# Stable Topic Modeling for Web Science: Granulated LDA

Sergei Koltcov skoltsov@hse.ru
Sergey I. Nikolenko sergey@logic.pdmi.ras.ru
Olessia Koltsova ekoltsova@hse.ru
Svetlana Bodrunova visual@jf.pu.ru

Laboratory for Internet Studies, National Research University Higher School of Economics, St. Petersburg, Russia

## ABSTRACT

Topic modeling is a powerful tool for analyzing large collections of user-generated web content, but it still suffers from problems with topic stability, which are especially important for social sciences. We evaluate stability for different topic models and propose a new model, granulated LDA, that samples short sequences of neighboring words at once. We show that gLDA exhibits very stable results.

## CCS Concepts

•**Computing methodologies** → **Topic modeling;**

## Keywords

topic modeling, latent Dirichlet allocation, Gibbs sampling

## 1. INTRODUCTION

In social sciences, topic modeling can be used to concisely describe a large corpus of documents, uncovering the actual topics covered in this corpus (via the word-topic distributions) and pointing to specific documents that deal with topics a researcher is interested in (via the topic-document distributions) and to mine latent variables from the documents. *Topic stability* is also a very important problem for real life applications of topic modeling, especially in social sciences. For a practical application of topic models it is highly desirable to have stable results: a social scientist is often interested in whether a topic is "there" in the dataset, and it would be hard to draw any conclusions if the topic was "blinking" in and out depending on purely random factors. Besides, it would be hard to rely on a study that cannot be reliably reproduced even in principle. In this work, we introduce a new modification of the basic latent Dirichlet allocation (LDA) model called *granulated LDA* (GLDA) that assumes that topics cover relatively large contiguous subsets of a document and automatically assigns the same topic to a whole window of words once the anchor word has been sampled in this window. We show that GLDA produces much more stable results while preserving approximately the same or better topic quality as classical topic models.



*WebSci '16*



ACM ISBN 978-1-4503-4208-7/16/05.

DOI: http://dx.doi.org/10.1145/2908131.2908184

## 2. TOPIC MODELING

Let $D$ be a collection of documents, and let $W$ be the set of all words in them (vocabulary). Each document $d \in D$ is a sequence of terms $w_1, \dots, w_{n_d}$ from the vocabulary $W$. The basic assumption of all probabilistic topic models is that there exists a finite set of topics $T$, each occurrence of a word $w$ in a document $d$ is related to some topic $t \in T$, and specific words occurring in a document depend only on the corresponding topic occurrences and not on the document itself: $p(w \mid d) = \sum_{t \in T} p(w \mid t)p(t \mid d) = \sum_{t \in T} \phi_{wt}\theta_{td}$, where $\phi_{wt} = p(w \mid t)$ is the distribution of words in a topic and $\theta_{td} = p(t \mid d)$ is the distribution of topics in a document. To train a topic model, one has to find multinomial distributions $\phi_{wt}$, $t \in T$, and $\theta_{td}$, $d \in D$, which we denote as matrices $\Phi = (\phi_{wt})_{wt}$ and $\Theta = (\theta_{td})_{td}$ respectively.

There are several approaches to topic modeling: *probabilistic latent semantic analysis* (pLSA) model optimizes the total log-likelihood with the EM algorithm, *latent Dirichlet allocation* (LDA) [1] adds Dirichlet priors to the $\theta$ and $\phi$ distributions, and *additive regularization of topic models* (ARTM) [7] adds regularizers explicitly to the objective function In any case, topic modeling basically approximates $F = (F_{dw})$ of size $|D| \times |W|$ by a product of $\Theta$ and $\Phi$ of size $|D| \times |T|$ and $|T| \times |W|$. Obviously, if $F = \Theta\Phi$ is a solution of this problem then $F = (\Theta S)(S^{-1}\Phi)$ is also a solution for any nondegenerate $S$. In practice this means that by running the same algorithm on the same dataset we get very different matrices $\Phi$ and $\Theta$, which is obviously an undesirable property. Hence, regularization is important in topic models, but regularizers for improving topic stability have virtually never been studied, except perhaps for *semi-supervised LDA* [2,6], already applied to social sciences, where one singles out topics related to specific subjects in question by defining a set of seed words and restricting topic samples to a subset of topics for these seed words.

## 3. GRANULATED LDA

In this work, we introduce the *granulated sampling* approach which is based on two ideas. First, we recognize that there may be a dependency between a pair of unique words, but, unlike the convolved Dirichlet regularizer model, we do not express this dependency as a predefined matrix. Rather, we assume that a topic consists of words that are not only described by a Dirichlet distribution but also often occur together; that is, we assume that words that are

characteristic for the same topic are often colocated inside some relatively small window. We view each document as a granulated surface consisting of granulas (topic occurrences) that can be sequences of consecutive words of any length, assuming that all words inside a granula belong to the same topic, and the only new model parameter is the width of the granula (sampling window). The idea is to capture the intuition that words that are located close to each other in the document usually relate to the same topic; i.e., topics in a document are not distributed as independently sampled random variables but rather as relatively large contigious streaks, or *granulas*, of words belonging to the same topic.

