Four Keys to Topic Interpretability in Topic Modeling

Andrey Mavrin[1], Andrey Filchenkov[1(✉)], and Sergei Koltcov[2]

[1] ITMO University, St. Petersburg, Russia
andreyshambala@gmail.com, afilchenkov@corp.ifmo.ru
[2] National Research University Higher School of Economics, St. Petersburg, Russia
skoltsov@hse.ru

Abstract. Interpretability of topics built by topic modeling is an important issue for researchers applying this technique. We suggest a new interpretability score, which we select from an interpretability score parametric space defined by four components: a splitting method, a probability estimation method, a confirmation measure and an aggregation function. We designed a regularizer for topic modeling representing this score. The resulting topic modeling method shows significant superiority to all analogs in reflecting human assessments of topic interpretability.

Keywords: Topic modeling
Additive regularization for topic modeling · Topic interpretability
Human assessment

1 Introduction

Topic modeling is a domain of machine learning that has been actively developing since the late 1990s. Its main goal is to determine given a set of text documents, to which topics each document relates, as well as what terms each topic consists of. Topic modeling allows effectively solving of such tasks as clustering and classification of text documents [19], topical search of documents and related objects [17], building of topical profiles of users of various Internet resources [9], analysis of news flows [11] and many others.

In many cases, in the above-mentioned areas of topic modeling application requires a person to interact directly with the topic model. In these cases, the concept of "topic" has to correspond to the human notion of it. In particular, words that form a specific topic must be semantically related. The task of assessing the topic interpretability in topic models has been actively studied since the end of 2010, when the methods of expert assessment of interpretability [4] were first proposed, and later interpretability scores were suggested [1,12,14].

The goal of this work is improving interpretability of topics. To do so, we use additive regularization for topic modeling (ARTM) approach by proposing a regularizer that supports topic interpretability. For this purpose, we explore

D. Ustalov et al. (Eds.): AINL 2018, CCIS 930, pp. 117–129, 2018.
https://doi.org/10.1007/978-3-030-01204-5_12

interpretability scores in an interpretability score parametric space and find the one, which is the best to reflect human assessment of topic interpretability.

The rest of the paper is structured as follows. In Sect. 2, we briefly describe ARTM approach and several regularizers, with which we will compare our work. In Sect. 3, we describe the parametric space of interpretability score, as well as present a regularizer corresponding to such space. In Sect. 4 we briefly describe details of the method implementation and experimental evaluation. Results and their discussion is presented in Sect. 5. Section 6 concludes.

2 Related Work

2.1 Topic Modeling and Additive Regularization

The probabilistic topic model (TM) of a document collection is a set of topics, each of which is a probability distribution on the set of words encountered in the collection, and a set of probability distributions on a set of topics for each document [20]. Since the notation in topic modeling domain has not been changed during recent years, and the size of paper is limited, we will skip the notation assuming that a reader is familiar with it. We will follow [5,21,22].

Many approaches for topic modeling were suggested: Latent Semantic Analysis (LSA) [16], Probabilistic Latent Semantic Analysis (PLSA) [8], Latent Dirichlet Allocation (LDA) [2]. They were generalized under an approach suggested in 2014 by Konstantin Vorontsov [21] called additive regularization of topic models (ARTM). The main idea of this approach is to maximize model likelihood jointly with additional criteria called regularizers that represent additional constraints.

2.2 Topic Interpretability

Interpretability of topics obtained as the result of topic modeling began to be actively considered in 2009, when a method for assessing the interpretability of the topic by a person called word intrusion was proposed [4]. Intuitively, the assessment of topic interpretability is whether a person can understand how the words representing a topic are related to each other and what is a general concept to which they relate. The word intrusion method evaluating of the topic interpretability by a respondent is as follows. Each topic is presented in the form of six words, five of which are the most probable words in the topic, and the sixth word is chosen randomly from words in this topic having a low probability. The task of the respondent is to correctly determine the intruder. The interpretability of the topic is estimated by the number of respondents who found the intruder.

Due to the assessing of the topic interpretability is a very expensive and time-consuming procedure, it would be desirable to be able to evaluate interpretability without human participation. Researchers have suggested several scores for estimating the topic interpretability discussed below.

