Stable Topic Modeling with Local Density Regularization

Sergei Koltcov[1], Sergey I. Nikolenko[1,2], Olessia Koltsova[1], Vladimir Filippov[1], and Svetlana Bodrunova[3,1]

[1] National Research University Higher School of Economics, St. Petersburg
[2] Steklov Institute of Mathematics at St. Petersburg, Russia
[3] St.Petersburg State University

Abstract. Topic modeling has emerged over the last decade as a powerful tool for analyzing large text corpora, including Web-based user-generated texts. Topic stability, however, remains a concern: topic models have a very complex optimization landscape with many local maxima, and even different runs of the same model yield very different topics. Aiming to add stability to topic modeling, we propose an approach to topic modeling based on local density regularization, where words in a local context window of a given word have higher probabilities to get the same topic as that word. We compare several models with local density regularizers and show how they can improve topic stability while remaining on par with classical models in terms of quality metrics.

1 Introduction

Over the last decade, topic modeling has become one of the standard tools in text mining. In social sciences, topic models 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) [22,23]. Apart from exploratory analysis of large text corpora, topic modeling can also be used to mine latent variables from the documents such as [12, 18].

These applications of topic modeling raise a number of problems regarding the evaluation of topic modeling results. First, it still remains an open problem to evaluate how "good" a topic is; the gold standard here is usually human interpretability, and the goal is to devise automated techniques that would come close to human estimates. Modern metrics include ones based on coherence [8,19] and its modifications [22], pointwise mutual information [6, 19, 21], and topics designed to match word intrusion and topic intrusion experiments [16].

However, apart from the actual quality of the resulting topics, *topic stability* is also a very important problem for real life applications of topic modeling, especially in social sciences. The likelihood function of a topic model is usually very complex, with plenty of local maxima. If we considering inference in a topic model as stochastic matrix decomposition, representing the word-document matrix as a stochastic product of word-topic and topic-document matrices, we see

that for every solution (Θ, Φ) there is an infinite number of equivalent solutions $(\Theta S, S^{-1}\Phi)$ for any invertible S; e.g., all permutations of the same topics are obviously equivalent. And there are plenty of substantially different solutions corresponding to different local maxima of the model posterior; the model may arrive to different local maxima depending on the randomness in initialization and sampling. For a practical application of topic models in social sciences, such as studies of Web content, 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. Hence, it becomes especially important to develop topic models that produce stable, reproducible topic solutions, hopefully not at the cost of their quality (i.e., topic interpretability).

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 assigns the same topic with high probability to a 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 topic quality as classical topic models.

The paper is organized as follows. In Section 2, we introduce the topic models that we will consider below and the two approaches to inference in topic models. Section 3 contains a brief overview of regularization in topic models. Section 4 introduces our new approach to topic modeling, granulated LDA (GLDA). In Section 5 we show experimental results that prove that granulated LDA has solutions with similar quality or better than regular topic models but that are much more stable; we conclude with Section 6.

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 $w_1, \ldots, w_{n_d}$ from the vocabulary W. The basic assumption of all probabilistic topic models is that there exists a finite set of topics T, and each occurrence of a word w in a document d is related to some topic $t \in T$, and the actual word depends only on the corresponding topic instance and not on the document itself or other words. Formally, we assume that the probability that a word w occurs in document d can be decomposed as

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) = \sum_{t \in T} \phi_{wt} \theta_{td},$$

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. The problem of training a topic model on a collection of documents is, thus, the problem of finding the set of latent topics T, i.e., the set of multinomial distributions ϕ_{wt}, $t \in T$, and the

set of multinomial distributions θ_{td}, $d \in D$, which we represent by the matrices $\Phi = (\phi_{wt})_{wt}$ and $\Theta = (\theta_{td})_{td}$ respectively.

There are two main approaches to solving this problem, i.e., reconstructing Φ and Θ. In the first approach, the total log-likelihood

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{wd} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max$$

is maximized with an expectation-maximization (EM) algorithm under constraints $\theta_{td} \geq 0$, $\phi_w t \geq 0$, $\sum_{t \in T} \theta_{td} = 1$, $d \in D$, and $\sum_{w \in W} \phi_{wt} = 1$, $t \in T$; n_{wd} denotes the number of times word w occurs in document d. This setting is the *probabilistic latent semantic analysis* (pLSA) model [13].