Granulated Gibbs sampling is implemented as follows: we randomly sample anchor words in the document, sample their topics, but then set the topic of all words in a window around the current anchor word to the sampling result; we sample as many anchor words as there are words in the document. Formally, after the initialization of $\Theta$ and $\Phi$ matrices as in regular Gibbs sampling, we run the following algorithm: for every document $d \in D$, repeat $|d|$ times: (1) sample a word instance $j \in d$ uniformly at random; (2) sample its topic $z$ as in Gibbs sampling; (3) set $z_i = z$ for all $i$ such that $|i - j| \leq l$, where $l$ is a predefined window size. On the final inference stage, after sampling is over, we compute the $\Phi$ and $\Theta$ matrices as usual. Interesingly, this rather natural idea of granulas has not really been explored in topic models; the only similar approach known to us in prior work deals with using the additional information available in the text in the form of sentences and/or paragraphs [3].

# 4. EVALUATION

In our experiments, we have used a large dataset of 101481 blog posts from the *LiveJournal* blog platform. We have trained seven different models: (1) the basic probabilistic latent semantic analysis model (pLSA), implemented as the baseline ARTM model with no regularizers; (2) ARTM with $\Phi$ sparsity regularizer; (3) ARTM with $\Theta$ sparsity regularizer; (4) LDA with Gibbs sampling inference; (5) LDA with variational Bayes inference; (6) supervised LDA model with a vocabulary consisting of ethnonyms; this vocabulary was developed in a previous case study of user-generated content designed to study ethnic-related topics [2, 4, 6]; (7) granulated LDA with different window sizes, from $l = 1$ to $l = 4$. In all cases, we have trained the models with $T = 200$ topics, using two different algorithms for LDA since they may have different stability properties. For SLDA, GLDA, and LDA with inference based on Gibbs sampling, we have set the Dirichlet prior parameters to be $\alpha = 0.1$ and $\beta = 0.5$. Regularization coefficients for the ARTM models were tuned to give the best possible topics. In the experiments, we mostly strived for topic stability but we cannot afford to achieve stability at a significant loss of *topic quality*: topics of use for social sciences have to be readily interpretable.

For topic quality, we use the *coherence* and *tf-idf coherence* metrics that have been shown to be good proxies for human interpretability [5,6]. To evaluate topic stability, we use the following similarity metrics for two topics [4]: (1) symmetric Kullback–Leibler divergence between the probability distributions of two topics in a solution, defined as $\mathrm{KL}(\phi^1, \phi^2) = \frac{1}{2}\sum_w \phi^1_w \log \frac{\phi^1_w}{\phi^2_w} + \frac{1}{2}\sum_w \phi^2_w \log \frac{\phi^2_w}{\phi^1_w}$ and its normalized similarity version [4]; (2) Jaccard similarity of top words in two topics. We call two topics *matching* if their normalized Kullback-Leibler similarity is larger than 0.9 (a threshold chosen by hand so that the topics actually are similar), and we call a topic *stable* if there is a set of pairwise matching topics in every result across all runs [4]; in our experiments, we have run each model three times. Table 1 shows the results of our experimental evaluation, comparing the basic topic quality and topic stability metrics across several baseline topic models and granulated LDA with different window sizes. We have trained 200 topics for every model. Overall, we conclude that GLDA produces much more stable topics at virtually no loss to quality and interpretability.

| Topic model | Quality | | Stability | |
|---|---|---|---|---|
| | coh. | tf-idf coh. | stable | Jaccard |
| pLSA | -237.38 | -126.08 | 54 | 0.47 |
| pLSA + $\Phi$ spars. reg., $\alpha = 0.5$ | -230.90 | -126.38 | 9 | 0.44 |
| PLSA + $\Theta$ spars. reg., $\beta = 0.2$ | -240.80 | -124.09 | 87 | 0.47 |
| LDA, Gibbs sampling | -207.27 | -116.14 | 77 | 0.56 |
| LDA, variational Bayes | -254.40 | -106.53 | 111 | 0.53 |
| SLDA | -208.45 | -120.08 | 84 | 0.62 |
| GLDA, $l = 1$ | -183.96 | -125.94 | 195 | 0.64 |
| GLDA, $l = 2$ | -169.36 | -122.21 | 195 | 0.71 |
| GLDA, $l = 3$ | -163.05 | -121.37 | 197 | 0.73 |
| GLDA, $l = 4$ | -161.78 | -119.64 | 200 | 0.73 |

Table 1: Topic quality and stability.

# 5. CONCLUSION

We have introduced a novel modification of LDA, *granulated LDA*, that samples whole windows of neighboring words in a document at once. This model was intended to improve topic stability, and our experiments show that GLDA is indeed much more stable while preserving the same overall topic quality. This improvement is especially important for web science and digital humanities that seek not only interpretable topics, but essentially entire solutions that could serve as a basis to make reliable conclusions about the topical structure of text collections.

**Acknowledgments**. This work was supported by the Basic Research Program of the National Research University Higher School of Economics.

# 6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3(4–5):993–1022, 2003.

[2] S. Bodrunova, S. Koltsov, O. Koltsova, S. I. Nikolenko, and A. Shimorina. Interval semi-supervised LDA: Classifying needles in a haystack. In *Proc. MICAI 2013*, LNCS vol. 8625, pp. 265–274. Springer, 2013.

[3] R.-C. Chen, R. Swanson, and A. S. Gordon. An adaptation of topic modeling to sentences. http://rueycheng.com/paper/adaptation.pdf, 2010.

[4] S. Koltcov, O. Koltsova, and S. I. Nikolenko. Latent dirichlet allocation: Stability and applications to studies of user-generated content. In *Proc. WebSci 2014*, pp. 161–165, 2014.

[5] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proc. EMNLP 2011*, pp. 262–272, 2011.

[6] S. I. Nikolenko, O. Koltsova, and S. Koltsov. Topic modelling for qualitative studies. *Journal of Information Science*, 2015.

[7] K. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, 2014.