Renormalization Approach to the Task of Determining the Number of Topics in Topic Modeling

Sergei Koltcov and Vera Ignatenko(✉)

National Research University Higher School of Economics, 55/2 Sedova Street, St. Petersburg, Russia 192148
{skoltsov,vignatenko}@hse.ru

Abstract. Topic modeling is a widely used approach for clustering text documents, however, it possesses a set of parameters that must be determined by a user, for example, the number of topics. In this paper, we propose a novel approach for fast approximation of the optimal topic number that corresponds well to human judgment. Our method combines the renormalization theory and the Renyi entropy approach. The main advantage of this method is computational speed which is crucial when dealing with big data. We apply our method to Latent Dirichlet Allocation model with Gibbs sampling procedure and test our approach on two datasets in different languages. Numerical results and comparison of computational speed demonstrate a significant gain in time with respect to standard grid search methods.

Keywords: Renormalization theory · Optimal number of topics

1 Introduction

Nowadays, one of the widely used instruments for analysis of large textual collections is probabilistic topic modeling (TM). However, when using topic modeling in practice, the problems of selecting the number of topics and values of hyperparameters of the model arise since these values are not known in advance by practitioners in most applications, for instance, in many tasks of sociological research. Also, the results of TM are significantly influenced by the number of topics and inappropriate hyperparameters may lead to unstable topics or to topic compositions that do not accurately reflect the topic diversity in the data. The existing methods to deal with this problem are based on grid search. For instance, one can use standard metrics such as log-likelihood [1] or perplexity [2] and calculate the values of these metrics for different values of model parameters and then choose the parameters which lead to the best values of considered metrics. Another popular metric is semantic (topic) coherence [3]. A user has to select the number of most probable words in topic to be used for topic coherence calculation, then topic coherence is calculated for individual topics. Let us note

K. Arai et al. (Eds.): SAI 2020, AISC 1228, pp. 234–247, 2020.
https://doi.org/10.1007/978-3-030-52249-0_16

that there is no clear criterion for selecting the number of words and the authors [3] propose to consider 5–20 terms. Values of individual topic coherence are then aggregated to obtain a single coherence score [4,5]. After that one can apply a grid search for determining the best values of model parameters with respect to the coherence score. However, the above methods are extremely time-consuming for big data which is why optimization of the procedure of topic number selection is of importance.

The computational complexity of the existing grid-search-based methods calls for greedy solutions that can speed up the process without substantial loss of TM quality. In this work, we propose a significantly faster solution for an approximation of an optimal number of topics for a given collection. We refer to the number of topics determined by encoders as the 'optimal number' of topics. Our approach is based on renormalization theory and on entropic approach [7], which, in turn, is based on the search for a minimum Renyi entropy under variation of the number of topics. Details of the entropic approach are described in Subsect. 2.3. The author of [7] demonstrated that the minimum point of Renyi entropy lies approximately in the region of the number of topics identified by users. This approach also requires a grid search for model optimization, but the search itself is optimized based on the previous research and theoretical considerations. In work [8], it was demonstrated that the density-of-states function (to be defined further) inside individual TM solutions with different topic numbers is self-similar in relatively large intervals of the number of topics, and such intervals are multiple. Based on these facts and taking into account that big data allow applying methods of statistical physics, we conclude that it is possible to apply the renormalization theory for fast approximation of the optimal number of topics for large text collections. This means that calculation reduction in our approach is based on the mentioned self-similarity. We test our approach on two datasets in English and Russian languages and demonstrate that it allows us to quickly locate the approximate value of the optimal number of topics. While on the dataset consisting of 8,624 documents our approach takes eight minutes, the standard grid search takes about an hour and a half. Therefore, for huge datasets gain in time can vary from days to months. We should especially note that renormalization-based methods are suitable for finding approximate values of T only. However, the exact value can be found afterwards by grid search on a significantly smaller set of topic solutions, which compensates for the approximate character of the renormalization-based search.

Our paper consists of the following sections. Subsection 2.1 describes basic assumptions of probabilistic topic modeling and formulation of the task of topic modeling. Subsection 2.2 gives an idea of renormalization theory which is widely used in physics. Subsection 2.3 reviews Renyi entropy approach which was proposed in [7,9]. Subsection 2.4 describes findings of work [8] which are necessary for the application of renormalization theory to topic modeling. Section 3 describes the main ideas of our approach and its application to Latent Dirichlet Allocation model (LDA). Section 4 contains numerical experiments on the renor-

malization of topic models and comparison of obtained approximations of the optimal number of topics to the ground truth. Section 5 summarizes our findings.