These ideas were further developed in the already classical *latent Dirichlet allocation* (LDA) model [4]. LDA is a Bayesian version of pLSA: it assumes that multinomial distributions θ_{td} and ϕ_{wt} are generated from prior Dirichlet distributions, one with parameter α (for the θ distributions) and one with parameter β (for the ϕ distributions). LDA inference can be done either with variational approximations or with Gibbs sampling, first proposed for LDA in [11]. Here the hidden variables z_i for every word occurrence are considered explicitly, and the inference algorithm produces estimates of model parameters as Monte Carlo estimates based on samples drawn for the latent variables. Gibbs sampling is a special case of Markov chain Monte Carlo methods where sampling is done coordinatewise, hidden variable by hidden variable. In the basic LDA model, Gibbs sampling with symmetric Dirichlet priors reduces to the so-called *collapsed Gibbs sampling*, where θ and ϕ variables are integrated out, and z_i are iteratively resampled according to the following distribution: $p(z_i = t \mid \boldsymbol{z}_{-i}, \boldsymbol{w}, \alpha, \beta) \propto$

$$q(z_i, t, \boldsymbol{z}_{-i}, \boldsymbol{w}, \alpha, \beta) = \frac{n_{-i,td} + \alpha}{\sum_{t' \in T} (n_{-i,t'd} + \alpha)} \frac{n_{-i,wt} + \beta}{\sum_{w' \in W} (n_{-i,w't} + \beta)},$$

where $n_{-i,td}$ is the number of words in document d chosen with topic t and $n_{-i,wt}$ is the number of times word w has been generated from topic t except the current occurrence z_i; both counters depend on the other variables $\boldsymbol{z}_{-w}$. Samples are then used to estimate model variables: $\theta_{td} = \frac{n_{-i,td}+\alpha}{\sum_{t' \in T}(n_{-i,t'd}+\alpha)}$, $\phi_{wt} = \frac{n_{-i,tw}+\beta}{\sum_{w' \in W}(n_{-i,w't}+\beta)}$, where ϕ_{wt} denotes the probability to draw word w in topic t and θ_{td} is the probability to draw topic t for a word in document d.

After it was introduced in [4], the basic LDA model has been subject to many extensions, each presenting either a variational or a Gibbs sampling algorithm for a model that builds upon LDA to incorporate some additional information or additional presumed dependencies. One large class of extensions deals with imposing new structure on the set of topics that are independent and uncorrelated in the base LDA model, including *correlated topic models* (CTM) [3], *Markov topic models* [17], *syntactic topic models* [7] and others. The other class of extensions takes into account additional information that may be available together with the documents and may reveal additional insights into the topical structure; this class includes models that account for timestamps of document

creation [27, 28], *semi-supervised LDA* that centers on specific topics [22], *DiscLDA* that uses document labels to solve a classification problem [15], and others. Finally, a lot of work has been done on nonparametric LDA variants based on Dirichlet processes, where the number of topics is also sampled automatically in the generative process; see [10] and references therein.

Additive Regularization of Topic Models (ARTM) [25, 26] is a recently developed novel approach to topic models that avoids complications of LDA inference (it is no easy matter to develop a new LDA extension) while preserving the capabilities for extending and improving LDA. ARTM has several conceptual differences from the Bayesian approach [25]: in ARTM, regularizers are explicit, adding new regularizers is relatively easy, and inference is done via the regularized EM algorithm. We add regularizers $R(\Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$ to the basic pLSA model, where $R_i(\Phi, \Theta)$ is some regularizer with nonnegative regularization coefficient τ_i. Then the optimization problem is to maximize $L(\Phi, \Theta) + R(\Phi, \Theta)$, where $L(\Phi, \Theta)$ is the likelihood, and the regularized EM algorithm amounts to iterative recomputation of the model parameters as follows:

$$p_{dtw} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}, \; \phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)_+, \; \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+ .$$

In this work, we use ARTM models with standard sparsity regularizers added to the Φ and Θ matrices.