Pointwise Mutual Information. Idea of this score (more well-known as UCI) [13] is to assess the topic interpretability by associating all the pairs of

words in a topic. Such association is estimated on some large external corpus. It is assumed that the topic is represented by the ten most likely words in this topic. The formula of the topic interpretability is as follows:

$$PMIScore(w) = \text{median}\{PMI(w_i, w_j), i, j \in \{1..10\}\},$$

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)},$$

where $p(w)$ is word probability estimated on an external corpus, $p(w_i, w_j)$ is a joint probability of a pair of words estimated with a sliding window of size 10 scanning the external corpus.

UMass. This score [12] is quite similar to the UCI, however in this case the function estimating the association between a pair of words is not symmetric. In addition, it does not use external corpus, evaluating the coherence of words in the collection of documents on which TM was built. It is also assumed that a topic is described with M of its most probable words. It is defined as follows:

$$C(t, V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})},$$

where $D(v)$ is frequency of word v among the documents, $D(v, v')$ is the joint frequency of pair (v, v') among documents, $V^{(t)} = (v_1^{(t)}, \ldots, v_M^{(t)})$ is a list of M most probable words in topic t. The unit in the numerator under the logarithm prevents the value under the logarithm from being converted to zero.

Context Vectors. The main idea of this score [1] is usage of vector representation of words in the subject. It is also assumed that a topic is represented with n most probable words. The proposed score is defined as:

$$Coherence_{Sim}(T) = \frac{\sum_{1 \le i \le n-1, i+1 \le j \le n} Sim(\boldsymbol{w}_i, \boldsymbol{w}_j)}{\binom{n}{2}}$$

$$Sim(\boldsymbol{w}_i, \boldsymbol{w}_j) = \frac{\boldsymbol{w}_i \cdot \boldsymbol{w}_j}{||\boldsymbol{w}_i|| \cdot ||\boldsymbol{w}_j||},$$

where $\boldsymbol{w}_i$ is a vector representation of word $w_i \in T$. The vector representation is learned on an external corpus with so-called word context, which is defined as 10 words closest to each of the word occurrences into the outer body (5 on each side). Thus, every occurrence of word w in the chosen external corpus results in 10 new components in the vector representation of w. Value of w component associated with word f is evaluated as $PMI(w, f)^{\gamma}$, where γ is the parameter that makes the components of the vector with a high value to be more meaningful.

However, it is easy to see that with this approach the dimensionality of the vectors turns out to be too large, therefore it is suggested to limit the dimension

by choosing only β_{w_i} of the most connected (maximal) components, where β_{w_i} is computed using the following formula [10]:

$$\beta_{w_i} = (\log(c(w_i)))^2 \cdot \frac{\log_2(m)}{\delta},$$

where δ is a regularization coefficient and m is the external corpus size.

2.3 Interpretability in ARTM

A regularizer for ARTM is known, which directly maximizes the coherence between words in a topic [22]. It uses a previously computed matrix of connectivity between the words C, where C_{uv} is the joint estimate of pair of words $(u, v) \in Q \subset W^2$. This regularizer, which minimizes the sum of divergences between each distribution of ϕ_{vt} and its estimate for all words that occur with v, looks like this:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{(u,v) \in Q} C_{uv} n_{ut} \ln \phi_{vt} \to \max.$$

However, application of this regularizer meets some difficulties. Given a sufficiently large volume of the collection, on which topic model is built, it is not possible to evaluate the joint occurrence for each pair of words in the collection due to the very large size of the set of all pairs of words. A choice of some subset of pairs of words must have some logical justification, which also causes difficulties. This is why we did not include this approach in comparison.

Next modification is *word embedding coherence* (WEC) [15], which is:

$$\mathrm{Coh}_{\mathrm{we}}(t) = \frac{1}{n(n-1)} \sum_{\tilde{w_i}^{(t)} \neq \tilde{w_j}^{(t)}} d(v(\tilde{w_i}^{(t)}), v(\tilde{w_j}^{(t)})),$$

where $v : \mathcal{W} \to \mathbb{R}^d$ is a mapping from tokens to d-dimensional vectors and $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a distance function.

3 Interpretability Scores and an Additive Regularizer

First, we describe a parametric space of interpretability scores, in which we will further search for the best metric. Second, we present a new regularizer for TMs, which maximizes the interpretability of the main words in topics.

