Finding a Complete Set of Topics Using Initialization, Regularization & Multiple Model Training

Vasiliy Alekseev, Konstantin Vorontsov
wasya.alekseev@gmail.com, k.v.vorontsov@phystech.edu

Abstract

To fully investigate a text collection, one may have to train a big number of topic models in search of a good one. In this article we are dealing with the problem of the effective and complete finding of topics in a collection of text documents. We propose to conduct the experiment in such a way, so as to gradually accumulate good topics during multiple model training. We show, that using topic interpretability estimation measures, non-random topic model initialization and regularization it is possible to speed up the process of making a complete set of topics.

Key words: topic modeling, topic, topic interpretability, topic quality measures, topic coherence, topic purity, instability of topic models, full set of topics, initialization of topic model, regularization topic models, text analysis, BigARTM, CDC, Arora.

1 Introduction

Topic modeling — is an automatic way to analyze text document collections, which aims at finding *topics*, which are covered in the collection of text documents. Thus, topics are hidden. Moreover, the very concept of a *topic* can be defined in different ways depending on the task at hand. In statistical topic modeling, each topic is considered as a probability distribution over all words in vocabulary. Nowadays topic modeling is used in various fields, for example, in document categorization (Rubin et al., 2012), exploratory search (Ianina et al., 2017), social network analysis (Varshney et al., 2014).

The ideas and hypotheses adopted in topic modeling ultimately allow to reduce the original problem of finding topics in documents to the matrix decomposition problem, which is solved by an iterative algorithm. However the matrix decomposition problem is *incorrectly posed*: it has infinitely many solutions.

The result of the iterative algorithm depends on the initialization of the model, so the solution is also *unstable* (Steyvers and Griffiths, 2007). Some topics may be similar for many topic models with different initializations, some require certain initialization of topic model, and some topics may be uninterpretable (Ianina et al., 2017). Many researchers are working on evaluation of topic model stability (De Waal and Barnard, 2008; Koltcov et al., 2014; Greene et al., 2014).

So the question is, if there is a complete set of interpretable topics for the text collection. And how to find such topics set if exists. In (Balagopalan, 2012) the authors propose training several topic models with different initializations, and then clustering topics of all models to merge similar ones. And the centers of the resulting topic clusters may be chosen as initial approximation of topics of topic model to be trained. It is hypothesized that the size of topic cluster (that is, how often the topic was found by different topic models) is the larger, the more often the topic is covered in documents of the collection.

In work (Koltcov et al., 2016) the attention is drawn to the fact that words that are often found close to each other in text should relate to similar topics. This assumption is consistent with the hypothesis of topic segment structure of natural language text, according to which the words of the topics are distributed throughout the text not by chance, mixed with the words of other topics, but by groups, segments (Alekseev et al., 2018).

In (Vorontsov and Potapenko, 2015) the authors propose an approach to train topic models, called Additive Regularization of Topic Models — ARTM — which flexibility allows to implement in topic models various properties: regularizers allow to reduce the possible set of solutions of the matrix decomposition problem to those solutions that satisfy certain conditions. For example, with the help of regularizers, one can want a model to have topics that differ from one another, or that each topic has only a small number of its most probable words, or, conversely, so that some topics have as many words as possible — so called background topics, contain-

ing words of general vocabulary. Regularization is used to obtain a solution with desired properties, and may also help to increase the stability of model topics.

The following questions are addressed in this paper:

- How many models required to find a complete set of topics?
- How to minimize the number of required models?
- How to select good topics automatically from each model?

2 Topic Bank

As already noted, due to instability and incompleteness of topic models, such task as *researching data*, searching for topics in a collection of documents, can take a lot of time: in addition to step of preprocessing data for models to work (which cannot be avoided), one have to train several topic models, select parameters, evaluating the quality of final topics. Once a decent model is obtained, even then it is not guaranteed that its topics form a complete set.

Ideally, one wishes to train *just one model* with all its topics being good and different and which fits the data. Unfortunately, this is not yet possible. However, it is possible at least partially *to speed up, make more efficient and organized* the process of collecting good topics.

For more convenient work with topics of models it is suggested to use *topic bank*: good topics gradually accumulate in the bank, until together they form a complete set of topics. The main purpose of topic bank is minimizing the number of models that need to be trained to obtain a complete set of topics.