3 Regularization in topic models

Whatever the inference method, the basic topic modeling problem is equivalent to stochastic matrix decomposition, where a large sparse matrix $F = (F_{dw})$ of size $|D| \times |W|$ that shows how words $w \in W$ occur in documents$d \in D$ is approximated by a product of two smaller matrices, Θ of size $|D| \times |T|$ and Φ of size $|T| \times |W|$. Note that almost by definition, the solution of this problem is not unique: if $F = \Theta\Phi$ is a solution of this problem then $F = (\Theta S)(S^{-1}\Phi)$ is also a solution for any nondegenerate $|T| \times |T|$ matrix S (for a simple example, note that we can permute topics freely, and nothing changes). In terms of the inference problem, this multitude of solutions means that an inference algorithm will converge to different solutions given different random factors in the algorithms and different starting points. In practice, by running the same algorithm on the same dataset we will get very different matrices Φ and Θ, which is obviously an undesirable property for applications.

In optimization theory, problems with non-unique and/or unstable solutions are called *ill-posed*, and a general approach to solving these problems is given by Tikhonov regularization [24]. In terms of the model definition, regularization can be viewed as extending the prior information which lets one reduce the set of solutions. Regularization is done either by introducing constraints on Φ and Θ matrices [20] or by modifying the sampling procedure [1].

In what follows, we give examples of regularizers from prior art that are relevant to the regularizer we propose in this work. First, the work [20] proposes

to introduce a regularization procedure that uses external information on the relations between words. This information, possibly from an external dataset, is expressed as a $|W| \times |W|$ covariance matrix C; formally, this adds the prior $p(\phi_t \mid C) \propto (\phi_t^\top C \phi_t)^\nu$ for some regularization parameter ν, the total log posterior looks like

$$L = \sum_{i=1}^{W} N_{it} \log \phi_{it} + \nu \log \phi_t^\top C \phi_t,$$

and the ϕ matrix is now updated as

$$\phi_{wt} \propto \frac{1}{N_t + 2\nu}(N_{wt} + \frac{2\nu\phi_{wt}\sum_{i=1}^{W} C_{iw}\phi_{it}}{\phi_t^\top C \phi_t}).$$

Another regularizer proposed in [20] is based on the idea that ϕ_{wt} depends on some matrix C which, in turn, expresses the dependencies between pairs of unique words. In other words, now a topic is defined as a collection of related words with probability distribution ψ_t, but the probability distribution of their occurrences is $\phi_t \propto C\psi_t$. The total log posterior is now

$$L = \sum_{i=1}^{W} N_{it} \log \sum_{j=1}^{W} C_{ij}\psi_{jt} + \sum_{j=1}^{W} (\gamma - 1) \log \psi_{jt}$$

under the constraints that $\sum_{j=1}^{W} \psi_{jt} = 1$. One can update the Ψ matrix similar to the updates of Φ and Θ matrices:

$$\phi_{wt} \propto \sum_{i=1}^{W} \frac{N_{it} C_{iw}}{\sum_{j=1}^{W} C_{ij}\psi_{jt}} + \gamma.$$

However, in both cases one has to know the C matrix in advance; C is a very large matrix that should incorporate prior knowledge about every pair of words in the dataset, which represents a major obstacle to using these regularizers.

Another direction of LDA extensions that has been intended, at least in part, to improve the stability of topic solutions, is the direction of *semi-supervised LDA* (SLDA) and related extensions. Semi-supervised LDA is based on a special kind of regularizer; the idea is that in real life applications, especially in social science, it often happens that the entire text corpus deals with a large number of different unrelated topics while the researcher is actually interested only in a small subset of them. In this case, it is desirable to single out topics related to the subjects in question a make them more stable. If the subject are given as a set of seed words, the semi-supervised LDA model simply fixes the values of z for certain key words related to the topics in question; similar approaches have been considered in [1, 2]. For words $w \in W_{\text{sup}}$ from a predefined set W_{sup}, the values of z are known and remain fixed to $\tilde{z}_w$ throughout the Gibbs sampling process:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} [t = \tilde{z}_w], & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise.} \end{cases}$$

Otherwise, the Gibbs sampler works as in the basic LDA model; this yields an efficient inference algorithm that does not incur additional computational costs.