3.1 Parametric Space of Interpretability Scores

We assume that a score estimating the quality of topic interpretability can be represented in the form of four relatively independent components [18] $(\mathscr{S}, \mathscr{P}, \mathscr{C}, \mathscr{A})$, which will be described in detail later. The input of a interpretability score are a topic and n of the most probable words $W =$

$\{w_1, w_2, \ldots, w_n\}$. The first component $\mathscr{S}$ of the score is a method of **s**plitting the most probable words into pairs (W', W^*), where $W', W^* \subset W$. The second component $\mathscr{P}$ is a method of estimating word **p**robability, which is a function $P : W' \to [0, 1]$. It is computed using an external collection of documents, which differs from the one, on which the topic model is built. Intuitive requirement for this collection is a presence of large amount of non-specific information. An example of such a collection is a set of Wikipedia articles. The third component $\mathscr{C}$ of the score is function $C : (W', W^*) \to \mathbb{R}$, which is the so-called **c**onfirmation measure. It shows how much the subset W' supports the subset W^*. The fourth component $\mathscr{A}$ is an **a**ggregating function that converts a set of real numbers into single real number.

Thus, the whole process of computing interpretability score of a topic can be described as follows. First, topic W is split into a set of pairs $\{(W, W^*)\}$ by means of $\mathscr{S}$. Then for each pair from the resulting set, confirmation measure $\mathscr{C}$ is computed using $\mathscr{P}$. Finally, the set of real numbers obtained with $\mathscr{C}$ is transformed by means of $\mathscr{A}$ into a single real number, which represents the quality of the topic interpretability. The scheme for evaluating interpretability in the manner described above is presented in Fig. 1.

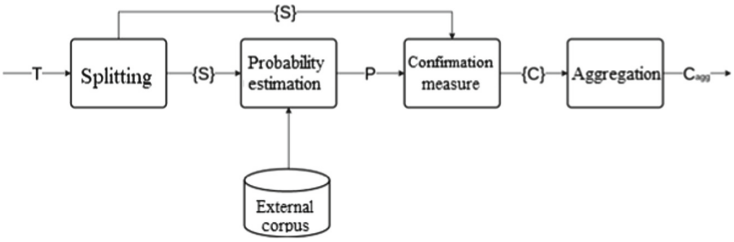

Fig. 1. Scheme of computing an interpretability score

Splitting Method $\mathscr{S}$. To estimate the interpretability of a topic, the set of words representing the topic is divided into pairs, for which their probabilistic "compatibility" is estimated. The most straightforward way of splitting is the simple principle of "every word with every other", S^{one}_{one}. This splitting is used, for example, for the UMI and UMass measures. Further options for splitting include those, in which each word is combined only with each subsequent or with each previous one, S^{one}_{suc} and S^{one}_{pre}. A smarter way of splitting is not only into pairs of single words, but also using subsets of more than one element [6]: we use one versus all other S^{one}_{all}, one versus some subset of other words S^{one}_{any} and two non-intersecting subsets of words S^{any}_{any}.

Probability Estimation Method $\mathscr{P}$. This component determines how the probability of a word is estimated by the external collection of documents. The simplest estimation method, which is used, for example, in the UMass metric, is a method called a "boolean document". The probability of a word is estimated

as the number n_w of documents, in which this word occurs, divided by the total number n of documents in the collection. It is worth noting that such estimate of the probability does not take into account the distance between occurrences of words, but only the fact of their appearance in the document.

An alternative approach is the so-called "sliding window". The idea is that a window of fixed size n moves through the external collection of documents. The probability of word w in this case is the number of steps on which w was in the window divided by the total number of steps. In this case, the distance between several words in the text matters, when joint probability is estimated. We will choose from windows of sizes 10, 50, 100 and 200.

Confirmation Measure $\mathscr{C}$. A confirmation measure receives a pair of topic most probable word subsets and uses a probability estimate method considered earlier to calculate how much one subset of the pair is associated with the other. The options, which will be used as elements of the component in the parametric score space, are presented in Table 1.

