, we point out the topic balancing problem. At this moment problem of training topic models on the unbalanced collections is not studied thoroughly and is far from the comprehensive solution. We examine previously suggested approach to the topic balancing and propose a balancing procedure, based on the a priori ratio between topic capacities.

## 2 Problem statement

### 2.1 Topic modelling introduction

Let $D$ denote the text corpora, $W$ denote the set of words in the corpora, or the corpora vocabulary, and $T$ denote the set of the topics. Every document $d \in D$ is presented as a token sequence $(w_1, w_2, \ldots, w_{n_d})$ of length $n_d$ from the vocabulary of size $n$. In the models, based on the "bag-of-words" hypothesis, the more compact way to represent a document is to consider the document as a vocabulary multiset, where each token $w \in d$ occurs $n_{dw}$ times in the document.

Topic model describes conditional probabilities $p(w|d)$ of the appearance of the tokens $w$ in the documents $d$ through the probabilities of the to-



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 59–65

kens in the topics $\varphi_{wt} = p(w|t)$ and topics in the documents $\theta_{td} = p(t|d)$. To build a probabilistic generative model, we consider further hypotheses fulfilled:

- conditional independence hypothesis: each topic generates tokens regardless of the document;
$$p(w|d,t) = p(w|t)$$
- "bag-of-words" hypothesis: words order in the document does not affect desired distributions;
- a finite set of topics $T$ exist in the corpora, and each token occurrence in each document refers to some latent topic from $T$.

According to the law of total probability and the assumption of conditional independence

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td}$$

This probabilistic model describes how the collection $D$ is generated from the known distributions $p(w|t)$ and $p(t|d)$. Learning a topic model is an inverse problem: obtaining tokens–topics and topics–documents distributions $p(w|t)$ and $p(t|d)$ given a corpora $D$. This problem is equivalent to finding a stochastic matrix decomposition of counter matrix as a product $F \approx \Phi\Theta$, where matrix $\Phi$ represents *tokens probabilities for the topics* and $\Theta$ represents *topic probabilities for the documents*:

$$F = (\hat{p}(w|d))_{W \times D},\ \hat{p}(w|d) = \frac{n_{wd}}{n_d}$$

$$\Phi = (\varphi_{wt})_{W \times T},\ \varphi_{wt} = p(w|t)$$

$$\Theta = (\theta_{td})_{T \times D},\ \theta_{td} = p(t|d)$$

In pLSA the topic model is learned by log-likelihood maximization through EM-algorithm

$$L(\Phi, \Theta) = \sum_{d \in D, w \in d} n_{dw} \log \sum_{t \in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta} \tag{1}$$

Further details can be found in the Appendix A.

Since the matrix product $\Phi\Theta$ is defined up to a linear transformation, solution of the problem is not unique and, therefore, is unstable. Additional objectives called *regularizers*, depending on the $\Theta$ and $\Phi$ matrices, can be included in the log-likelihood along with their non-negative *regularization coefficients* $\tau$ to reduce the solution domain. Likelihood maximization problem (1) with $r$ regularizers then takes the following form:

$$L(\Phi, \Theta) + \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta} \tag{2}$$

Solution of the problem therefore transforms to

$$p_{\hat{t}dw} = \frac{\varphi_{w\hat{t}}\theta_{\hat{t}d}}{\sum\limits_{t \in T} \varphi_{wt}\theta_{td}}$$

$$\varphi_{wt} = \underset{w \in W}{\text{norm}} \left( n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)$$

$$\theta_{wt} = \underset{t \in T}{\text{norm}} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

where

$$n_{wt} = \sum_{d \in D} n_{dw} p_{tdw},\ n_{td} = \sum_{w \in d} n_{dw} p_{tdw}$$

Regularization approach and theorem proofs can be found in (Vorontsov and Potapenko, 2015)

### 2.2 Topic balancing problem statement

Let $n_t = \sum\limits_{d \in D} p(t|d) n_d$ denote the *topic capacity* of the topic $t$. Let $k = \frac{n_{t_{max}}}{n_{t_{min}}}$ denote the *imbalance degree* of the model; with $p(t) = \frac{n_t}{n}$ denoting the *topic probability* and $N(\hat{t}) = |\{d \in D | \underset{t}{\text{argmax}}\ \theta_{td} = \hat{t}\}|$, we can denote *documents imbalance degree* $k = \frac{N_{t_{max}}}{N_{t_{min}}}$ too. Probabilistic topic models, based on the matrix factorization, tend to spread documents by topics uniformly and extract topics with the equal capacity. In order to maximize log-likelihood, model should engage all inner parameters for data description. Reducing the topics number, meaning reducing the number of available parameters, is unprofitable for the model in terms of EM-algorithm optimization, therefore strong proportion reduction of the particular topic is unprofitable too. Experiments show that in the pLSA and LDA models imbalance degree rarely exceeds 3-4.