In a straightforward extension, *interval semi-supervised LDA* (ISLDA), each key word $w \in W_{\text{sup}}$ is mapped to an interval of topics $[z_l^w, z_r^w]$, and the probability distribution is restricted to that interval. In the Gibbs sampling algorithm, we simply set the probabilities of all topics outside $[z_l^w, z_r^w]$ to zero and renormalize the distribution inside:

$$p(z_w = t \mid \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) \propto \begin{cases} I_{z_l^w}^{z_r^w}(z) \frac{q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}{\sum_{z_l^w \le t' \le z_r^w} q(z_w, t', \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta)}, & w \in W_{\text{sup}}, \\ q(z_w, t, \boldsymbol{z}_{-w}, \boldsymbol{w}, \alpha, \beta) & \text{otherwise,} \end{cases}$$

where $I_{z_l^w}^{z_r^w}$ denotes the indicator function: $I_{z_l^w}^{z_r^w}(z) = 1$ iff $z \in [z_l^w, z_r^w]$. Interval semi-supervised LDA has been used in case studies related to social sciences in [5, 22]; these works show that SLDA and ISLDA not only mine more relevant topics than regular LDA but also improve their stability, providing consistent results in the supervised subset of topics. In this work, we present a new LDA extension which provides even more stable results at no loss to their quality.

4 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 it as a predefined matrix. Rather, we assume that a topic consists of words that 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. 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 contiguous streaks, or *granulas*, of words belonging to the same topic. Figure 1 illustrates the basic idea, showing a granulated surface as it is usually understood in physics (bottom right) and a sample partially granulated text that might result from the granulated LDA model (on the left).

Interesingly, the 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. The work [9] adds a sentence layer to the basic LDA model; in sentence-layered LDA, each sentence is governed by its own topic distribution. Sentence and paragraph boundaries are also often used in LDA extensions dealing with sentiment analysis: it is often assumed that a single sentence or paragraph deals with only one aspect; see, e.g., the Aspect and Sentiment Unification Model (ASUM) [29] that extends the basic Sentence LDA (SLDA) model However, we are not aware of topic models that would use naturally arising granulas of fixed or variable size and assume that a granula is covered by

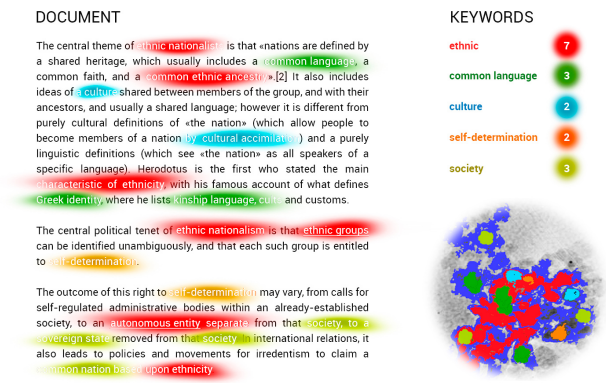

Fig. 1: Illustration for granulated LDA: granulated surface and granulated text.

the same topic. One could say that GLDA is in essence equivalent to a certain cooccurrence-based regularizer, but without the need to compute the entire cooccurrence matrix, everything is local.

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 local context window with the use of the anchor word's sampling result. We sample as many anchor words as there are words in the document.

On the other hand, the topical distribution of words inside a window (granula) can have its own distribution, different from the distribution imposed by Dirichlet priors. By modifying the distribution function inside a window (local density) and changing the window size, we can influence the model's regularization. Thus, we regularize the topic model as follows: having sampled an anchor word $z_j = z$ in the middle of a window, we then set the topics of nearby words z_i, $|i-j| \le l$, as $z_i = zK\left(\frac{|i-j|}{l}\right)$ for some kernel function K. The kernel function should satisfy $K(0) = 1$ and be monotone nonincreasing towards the ends of the window, modifying the distribution of topics inside a local window. We have compared three different kernels:

(1) step kernel $K(r) = 1$, when all topics in the window are set to z;
(2) Epanechnikov kernel $K(r) = 1 - r^2$;
(3) triangular Epanechnikov kernel $K(r) = 1 - |r|$.

Thus, formally speaking, after the initialization of Θ and Φ matrices as in regular Gibbs sampling, we run the following algorithm:

– for every document $d \in D$, repeat $|d|$ times:
  - sample a word instance $j \in d$ uniformly at random;
  - sample its topic $z_j = z$ as in Gibbs sampling;
  - set $z_i = zK\left(\frac{|i-j|}{l}\right)$ for all i such that $|i-j| \le l$.