Table 1. Confirmation measures

Measure	Formula
difference, $\mathscr{C}_d$	$P(W'\|W^*) - P(W')$
ratio, $\mathscr{C}_r$	$\frac{P(W',W^*)}{P(W')P(W^*)}$
log-ration, $\mathscr{C}_{lr}$	$\log \frac{P(W',W^*)+\epsilon}{P(W')P(W^*)}$
normalized log-ratio, $\mathscr{C}_{nlr}$	$\frac{m_{lr}(W',W^*)}{-\log(P(W',W^*)+\epsilon)}$
likelihood, $\mathscr{C}_l$	$\frac{P(W'\|W^*)}{P(W'\|\neg W^*)+\epsilon}$
log-likelihood, $\mathscr{C}_{ll}$	$\log \frac{P(W'\|W^*)+\epsilon}{P(W'\|\neg W^*)+\epsilon}$
conditional, $\mathscr{C}_c$	$\frac{P(W',W^*)}{P(W^*)}$
logarithmic conditional, $\mathscr{C}_{lc}$	$\log \frac{P(W',W^*)+\epsilon}{P(W^*)}$
Jaccard, $\mathscr{C}_j$	$\frac{P(W',W^*)}{P(W'\vee W^*)}$
logarithmic Jaccard, $\mathscr{C}_{lj}$	$\log \frac{P(W',W^*)+\epsilon}{P(W'\vee W^*)}$
Fitelson [7], $\mathscr{C}_f$	$\frac{P(W'\|W^*)-P(W'\|\neg W^*)}{P(W'\|W^*)+P(W'\|\neg W^*)}$

Aggregation Function $\mathscr{A}$. As an aggregation function, we take the arithmetic mean $\mathscr{A}_{am}$, median $\mathscr{A}_{med}$, the geometric mean $\mathscr{A}_{gm}$ and the harmonic mean $\mathscr{A}_{hm}$.

3.2 Regularizer for ARTM

In this Subsection, we describe a new regularizer for ARTM, adding of which will lead to maximizing of a score from the parametric space, maximizing thus the interpretability of the topics.

First, recall that the problem of maximizing the interpretability of topics stands first of all if a person needs to interact directly with a topic model and analyze it. When a person interacts with topics, it is incontinent for a person to consider a topic as a distribution on the whole set of words. Topic models are usually built for large collections of documents, and the dictionary of such collections is so large that a person is not able to process it visually, let alone process a certain number of probability distributions in this dictionary. In this case, the common practice is to present a topic in the form of n of the most probable words. Most often, n is assumed to be 10. Thus, the main idea of the proposed regularizer is to optimize the quality of interpretability of exactly the ten most probable words in the topic.

Let $Top_t = \{w_1, w_2, \ldots, w_{10}\}$ be the ten most probable words of topic t, and $C(u, v)$ be the adjusted confirmation measure for the pair words (u, v), taken from the parametric space of interpretable scores. Then the regularizer, which for each word v from Top_t minimizes the sum of divergences between the distribution of ϕ_{vt} and the confirmation measure for all the remaining words from Top_t, looks like this:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{(u,v) \in Top_t^2} C(u, v) \hat{n}_{ut} \ln \phi_{vt} \to \max,$$

where τ is a regularization coefficient. Further, the resulting regularizer casts the following modified formula for M-step in EM-algorithm:

$$\begin{cases} \phi_{wt} \propto \hat{n}_{wt} + \tau \sum_{v \in Top_t \setminus w} C(w, v) \hat{n}_{vt} & \text{if } w \in Top_t, \\ \phi_{wt} \propto \hat{n}_{wt} & \text{otherwise.} \end{cases}$$

We must note that this regularizer is in general a modification of coherence [12] with specified $C(u, v)$. The confirmation measure must be adjusted with a constant so that its values are to some degree symmetric with respect to zero. That is, for poorly connected words the measure should take negative values, and for well-connected words the value should be positive. Then the presented regularizer can be understood as follows: on each iteration of the algorithm, for each of the ten most probable words of topic t, its relative interpretability to other words from Top_t is estimated; if word w is semantically well-connected with the remaining words $v \in Top_t$ and $C(w, v)$ is positive for the most words, then the probability estimate ϕ_{wt} increases and improves w probability estimation, allowing it to remain in Top_t; if word w is badly related to the rest of $v \in Top_t$, then most values of $C(w, v)$ take negative values, and then the probability estimation ϕ_{wt} decreases, which is likely to cause word w to be excluded from the most probable words of t.