Let us explain what is meant by the term *complete set of topics*. Not in the form of definition, but rather in the form of properties, which are expected from complete set of topics. Topics from complete set

- collectively comprise such matrix decomposition that maximizes the likelihood of the topic model (that is, together topics make up a complete model that describes the data well)
- are interpretable
- are diverse

In this paper, a linear relationship between the topics of the complete set is allowed. So the diversification requirement may not be partially implemented. Maintaining a complete set in a linearly independent state seems to be the topic of separate research studies.

Even if model training process is managed to be built so that the result models are good, due to the incompleteness of topic models one still need many model training iterations and a selection of good topics from each model to make a complete set. Some of model topics can be uninterpretable. Furthermore, topics vary from model to model (both good and bad). In this work the assumption is made that *by multiple model training it is possible to obtain a complete set of good topics*, where complete set definition is given above.

Topic bank is like a *wrapper* over topic modeling, which allows to reduce the number of trained models to build a complete set of topics.

Gradual selection of topics can also help in the task of determining the number of topics in the collection: when one can no longer add a good topic in the bank with increasing the overall topic set likelihood. Even if the topic is good, but is a duplicate of a good topic in the bank, then adding it to the topic bank will not increase its likelihood (or more precisely likelihood of the model which topics are exactly those stored in the bank). Thus, topic bank can help both in finding good topics and in determining their number.

Also it is proposed to apply initialization and regularization of topic models to further optimize the number of training stages. Ideally, the correct initialization and regularization should help train a model which topics comprise a complete set.

Initialization of models, increasing the quality of final topics

Since model instability is caused by random initialization, one of the ways to increase stability is to make initialization *meaningful*, that is to choose an initial approximation for the model better than random one.

For example, in the works (Arora et al., 2012a,b) the authors introduce the concept of *anchor words*: words that can be used to immediately decide whether document belongs to some particular topic or not, that is, anchor words belong to only one topic. The requirement of topics having anchor words imposes additional restrictions on the matrix decomposition problem. However, it was shown (Arora et al., 2013) that the task of finding anchor words is easier than the matrix decomposition problem. The anchor words allow to achieve a local minimum of the topic modeling optimization problem in just a few iterations.

It is worth noticing that that not all topics can have anchor words: a single word may not be enough to define a topic, especially when it comes to topics with a large number of general vocabulary words or parent and child topics. To use such a model as the *initialization* matrix Φ seems to make sense, because such matrices will contain at least *part of the desired structure*, information

about the topics of the collection. It is also worth noting that the definition of an anchor word can be approached in different ways. So, one can assume that each topic can have only one anchor word or that there can be several of them for each topic.

In the work (Dobrynin et al., 2004) authors offer a different approach. The concept of *context of a word* is introduced, as words that are often found together with a given word not far from each other in text. Such a concept is based on the following hypothesis: words that most accurately characterize a topic are usually found together in text, their relative positions are not random. This allows one to search for the initial approximation of Φ as follows: split the documents of the original collection D into segments (for example, paragraphs or sentences). Estimate the probabilities of joint close encounters $p(w_1 \mid w_2)$ in the text for all pairs of words, select among all words those that occur quite often together with *only a small number* of other words and to cluster the probability vectors of joint encounters of such words. If each topic is represented in the text by segments, then as a result of clustering the selected words should be combined into one cluster — the topic. And the center of this cluster can approximate the topic as a column in matrix Φ.

3 Topic Modeling

Topic model can be represented as

$$p(w \mid d) = \sum_t p(w \mid t)p(t \mid d) = \sum_t \phi_{wt}\theta_{td}$$

A topic in topic modeling can be thought of as a distribution on a set of words $p(w \mid t)$, $w \in W$. Moreover, each topic is also characterized by its distribution on a set of documents: $p(t \mid d)$, $d \in D$.

So each trained topic model can be described using distributions Φ and Θ:

$$\Phi \equiv (\phi(w \mid t))_{W\times T} \equiv (\phi_{wt})_{W\times T}$$

$$\Theta \equiv (\theta(t \mid d))_{T\times D} \equiv (\theta_{td})_{T\times D}$$

The matrices Φ and Θ stochastic: their columns are non-negative, representing discrete distributions. They are found together by solving the matrix decomposition problem of the known matrix of word frequencies in documents.

There are different kinds of topic models. One of the very first, and at the same time one of the simplest and most understandable — the PLSA (Hoffman, 1999) model — maximizes the likelihood of the collection (more precisely, the log likelihood $\ln p(\Phi, \Theta)$)

$$\mathcal{L}(\Phi, \Theta) \equiv \sum_{d\in D}\sum_{w\in W_d} n_{wd} \log \sum_t \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

The point of local extremum of the formulated problem satisfies a system of equations that can be solved by iterative methods, updating Φ and Θ (Vorontsov et al., 2015) at each iteration. In this case, one need to *initialize* the matrix Φ — so the method of successful initialization is a separate issue.