On the final inference stage, after sampling is over, we compute the Φ and Θ matrices as usual (see Section 2).

Note that unlike regular Gibbs sampling, we do not go over all words in the document but randomly sample anchor words. As a result of this process, words that are often found close together in different documents (inside a given window size) will be more likely to fall in the same topic.

5 Evaluation

In our experiments, we have used a dataset of 101481 blog posts from the *LiveJournal* blog platform with 172939 unique words in total; *LiveJournal* is a platform of choice for topic modeling experiments since the posts are both user-generated and much longer than a typical tweet or *facebook* post. We have trained six baseline models and several varieties of GLDA:

(1) the basic probabilistic latent semantic analysis model (pLSA);
(2) ARTM model with Φ sparsity regularizer;
(3) ARTM model with Θ sparsity regularizer;
(4) basic LDA model with inference based on Gibbs sampling [11];
(5) basic LDA model with inference based on the variational Bayes [4];
(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 [5, 14, 22];
(7) granulated LDA with three different windows: step, Epanechnikov, and triangular, and different window sizes, from $l = 1$ to $l = 3$;

In all cases, we have trained the models with $T = 200$ topics. Note that we train LDA with two different inference algorithms 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 $\beta = 0.5$, values that have been previously tuned for our datasets [14]. 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*: useful topics have to be readily interpretable. For evaluation, we use the *coherence* and *tf-idf coherence* metrics. Coherence has been proposed as a topic quality metric in [8, 19]. For a topic t characterized by its set of top words W_t, coherence is defined as $c(t, W_t) = \sum_{w_1,w_2 \in W_t} \log \frac{d(w_1,w_2)+\epsilon}{d(w_1)}$, where $d(w_i)$ is the number of documents that contain w_i, $d(w_i, w_j)$ is the number of documents where w_i and w_j cooccur, and ϵ is a smoothing count usually set to either 1 or 0.01. A recent work [22] proposed a modification of the coherence metric called *tf-idf coherence*:

$$c_{\text{tf-idf}}(t, W_t) = \sum_{w_1,w_2 \in W_t} \log \frac{\sum_{d: w_1,w_2 \in d} \text{tf-idf}(w_1, d)\text{tf-idf}(w_2, d) + \epsilon}{\sum_{d: w_1 \in d} \text{tf-idf}(w_1, d)},$$

where the tf-idf metric is computed with augmented frequency,

Topic model	Topic quality metrics		Topic stability metrics	
	coherence	tf-idf coherence	stable topics	Jaccard
pLSA	-238.522	-126.934	54	0.47
pLSA + Φ sparsity reg.	-231.639	-127.018	9	0.44
PLSA + Θ sparsity reg.	-241.221	-125.979	87	0.47
LDA, Gibbs sampling	-208.548	-116.821	77	0.56
LDA, variational Bayes	-275.898	-112.544	111	0.53
SLDA	-208.508	-120.702	84	0.62
GLDA, step window, $l = 1$	-180.248	-123.231	195	0.64
GLDA, step window, $l = 2$	-171.038	-122.029	195	0.71
GLDA, step window, $l = 3$	-164.573	-121.582	197	0.73
GLDA, Epanechnikov window, $l = 1$	-226.394	-148.725	184	0.23
GLDA, Epanechnikov window, $l = 2$	-227.099	-174.475	192	0.33
GLDA, Epanechnikov window, $l = 3$	-206.347	-171.155	199	0.20
GLDA, triangular window, $l = 1$	-226.486	-148.147	162	0.16
GLDA, triangular window, $l = 2$	-234.096	-186.294	200	0.30
GLDA, triangular window, $l = 3$	-222.487	-184.187	200	0.68

Table 1: Overall metrics of topic quality and stability for granulated LDA and other models averaged over all runs of the corresponding model.

$$\text{tf-idf}(w,d) = \text{tf}(w,d) \times \text{idf}(w) = \left(\frac{1}{2} + \frac{f(w,d)}{\max_{w' \in d} f(w',d)}\right) \log \frac{|D|}{|\{d \in D : w \in d\}|},$$

where $f(w,d)$ is the number of occurrences of term w in document d. This skews the metric towards topics with high tf-idf scores in top words, since the numerator of the coherence fraction has quadratic dependence on the tf-idf scores and the denominator only linear. We have used both coherence and tf-idf coherence to evaluate topic quality in our solutions.