4 Experiment Setup

4.1 Finding Best Interpretability Score in the Parametric Space

Data Labeling. We obtain the topics, on which the score quality is evaluated, using various methods for building topic models, namely PLSA, LDA, and

ARTM. We learn them on a collection of documents representing posts on the blog platform LiveJournal (in Russian). The resulting set counts 1200 topics. Each of the topics is presented with its ten most probable words.

The obtained topics were demonstrated to two assessors. We ask them to estimate each topic, answering two questions. The first question is "Do you understand why these words turned out to be together in this topic?". The second question is "Do you understand what kind of event or phenomenon of life can be discussed in the texts on this topic?". Each answer should be an integer from 0 to 2, where 0 stands for "no", 1 stands for "partly", and 2 stands for "yes". After that each topic was estimated with the mean of two answers.

External Corpus for Learning $\mathscr{P}$. In order to learn word probability estimates, we use an external corpus, which is a collection of approximately 1.5 million preprocessed articles of the Russian Wikipedia. First, XML tags and punctuation were removed. Further, all the words were lemmatized by the means of pymorphy2. After that, stop words, numerals, English words, Roman numerals, service parts of speech were removed. Finally, due to the resulting collection was quite large, the index of this collection was built using the Apache Lucene library to improve the speed of work.

Selection Criteria. To assess the quality of the interpretability score, the Spearman rank correlation coefficient between the scores values and the respondents' answers is used. To ensure that the difference between the mean values of the expert estimates is not random, we used Student's t-test.

Experiment Pipeline. We examine each point of the parametric space, with which we estimated each of 1200 topics. We use Java 8 and Palmetto library [18], which implements many elements from which the parametric space components were composed.

4.2 Comparing Topic Models

Document Collection. We use the following document collections: (1) papers presented at conference "Intellectual Data Processing" in various years; (2) articles published in the newspaper "Izvestia" in 1997; (3) text corpus that was labeled within the project OpenCorpora [3].

Each collection was preprocessed in the same way as the external corpus, described in the previous Subsection. Two TMs are built for each of the collections. The first TM is built using such regularizers as the topic rarefaction, blurring of background topics and decorrelation of subject topics. The second TM is build using the very same regularizers and a new regularizer proposed in this work. Each of the six topic models consists of one hundred subject topics and ten background topics.

Topic Model Assessment Criteria. To evaluate how the adding of the new regularizer improves the quality of the interpretability of topics, we invited three assessors who estimated each topic in the way described in the previous Subsection. The most common criterion for the quality of TMs is perplexity, which

characterizes the discrepancy between the model $p(w|d)$ for word w observed in documents $d \in D$ and is determined through the log likelihood as follows:

$$\mathscr{P}(D;p) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} \ln p(w|d)\right).$$

Also, following the assumption that the topics contain a relatively small number of words from the collection dictionary, and a relatively small number of topics are represented in the documents from the collection, the sparsity of the matrices Φ and Θ is used as an important characteristic of topic models.

Implementation Details. We used BigARTM to implement the TM additive regularization. The source code of this library was supplemented with the regularizer, described in Sect. 3.

5 Results

5.1 Comparison of Interpretability Scores

As a result of the experiment, we found that the highest Spearman coefficient in the interpretability score parametric space was shown by the following score: $(\mathscr{S}^{one}_{one}, \mathscr{P}_{sw(200)}, \mathscr{C}_d, \mathscr{A}_{am},)$ where $\mathscr{S}^{one}_{one}$ is splitting "each word with each other", $\mathscr{P}_{sw(200)}$ is sliding window probability estimation of size 200 words, $\mathscr{C}_d(W', W^*) = P(W'|W^*) - P(W')$, and $\mathscr{A}_{am}$ is the arithmetical mean. It is important to note that in order to use this score in the regularizer presented in Sect. 3, no additional regulation of the confirmation measure by means of some scalars, due to $\mathscr{C}_d$ takes values in range $[-1, 1]$.

We present comparison of the Spearman correlation coefficient (SCC) between the human assessments and all the discussed scores in the Table 2. One can see that the score of the parametric space is superior to all the presented analogs. From this we conclude that the selected interpretability score models human interpretability assessment better than the known scores.