Similar problem arises in the multiclass classification with imbalanced data, where classifying model prefers predicting the label of the most common class for every object to reduce the number of errors in classification. The standard approach to imbalanced data problem is a class weighting. It can help to provide some bias towards the minority classes while training the model, and thus help in

improving performance of the model while classifying various classes. Documents imbalance leads to overweight of the vocabulary of predominant topics in the collection. This effect exaggerates "word burstiness" in the model (Doyle and Elkan, 2009; Lei et al., 2011) in terms of documents: if a collection has disproportion of topics, a document is likely to belong to the widely represented topic.

Let us call the model *imbalanced* if it can extract and maintain topics with the imbalance degree $k$ up to 10. In this article, we examine different ways of balancing topics in topic models and building imbalanced models.

## 3 Topic balancing hypotheses

### 3.1 Iterative renormalization of parameter in the Dirichlet distribution

While formulating the probabilistic generative model in terms of LDA, topic distributions over words and document distributions over topics are generated from prior Dirichlet distributions. A learning algorithm for LDA can also be considered as an EM-like algorithm with modified M-step (Asuncion et al., 2009). The most simple and frequently used modification is the following:

$$\varphi_{wt} \propto n_{wt} + \beta_w, \; \theta_{td} \propto n_{td} + \alpha_t$$

Thus probabilities of words in topics and probabilities of topics in documents are estimated with apriori shift. This LDA modification is covered by the ARTM framework through the LDA regularizer

$$R(\Phi, \Theta) = \sum_t \sum_w (\beta_w - 1) \log \varphi_{wt} + \\ + \sum_d \sum_t (\alpha_t - 1) \log \theta_{td}$$

and parameters of Dirichlet distributions can be manually adjusted.

We put forward a hypothesis that increasing Dirichlet parameters in proportion to the topic capacities similar to the classes weighing in unbalanced classification can countervail tendency of the EM-algorithm to decrease the capacity of the big topics and increase the capacity of the small topics.

For the modelling experiment we chose synthetic collection which consists of the two themes — business and music — with 1000 and 150 documents respectively. Two pairs of models were built to compare modelling results and evaluate balancing opportunity. First models were trained with two topics with and without renormalization, second — with six topics. In the second pair, the separation of topics was evident through each topic size and top-tokens: five topics had top-tokens from a big theme (with $\sim 200$ documents in each topic), the last one topic had top-tokens from a small theme. However, better topics were obtained with balanced Dirichlet parameters. In the first pair of models we implied that through the process of rebalancing Dirichlet parameters we could obtain two topics with $\sim 150$ and $\sim 1000$ documents each and different top-tokens. This hypothesis was not fully confirmed in the experiment: without the parameter renormalization EM-algorithm had converged to the topics with almost similar topic capacities, with parameter renormalization model maintained documents imbalance degree equal 2 instead of 7. Results of the experiment can be seen in Figure 1.

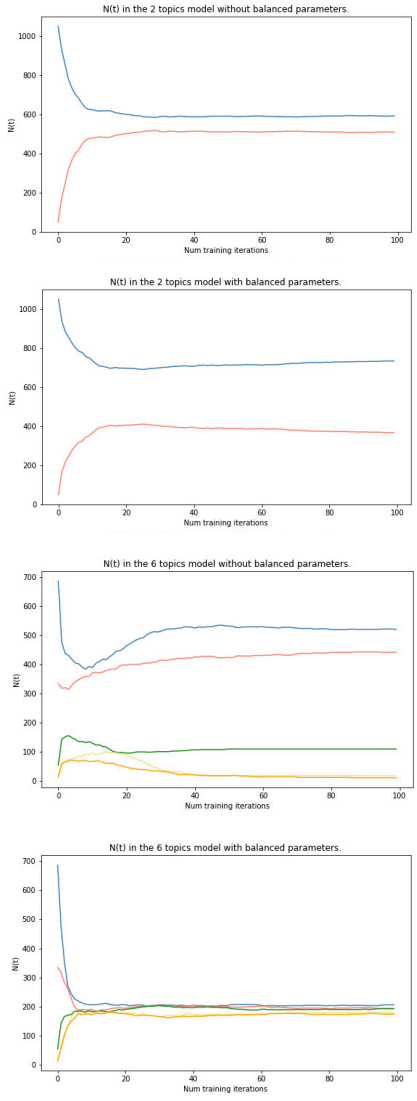


Figure 1: Results of LDA renormalization.