To evaluate topic stability, we have used the following approach. First, we introduce two natural similarity metrics for two topics [14]: symmetric Kullback–Leibler divergence between the probability distributions of two topics in a solution, defined as $\text{KL}(\phi^1, \phi^2) = \frac{1}{2}\sum_w \phi^1_w \log \frac{\phi^1_w}{\phi^2_w} + \frac{1}{2}\sum_w \phi^2_w \log \frac{\phi^2_w}{\phi^1_w}$, together with its normalized version [14] $\text{NKLS}(t_1, t_2) = 1 - \frac{\text{KL}(t_1,t_2)}{\max_{t'_1,t'_2} \text{KL}(t'_1,t'_2)}$, and Jaccard similarity of two sets of top words in two topics: for a given threshold T, we denote by Top^T_ϕ the set of T words with largest probabilities in a topic distribution ϕ and compute $J^T(\phi_1, \phi_2) = \frac{|\text{Top}^T_{\phi_1} \cap \text{Top}^T_{\phi_1}|}{|\text{Top}^T_{\phi_1} \cup \text{Top}^T_{\phi_1}|}$. 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 [14].

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, averaging results over three runs. We see that granulated LDA with the step window produces topics that have quality matching that of baseline topic models or even exceeding it, but the other two windows, Epanechnikov and triangular, do not work nearly as well. One should be careful about using coherence to draw steadfast conclusions in this case, though, because granulated LDA naturally lends itself to optimizing coherence: it artificially sets words that

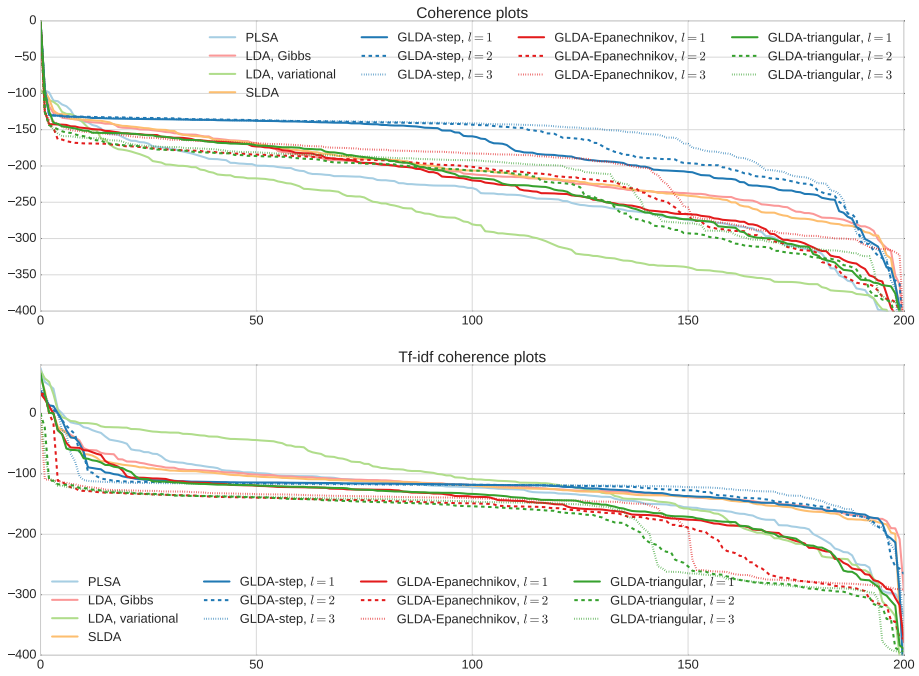

Fig. 2: Sorted topic quality metrics: coherence (top), tf-idf coherence (bottom).

cooccur in the same document (even in the same window) to the same topic. This effect is much less prominent for tf-idf coherence (many words in a window are likely to be common words with low tf-idf weights), and in tf-idf coherence we see GLDA with step window performing on par with other models. Figure 2 shows the distributions of coherence and tf-idf coherence metrics in more detail; namely, it shows the coherences (top) and tf-idf coherences (bottom) of all 200 topics for all models sorted in decreasing order, so a line higher on this plot means a better overall model. We can see that GLDA solutions, especially with the step window, hold up quite well compared with other models in our study.