In the optimized function one can include members that make the result topics satisfy some additional requirements. For example, one can require the model to have its topics very different from each other, or so that the words are distributed unevenly in the topics (making the most frequent words of the topic stand out more against the background words). The additive regularization approach for topic models (Vorontsov and Potapenko, 2014; Vorontsov et al., 2015; Vorontsov and Potapenko, 2015) implements the described idea — imposing additional restrictions on the resulting models by introducing additional regularization terms with non-negative weights τ_i into the optimized functional:

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

4 Methods of Evaluation Topic and Model Quality

As an intrinsic way to evaluate the quality of the model as a whole, one can use perplexity, which is closely connected with likelihood of the model $\mathcal{L}(\Phi, \Theta)$

$$\text{Perplexity} = e^{-\mathcal{L}(\Phi,\Theta)}$$

The higher the likelihood of the model, the lower the perplexity, and vice versa.

Let us introduce some of the methods to evaluate the quality of individual topics.

Purity — a possible ways to evaluate the quality of a topic based on the information in Φ matrix (Vorontsov and Potapenko, 2015).

$$\text{Purity}(t \mid \text{threshold}) = \sum_{w\in W_t} p(w \mid t)$$

Where $W_t = \{w \in W \mid p(t \mid w) > \text{threshold}\}$ — topic kernel.

In the works (Newman et al., 2010; Mimno et al., 2011; Lau et al., 2014) authors proposed a method for evaluating the quality of topics called *coherence*: when the decision about quality of a topic is made on the basis of how often word pairs of topic most common words appear close to each other in text (compared with the number of times when one and the other word is found in the text, but not necessarily close to each other). The mathematical expression for the introduced concept is the following:

$$\text{coh}(D, W, \phi_t \mid k) = \underset{w_i,w_j \in \text{ k top words}}{\text{Average}} \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$$

where $p(w_i, w_j)$, $p(w_i)$ — probabilities of meeting the word w_i or two words together w_i, w_j in a window of some size in the text. Probabilities are estimated using known word frequencies in documents. In general, coherence — is an attempt to automate the human way of assessing the quality of a topic. Authors proved that it correlates well with human judgements about topic quality.

In (Alekseev et al., 2018) the authors proposed an approach to quality evaluation of topics called *intra-text coherence*. When assessing the quality of the topic, distribution of words in the overall collection is taken into account, not only most probable words of the topic. The idea of such an approach is based on the hypothesis of the segment structure of natural language text: when topics are presented in the text as segments, rather than words arranged randomly.

The method of assessing the quality of topics exclusively by a small number of the most frequent words of the topic (*top words*) has an attractive side: for example, the speed of counting (although one still have to view the entire text of the collection). The minus of the top-word approach is that reducing the topic to only a small number of words leads to the fact that a large amount of information about the topic is not used in assessing the quality. The differences in approaches between coherences are discussed in more detail in (Alekseev et al., 2018).

5 Experiments

5.1 On Model Data

The idea is to check if it is possible to restore all original topics of the dataset using multiple model training.

Synthetic dataset. Model training

Dataset is created in such a way: 10 topics, each with 10 words with equal probabilities. There are no common words for topics (which with nonzero probabilities refer to several topics). 100 documents are created for each topic (the main topic with a probability of 0.8 and several other topics). On each run, a topic model with 5 topics is trained over 10 iterations with some initialization of the word probability matrix in topics.

Results

A topic was considered to be found by a model if 7 out of 10 top words and the first 2 top words of this topic are correct. The word order in the topic was not taken into account, because the words in the original synthetic topics were equally probable.

At the end of 24^{th} model training, *all original topics were found* (in the sence given above).

Moreover, the topics were found at different frequencies: one topic was found by 22 models, several topics were found only by 1 model. On average, one topic was found by 6 models. Out of the 5 topics of each model, on average 2 of them turned out to be good.

5.2 On Real Data

Here, the main question under concern is whether it is possible to speed up the process of collecting complete set of topics using topic bank.

Dataset

Dataset which is used in the experiment — a collection of popular science articles "PostNauka"[1] (in Russian): 3446 documents, several dozens of topics (exactly 19). The number of topics is approximately known, because the articles are divided into sections. Topics have different number of articles.