### 3.2 Rebalancing $p(t|d, w)$

Referring a weighting classes approach in the unbalanced classification task, we considered possibility to rebalance $p(t|d, w)$ (4). We proposed dividing $n_{tdw}$ by $n_t$. However, the same experiment as with LDA model gave no positive results, and later, in the subsection, we are going to prove this hypothesis failure.

We show that dividing $p(t|d, w)$ by any value $Z_t$, which depends on $t$ only, does not change $\Phi$, but only leads to minor the topics redistribution in documents. Proof can be found in the Appendix B. We prove that during renormalization in the EM-algorithm, M-step formulas for $\Phi$ does not change, because normalizing multiplier $Z_t$ is reduced. Therefore, pLSA renormalization does not influence the topics.

### 3.3 $\Phi$ initialization

According to the (Wallach et al., 2009), $\Phi$ and $\Theta$ prior distribution, inserted into the model, could promote stability of the solution. We followed this assumption and conducted an experiment, in which $\Phi$ matrix was initialized not randomly, as in the unmodified topic models, but with the previously calculated probabilities according to the foregone distribution of documents by topics. We suppose that the "real" $\Phi$ initialization along with the $\Theta$, calculated from $\Phi$, are the optimal factorization of the counter matrix $F$ in terms of log-likelihood. Therefore, the overall topic balance and relative change of $\Phi$ matrix value must not be small enough ($\sim 1 - 3\%$).

For this experiment chose four synthetic collections with two themes about business and music: first collection consisted of 1000 and 10 documents per theme respectively, second consisted of 1000 and 100 documents, third consisted of 1000 and 300 documents, and fourth consisted of 1000 and 600 documents respectively.

The experiment was split into two levels: at the first level, we trained models without a priori $\Phi$ initialization, at the second level, beforehand calculated $\Phi$ matrix was used as an initial tokens–topics distribution for each model. All zero a priori probabilities in the calculated $\Phi$ matrix were replaced with the minimal possible probability value $\propto 10^{-5}$. Zero probabilities emerge when a word does not occur in any document of the foregone topic; hence we are not artificially limiting topic vocabularies by preserving zeroes. We were training

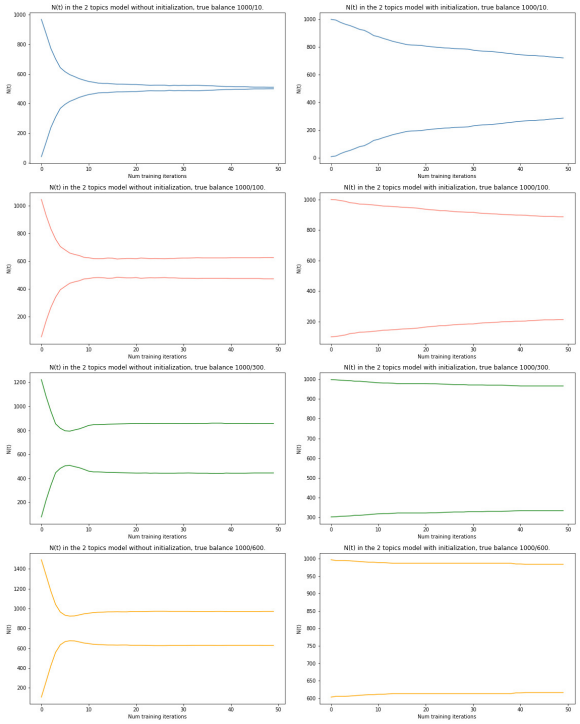


Figure 2: Results of a priori $\Phi$ initialization in pLSA model.

and comparing pairs of basic model with two topics and model with the initialization of the $\Phi$ matrix with two topics for each collection, eight models in sum. Regardless of the data collection, after first 10 training iterations, uninitialized models converged to the balanced solutions with almost equal $N(t)$, though initial initialization supported documents imbalance degree up to 6. This result is represented in Figure 2 through the topic's N(t). The left column represents the model without initialization, the right column represents the model with initialization with true topic's balance [10:1000, 100:1000, 300:1000, 600:1000] respectively.