Table 2. Comparison of interpretability scores

Score	SCC	Score	SCC
UCI	0.44538	Context vectors	0.62002
UMass	0.54474	WEC	0.50074
NPMI	0.53320	**ParamSpace**	**0.70330**

As a result, we found a score that maximizes the SCC, which outperforms other scores, and now we can use it for topic modeling.

5.2 Comparison of Topic Model Interpretability and Quality

Comparison Using the Human Assessments. In Table 3 one can see the arithmetic mean of human assessments for each of the built TMs. It is easy to see that addition of the regularizer has significantly improved the interpretability of topics for all collections of documents.

Table 3. Human assessment of the topic models

Collection	Without the regularizer	With the regularizer
ISP	2.357	2.490
Izvestia	2.503	2.863
OpenCorpora	1.950	2.183

The value of Student's t-test was 4.705, which exceeded the value of Student's distribution (2.59) at 299° of freedom and significance level 0.01, which allows rejecting the hypothesis about the equality of mean values.

Comparison Using Topic Model Quality Measures. Figures. 2, 3 and 4 show how the perplexity of TMs has been changing on each step of the EM algorithm. Blue is used by the TMs built using the proposed regularizer, red is used by the TMs built without it. The perplexity of TMs with the regularizer turned out to be noticeably higher, which may indicate that the proposed regularizer worsens the quality of TMs. However, it 2009 in one of the first articles devoted to the interpretability of TMs [4], authors showed that when the value of perplexity is high enough, perplexity and human interpretability assessments are directly dependent. In particular, it was shown that when the perplexity of a TM is reduced, human's interpretability assessments are also reduced. This corresponds to our experiment results described above.

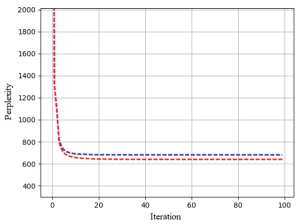

Fig. 2. Perplexity of TMs on IDP corpus

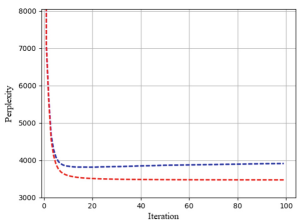

Fig. 3. Perplexity of TMs on Izvestia corpus

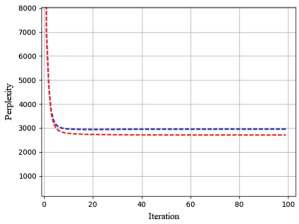

Fig. 4. Perplexity of TMs on OpenCorpora

Figures 5, 6 and 7 show how the sparsity of Φ has been changing during the EM-iterations of the algorithm. It is easy to see that the proposed regularizer somewhat worsens the sparsity of Φ, which looks logical enough given the

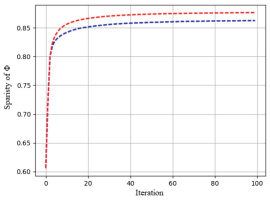

Fig. 5. Sparsity of Θ of TM on IDP corpus

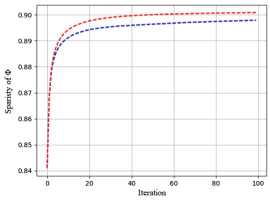

Fig. 6. Sparsity of Θ of TMs on Izvestia corpus

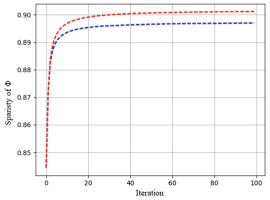

Fig. 7. Sparsity of Φ of TMs on OpenCorpora

structure of the proposed regularizer. However, it caused only a small decrease of the sparseness of Φ (by no more than 2%), which prevents stating that the introduced regularizer significantly worsened the quality of TMs.

The change in the sparsity of Θ during the iterations of the EM algorithm can be traced on Figs. 8, 9 and 10. Interestingly, the introduction of the proposed regularizer somewhat improved the sparsity of the Θ for each of the collections, but not so much as to say that the number of zero elements in the Θ matrix became comparatively large.