Experiment

Initially, topic bank is empty. On the described dataset, topic models are trained. The model is initialized (either all topics are initialized randomly, or the initial approximation for half of the topics is found using Arora or CDC methods). Each model is trained during 20 iterations. There are 100 topics in each model. Likelihood, and all topic quality measures are calculated. Next among the topics of the trained model those ones are chosen, whose quality is not less than the 90 percentile among the qualities of all the topics of the model. Selected topics are considered as good. Then among the selected good topics new ones are found, that is those to which there are no close topics in the bank yet. The distance between the topics was calculated according to the Jacquard measure, and during the comparison only words with probability more than uniform $p(w \mid t) > \frac{1}{W}$ were taken into account. The threshold was set equal to 0.5: if the distance between model topic and the nearest topic in the bank is not less than the threshold, then the topic is considered new. All new good topics are added to the bank.

This was the process of selecting good topics in order to create a complete topic set.

Results

The results of the experiment are presented on figures 1a, 1b, 1c, 2a, 2b, 2c.

Let us explain plot identificators on figures 1a, 1b. Name *plain bank* means the set of topics stored in topic bank when there is no regularization and initialization is random during the training process, *one model* means the average value of perplexity for range of trained models with no regularization and with random initialization, *reg Decorr* represents the bank state if trained models are also regularized with with topic decorrelation regularizer, *reg Complex* represents bank creation with two regularizers while training: decorrelation and smoothness of topics, *init Arora* means that instead of random initialization Arora algorithm is used to initialize half of the topics in each trained

[1] postnauka.ru

model, *init CDC* means initialization of half of each model's topics using CDC algorithm. So, looking at the plots 1a, 1b one can see that automatic topic quality evaluation together with proper regularization can boost the process of finding the complete topic set: making topic bank with regularization allowed to get better overall likelihood compared to complete set collected without regularization. And the value of likelihood converged faster: it took fewer models to get better result.

Looking at 1b, 1c, one can deduce, that using multiple model training with good topics selection it is possible to determine an approximate number of topics in text collection. When perplexity stabilizes 1b it means that no more new topics can be found. So the training process can be stopped at that point. Using the number of models at stopping point on 1b one can see what number of topics corresponds to this number on 1c: this is the approximate number of topics in the complete set. Thus, the number of topics appeared to be in range $[50, 150]$. This is not an accurate estimate, but in order of magnitude, the value coincides with the number of topics in the dataset 19.

Plots 2a, 2b, 2c show that topics in the complete set actually do have better values of top-tokens and intra-text coherence compared to topics in trained models.

6 Conclusion

The work explored how to improve the quality of topic models using multiple model training. It is shown that methods for evaluating the quality of topics help to create a complete set of topics whose likelihood is comparable, albeit less, to the likelihood of an ordinary model. It is also shown that the use of regularizers in training process helps to reduce the number of trained models compared to training without regularizers.

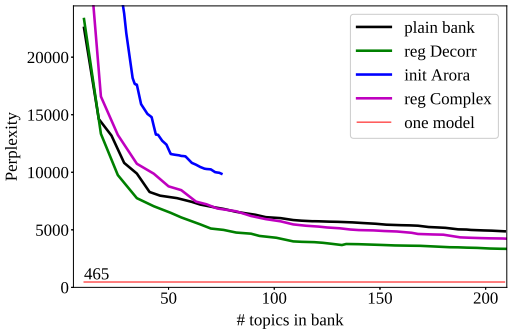
Finding complete set of topics also helps to determine the approximate number of topics in text collection.

The question still remains how to build a model in which all the topics are good, different, and more good topics can no longer be found. There is no ready-made recipe yet, and this is a topic for further research.

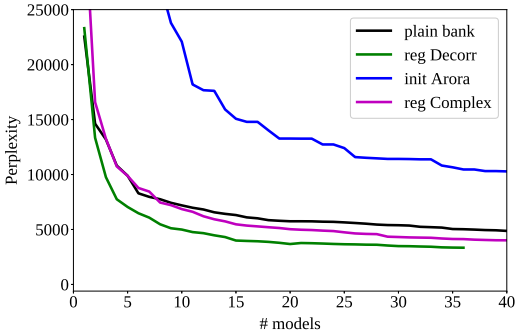
References

Vasiliy Alekseev, Victor Bulatov, and Konstantin Vorontsov. 2018. Intra-text coherence as a measure of topic models' interpretability. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue*, pages 1–13.