## 4 Topic prior regularizer

### 4.1 Description of the regularizer

According to our experiments and modelling experience, log-likelihood functional optimization does not preserve topic balance in models and does not converge to the optimal solution from the user's point of view. We want an optimal solution to allow topics with relatively small topic capacities or topics with relatively small $p(t|d)$ for the most of corpora documents. Optimality in such terms can be achieved in a solution, where some topic variables, or degrees of freedom, are not fully utilized. Current functional during the optimization

via EM-algorithm tends to redistribute $p(t|d)$ in the most efficient way, without degenerate distributions. Thus topic capacities obtain similar values during the training process.

We formed the hypothesis from our experiments, that additional shift in tokens–topics $\Phi$ may influence the EM-algorithm as a restriction of the degrees of freedom, supporting topics imbalance. By setting relative collection balance in $\Phi$ in advance, we can control possible collection balance after the training process. During the optimization, all $\varphi_{wt}$ are specified according to the tokens distribution in documents. We implemented this hypothesis in a new ARTM regularizer $R_{TopicPrior}$ called *TopicPriorRegularizer* with the parameter $\beta$ to describe a priori topic balance in the collection.

$$R_{TopicPrior}(\Phi, \Theta) = \sum_t \sum_w \beta_t \log \phi_{wt}$$

To better understand the $R_{TopicPrior}$ influence on the EM-algorithm, we calculated the $R_{TopicPrior}$ partial derivative:

$$\frac{\partial R}{\partial \Phi_{wt}} = \frac{\beta_t}{\varphi_{wt}}$$

and modified log-likelihood in case of one additional regularizer with regularization coefficient $\tau$, determining regularizing strength:

$$\varphi_{wt} \propto n_{wt} + \tau \beta_t$$

In most of the cases, we lack knowledge about topic capacities in the researched data collection, therefore we cannot set precise $\beta$ value. We generalize our regularization approach and propose $R_{TopicPriorSampled}$ regularizer, where $\beta$ parameter is being sampled from the Dirichlet distribution with the parameter $\gamma \in \mathbb{R}^1$. $\gamma$ is responsible for the estimated data sparsity, thus $\gamma = 1$ stands for the random topic capacities in a model, $\gamma \ll 1$ stands for the equal topic capacities, $\gamma \gg 1$ stands for the significantly uneven topic capacities.

$$\beta \sim \mathrm{Dir}(\gamma),\ \gamma \in \mathbb{R}^1$$

### 4.2 Modelling experiments

For the first modelling experiment we chose synthetic collection with the two themes — business and music — with 1000 and 100 documents respectively. We build two models with two topics in each and train them for 15 epochs, however, the second model is trained with the $R_{TopicPrior}$, where $\beta = [0.1, 0.9]$. After training we evaluate both

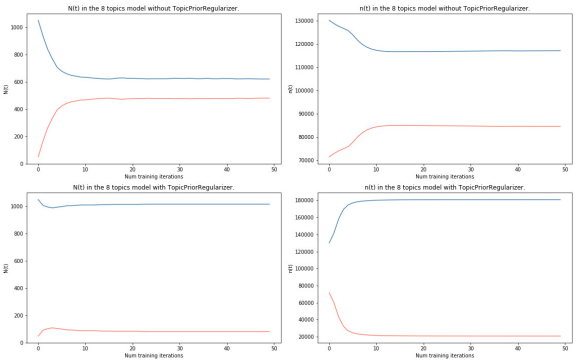


Figure 3: Results of unregularized and regularized pLSA model training with 2 topics.