The primary gains of our new model lie in topic stability. Table 1 shows the number of stable topics for every model and average Jaccard similarity (w.r.t. to 100 top words in each topic) between pairs of matching topics. We see that granulated LDA indeed produces very stable results: in all runs of granulated LDA with all window variants almost all topic were stable, and the average Jaccard similarity between them is also much higher than in other models in the case of a step window. Overall, we conclude that GLDA with step window produces much more stable topics at virtually no loss to quality and interpretability.

6 Conclusion

In this work, we have introduced a novel modification of the latent Dirichlet allocation model, granulated LDA, that samples whole windows of neighboring words in a document at once. This model was intended to improve the stability of the topic model results, and in the experimental evaluation we have shown that the results of GLDA are 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. In further work, we plan to extend and improve upon the basic idea of granulated LDA, experimenting with variations of this model. We hope that designing topic models with an eye to topic stability will prove to be a promising new venue of research.

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

References

1. D. Andrzejewski and X. Zhu. Latent Dirichlet allocation with topic-in-set knowledge. In *Proc. NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, pages 43–48, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
2. D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. In *Proc. 26th Annual International Conference on Machine Learning*, ICML '09, pages 25–32, New York, NY, USA, 2009. ACM.
3. D. M. Blei and J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18, 2006.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4–5):993–1022, 2003.
5. S. Bodrunova, S. Koltsov, O. Koltsova, S. I. Nikolenko, and A. Shimorina. Interval semi-supervised lda: Classifying needles in a haystack. In *Proc. 12th Mexican International Conference on Artificial Intelligence*, volume 8625 of *Lecture Notes in Computer Science*, pages 265–274. Springer, 2013.
6. G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, pages 31–40, 2013.
7. J. L. Boyd-Graber and D. M. Blei. Syntactic topic models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, pages 185–192. Curran Associates, Inc., 2008.
8. J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 20, 2009.
9. R.-C. Chen, R. Swanson, and A. S. Gordon. An adaptation of topic modeling to sentences. http://rueycheng.com/paper/adaptation.pdf, 2010.
10. X. Chen, M. Zhou, and L. Carin. The contextual focused topic model. In *Proceedings of the* 18^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 96–104, New York, NY, USA, 2012. ACM.

11. T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 (Suppl. 1):5228–5335, 2004.
12. J. Grimmer and B. M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013.
13. T. Hoffmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.
14. S. Koltcov, O. Koltsova, and S. I. Nikolenko. Latent dirichlet allocation: Stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science (WebSci 2014)*, pages 161–165, 2014.
15. S. Lacoste-Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems*, 20, 2008.
16. J. H. Lau, D. Newman, and T. Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*, pages 530–539, 2014.
17. S. Z. Li. *Markov Random Field Modeling in Image Analysis*. Advances in Pattern Recognition. Springer, Berlin Heidelberg, 2009.
18. D. A. McFarland, D. Ramage, J. Chuang, J. Heer, C. D. Manning, and D. Jurafsky. Differentiating language usage through topic models. *Poetics*, 41(6):607–625, 2013.
19. D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
20. D. Newman, E. V. Bonilla, and W. Buntine. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems 24*, pages 496–504. Curran Associates, Inc., 2011.
21. D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
22. S. I. Nikolenko, O. Koltsova, and S. Koltsov. Topic modelling for qualitative studies. *Journal of Information Science*, 2015.
23. D. Ramage, E. Rosen, J. Chuang, C. D. Manning, and D. A. McFarland. Topic modeling for the social sciences. In *NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, December 2009.
24. A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems*. W.H. Winston, 1977.
25. K. Vorontsov. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304, 2014.
26. K. V. Vorontsov and A. A. Potapenko. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323, 2015.
27. C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *Proceedings of the* 24^{th} *Conference on Uncertainty in Artificial Intelligence*, 2008.
28. X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the* 12^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, 2006.
29. J. Yohan and O. A. H. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824, New York, NY, USA, 2011.