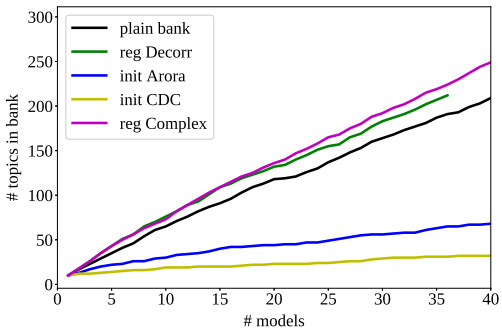
Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen



(a) Perplexity of the model depending on number of topics added to topic bank. Line representing initialization using CDC is not showed, because its perplexity is too high compared to the presented plots.

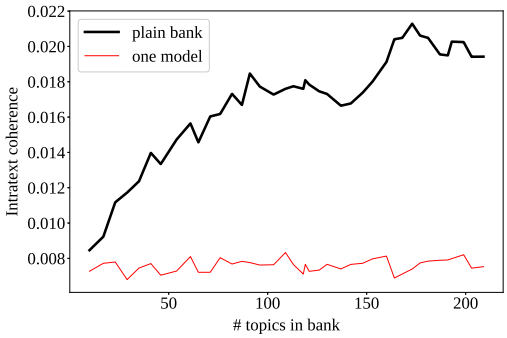


(b) Perplexity of the model depending on number of models used to find the complete set of topics. Regularization may help to converge faster

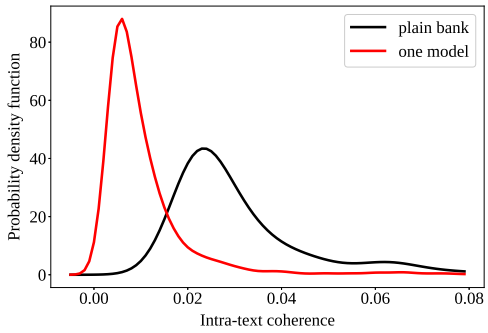


(c) Number of topics in bank depending on number of models used to find the complete set of topics. It is growing steadily, but the perplexity on the above plot stabilizes, which means, that although the bank is getting bigger, it doesn't receive new topics anymore.

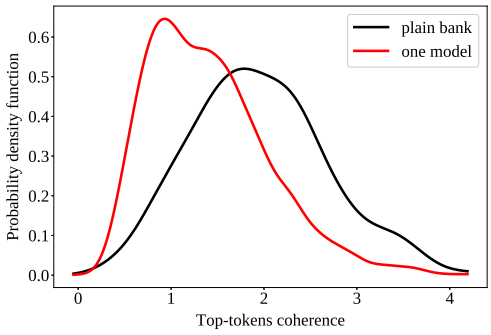
Figure 1: Topic bank creation using multiple model training. Good topics are selected from each trained model and added to the bank. Thus, topic bank is growing gradually.



(a) Intratext coherence depending on number of topics in bank. Complete set of topics shows higher values of coherence comparing ordinary models.



(b) KDE estimation of intra-text coherence values.



(c) KDE estimation of top-tokens coherence values.

Figure 2: Lower two plots show KDE estimation of intra-text and top-tokens coherence values of topics in bank and in models being trained. Average value of coherence among bank topics is higher

Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. 2012a. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM.

Sanjeev Arora, Rong Ge, and Ankur Moitra. 2012b. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE.

Arun Balagopalan. 2012. *Improving topic reproducibility in topic models*. University of California, Irvine.

Alta De Waal and Etienne Barnard. 2008. Evaluating topic models with stability.

Vladimir Dobrynin, David Patterson, and Niall Rooney. 2004. Contextual document clustering. In *European Conference on Information Retrieval*, pages 167–180. Springer.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 498–513. Springer.

Thomas Hoffman. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, 1999*, pages 50–57.

Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. 2017. Multi-objective topic modeling for exploratory search in tech news. In *Conference on Artificial Intelligence and Natural Language*, pages 181–193. Springer.

Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science*, pages 161–165. ACM.

Sergei Koltcov, Sergey I Nikolenko, Olessia Koltsova, Vladimir Filippov, and Svetlana Bodrunova. 2016. Stable topic modeling with local density regularization. In *International Conference on Internet Science*, pages 176–188. Springer.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Devesh Varshney, Sandeep Kumar, and Vineet Gupta. 2014. Modeling information diffusion in social networks using latent topic information. In *International Conference on Intelligent Computing*, pages 137–148. Springer.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer.

Konstantin Vorontsov and Anna Potapenko. 2014. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 29–46. Springer.

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