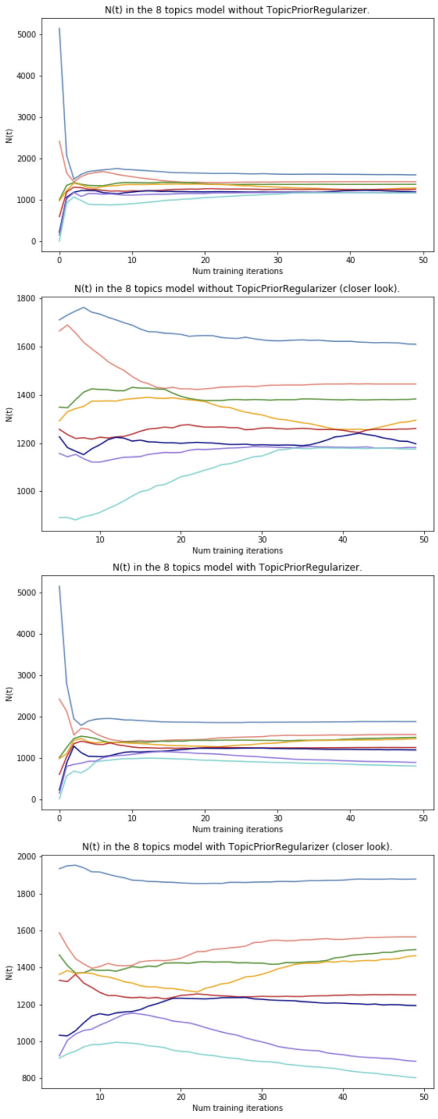


Figure 4: Results (N(t)) of unregularized and regularized pLSA model training with 8 topics.

models by their perplexity, top-tokens and $n(t)$ for every topic in the model. The second model had extracted a small theme as a distinct topic, while the first unregularized model has two similar topics. Training results are presented in Figure 3: the first row represents model without regularizer, the second row represents regularized model; the left column represents N(t) of the topics, the right column represents n(t) of the topics.

For the second modelling experiment we choose collection with the eight themes, balanced with the following documents proportion: $doc\_prop = [3000, 2000, 1500, 1000, 1000, 1000, 700, 350]$. Two models were trained on this collection: unregularized and regularized model, where regularizer was initialized with $\beta = \frac{doc\_prop}{sum(doc\_prop)}$. Figure Figure 4 and Figure 5 show better topics composition in the second model, compared to the first model results.

## 5 Discussion and conclusion

Learning an unbalanced topic model from unbalanced text collection is a non-trivial task for all of the existing modelling methods. In this paper we discussed the problem of training topic models with unbalanced text collections. No previous research provides a thorough analysis of this problem or an efficient training procedure for unbalanced models. After reviewing the problem, we proposed an approach to building topic models, able to maintain relatively high imbalance degree. We described our approach in terms of pLSA regularization and brought theoretical justification for the $R_{TopicPrior}$ regularizer.

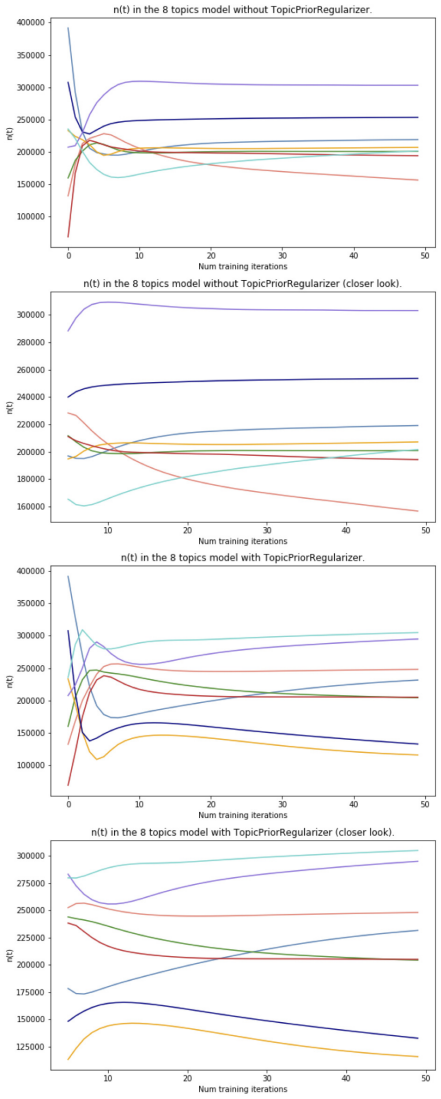


Figure 5: Results (n(t)) of unregularized and regularized pLSA model training with 8 topics.

## References


Amritanshu Agrawal, Wei Fu, and Tim Menzies. 2018. What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88.

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 27–34. AUAI Press.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Wim De Smet and Marie-Francine Moens. 2009. Cross-language linking of news stories on the web using interlingual topic modelling. In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pages 57–64. ACM.

Gabriel Doyle and Charles Elkan. 2009. Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Shaoze Lei, JianWen Zhang, Shifeng Weng, and Changshui Zhang. 2011. Topic model with constrainted word burstiness intensities. In *The 2011 International Joint Conference on Neural Networks*, pages 68–74. IEEE.

JW Uys, ND Du Preez, and EW Uys. 2008. Leveraging unstructured information using topic modelling. In *PICMET'08-2008 Portland International Conference on Management of Engineering & Technology*, pages 955–961. IEEE.

Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

## A pLSA and ARTM model optimization problem

In pLSA the topic model is learned by log-likelihood maximization through EM-algorithm

$$L(\Phi,\Theta)=\sum_{d\in D,w\in d} n_{dw}\log\sum_{t\in T}\varphi_{wt}\theta_{td}\to\max_{\Phi,\Theta} \tag{3}$$

with linear constraints of non-negativity and normalization:

$$\sum_{w\in W}\varphi_{wt}=1,\ \varphi_{wt}\geq 1;\quad \sum_{t\in T}\theta_{td}=1,\ \theta_{td}\geq 1$$

Solution of the pLSA problem satisfies the following system of equations with auxiliary variables $p_{tdw}$:

$$\begin{gathered} p_{\hat{t}dw}=\frac{\varphi_{w\hat{t}}\theta_{\hat{t}d}}{\sum\limits_{t\in T}\varphi_{wt}\theta_{td}} \\ \varphi_{wt}=\operatorname*{norm}_{w\in W}(n_{wt}),\ n_{wt}=\sum_{d\in D}n_{dw}p_{tdw} \\ \theta_{wt}=\operatorname*{norm}_{t\in T}(n_{td}),\ n_{td}=\sum_{w\in d}n_{dw}p_{tdw} \end{gathered} \tag{4}$$

Process of the calculation auxiliary variables $p_{tdw}$ is an E-step, while model parameters elaboration by the calculated $p_{tdw}$ is an M-step in the EM-algorithm.

## B Proof of rebalancing failure

We considered possibility to rebalance $p(t|d,w)$ in accordance with weighting classes approach. We proposed dividing $n_{tdw}$ by $n_t$.

We show that dividing $p(t|d,w)$ by any value $Z_t$, which depends on $t$ only, doesn't change $\Phi$, but only leads to minor the topics redistribution in documents. We put $R=0$ in (2) for the sake of simplicity.

Investigating M-step of the EM-algorithm, we write down log-likelihood with renormalizing factor $\frac{1}{Z_t}$:

$$\frac{1}{Z_t}\sum_{d\in D}\sum_{w\in d}\sum_{t\in T}n_{dw}\varphi_{wt}\theta_{td}\to\max_{\Phi,\Theta}$$

and then separate variables $\Phi$ and $\Theta$:

$$\sum_{w\in W}\sum_{t\in T}\frac{n_{wt}}{Z_t}\log\varphi_{wt}+ \\ +\sum_{d\in D}\sum_{t\in T}\frac{n_{td}}{Z_t}\log\theta_{td}\to\max_{\Phi,\Theta}$$

To solve this linear programming task, we apply Karush–Kuhn–Takker conditions. We write Lagrangian:

$$\begin{aligned}\mathcal{L}(\Phi,\Theta)=&\sum_{w\in W}\sum_{t\in T}\frac{n_{wt}}{Z_t}\log\varphi_{wt}- \\ &-\sum_{t\in T}\lambda_t\left(\sum_w\varphi_{wt}-1\right)+ \\ &+\sum_{d\in D}\sum_{t\in T}\frac{n_{td}}{Z_t}\log\theta_{td}- \\ &-\sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td}-1\right)\end{aligned}$$

and equate its derivations to zero:

$$\begin{gathered}\frac{\partial\mathcal{L}}{\partial\varphi_{wt}}=\frac{n_{wt}}{Z_t\varphi_{wt}}-\lambda_t=0 \\ \lambda_t\varphi_{wt}=\frac{n_{wt}}{Z_t}\quad\Rightarrow\quad\lambda_t=\frac{n_t}{Z_t} \\ \varphi_{wt}=\operatorname*{norm}_{w\in W}(n_{wt})\end{gathered}$$

and

$$\begin{gathered}\frac{\partial\mathcal{L}}{\partial\theta_{td}}=\frac{n_{td}}{Z_t\theta_{td}}-\mu_d=0 \\ \mu_d\theta_{td}=\frac{n_{td}}{Z_t}\quad\Rightarrow\quad\mu_d=\sum_{t\in T}\frac{n_{td}}{Z_t} \\ \theta_{td}=\operatorname*{norm}_{t\in T}\left(\frac{n_{td}}{Z_t}\right)\end{gathered}$$

M-step formulas for $\Phi$ does not change, because normalizing multiplier $Z_t$ is reduced. Therefore, pLSA renormalization has no influence on the topics.