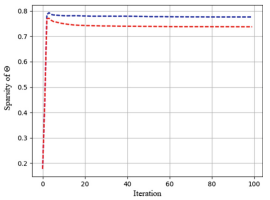

Fig. 8. Sparsity of Θ of TMs on IDP corpus

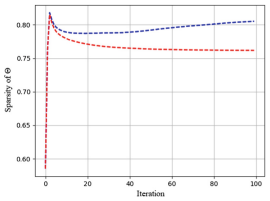

Fig. 9. Sparsity of Θ of TMs on Izvestia corpus

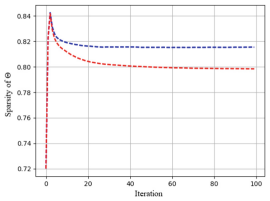

Fig. 10. Sparsity of Θ of TMs on OpenCorpora

To summarize, the addition of the proposed regularizer did not decreased the quality of TMs, but significantly increased their interpretability.

6 Conclusion

In this paper, we found the best interpretability score in an interpretability score parametric space composed of four components. Basing on this score, we proposed a regularizer for ARTM, which being added is capable of building interpretable topic models. The experiments showed that a topic model with the proposed regularizer significantly outperforms topic models without it having comparable results in term of topic model quality.

As a development of this work, one can consider, for example, the improvement of the semantic similarity of documents belonging to the same topic.

Acknowledgments. Authors would like to thank Anton Belyy and Konstantin Vorontsov for useful conversation. Andrey Mavrin and Andrey Filchenkov were supported by the Government of the Russian Federation (Grant 08-08). Sergei Koltsov was supported by the Basic Research Program at the National Research University Higher School of Economics (HSE).

References

1. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)-Long Papers, pp. 13–22 (2013)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
3. Bocharov, V., Bichineva, S., Granovsky, D., Ostapuk, N., Stepanova, M.: Quality assurance tools in the OpenCorpora project (2011)
4. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296 (2009)
5. Daud, A., Li, J., Zhou, L., Muhammad, F.: Knowledge discovery through directed probabilistic topic models: a survey. Front. Comput. Sci. China **4**(2), 280–301 (2010)
6. Douven, I., Meijs, W.: Measuring coherence. Synthese **156**(3), 405–425 (2007)
7. Fitelson, B.: A probabilistic theory of coherence. Analysis **63**(3), 194–199 (2003)
8. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
9. Hong, L., Davison, B.D.: Empirical study of topic modeling in Twitter. In: Proceedings of the First Workshop on Social Media Analytics, SOMA 2010, pp. 80–88. ACM (2010)
10. Islam, A., Inkpen, D.: Second order co-occurrence PMI for determining the semantic similarity of words. In: Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy, pp. 1033–1038. Citeseer (2006)
11. Jacobi, C., van Atteveldt, W., Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. Digit. Journalism **4**(1), 89–106 (2016)
12. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272. Association for Computational Linguistics (2011)
13. Newman, D., Karimi, S., Cavedon, L.: External evaluation of topic models. In: 2009 Australasian Document Computing Symposium. Citeseer (2009)
14. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics (2010)
15. Nikolenko, S.I.: Topic quality metrics based on distributed word representations. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1029–1032. ACM (2016)

16. Papadimitriou, C.H., Tamaki, H., Raghavan, P., Vempala, S.: Latent semantic indexing: a probabilistic analysis. In: Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 159–168. ACM (1998)
17. Perkio, J., Buntine, W., Perttu, S.: Exploring independent trends in a topic-based search engine. In: 2004 Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2004, pp. 664–668, September 2004
18. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 399–408. ACM (2015)
19. Rubin, T.N., Chambers, A., Smyth, P., Steyvers, M.: Statistical topic models for multi-label document classification. Mach. Learn. **88**, 157–208 (2012)
20. Steyvers, M., Griffiths, T.: Probabilistic topic models. Handb. Latent Semant. Anal. **427**(7), 424–440 (2007)
21. Vorontsov, K., Frei, O., Apishev, M., Romov, P., Dudarenko, M.: BigARTM: open source library for regularized multimodal topic modeling of large collections. In: Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) AIST 2015. CCIS, vol. 542, pp. 370–381. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26123-2_36
22. Vorontsov, K., Potapenko, A.: Tutorial on probabilistic topic modeling: additive regularization for stochastic matrix factorization. In: Ignatov, D.I., Khachay, M.Y., Panchenko, A., Konstantinova, N., Yavorskiy, R.E. (eds.) AIST 2014. CCIS, vol. 436, pp. 29–46. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12580-0_3