2 Background

2.1 Basics of Topic Modeling

Topic modeling takes a special place among machine learning methods since this class of models can effectively process huge data sets. In the framework of TM, several assumptions are expected to be met. First, the dataset contains a fixed number of topics. It means that a large matrix of occurrences of words in documents can be represented as a product of two matrices of smaller size which represent the distribution of words by topics and distribution of topics by documents, correspondingly. Second, documents and words are the only observable variables. Hidden distributions are calculated based on these variables. Thus, a document collection can be characterized by three numbers: D, W, T, where D is the number of documents, W is the number of unique words in the dataset, T is the number of topics, which is usually selected by users of TM. Third, currently, TM is constructed on the basis of the 'bag of words' concept. It means that topic models do not take into account the order of words in documents. Thus, probability of a word w in a document d can be written in the following form [10,11]: $p(w|d) = \sum_t p(w|t)p(t|d) \equiv \sum_t \phi_{wt}\theta_{td}$, where $\{p(w|t) \equiv \phi_{wt}\}$ refers to the distribution of words by topics, $\{p(t|d) \equiv \theta_{td}\}$ is the distribution of topics by documents. A more detailed description of TM formalism can be found in [7,9]. In fact, finding the hidden distributions in large text collection is equivalent to understanding what people write about without reading a huge number of texts, that is, to identifying topics that are discussed in the collection.

2.2 Basics of Renormalization Theory

Renormalization is a mathematical formalism that is widely used in different fields of physics, such as percolation analysis and phase transition analysis. The goal of renormalization is to construct a procedure for changing the scale of the system under which the behavior of the system preserves. Theoretical foundations of renormalization were laid in works [12,13]. Renormalization was widely used and developed in fractal theory since fractal behavior possesses the property of self-similarity [14,15]. To start a brief description of renormalization theory, let us consider a lattice consisting of a set of nodes. Each node is characterized by its spin direction, or spin state. In turn, a spin can have one of many possible directions. Here, the number of directions is determined by a concrete task or a model. For example, in the Ising model, only two possible directions are considered; in the Potts model, the number of directions can be 3–5 [16]. Nodes with the same spin directions constitute clusters. The procedure of scaling or renormalization follows the block merge principle where several nearest nodes are replaced by one node. The direction of the new spin is determined by the direction of the majority of spins in the block. A block merge procedure is conducted

on the whole lattice. Correspondingly, we obtain a new configuration of spins. The procedure of renormalization can be conducted several times. Following the requirement of equivalence between the new and the previous spin configurations, it is possible to construct a procedure of calculation of parameters and values of critical exponents, as described in [17]. Let us note that consistent application of renormalization of the initial system leads to approximate results, however, despite this fact, this method is widely used since it allows to obtain estimations of critical exponents in phase transitions, where standard mathematical models are not suitable. Renormalization is applicable if scale invariance is observed. Scale invariance is a feature of power-law distributions. Mathematically, self-similarity (or scale invariance) is expressed in the following way. Assume that $f(x) = cx^{\alpha}$, where c, α are constants. If we transform $x \to \lambda x$ (it corresponds to scale transformation) then $f(\lambda x) = c(\lambda x)^{\alpha} := \beta x^{\alpha}$, where $\beta = c\lambda^{\alpha}$, i.e., scale transformation leads to the same original functional dependence but with a different coefficient. In concrete applications, the parameter of power-law, α, can be found by different algorithms, such as 'box counting' or others.

2.3 Entropy-Based Approach

The entropic approach for analysis of topic models was proposed in [7,9] and is based on a set of principles. A detailed discussion of these principles can be found in [7,9]. In this work, we would like to briefly discuss some important observations related to the entropic approach which would be necessary for the formulation of our renormalization procedure. First, a document collection is considered as a statistical system, for which the free energy can be determined. Let us note that free energy is equivalent to Kullback-Leibler divergence. Further, the free energy (and, correspondingly, Kullback-Leibler divergence) can be expressed in terms of Renyi entropy through partition function (Z_q) [9]:

$$Z_q = \rho(\tilde{P})^q,$$

where $q = 1/T$ is a deformation parameter,

$$\rho = N/(WT) \tag{1}$$

is the 'density-of-states' function of the whole topic solution, N is the number of highly probable words with $p(w|t) > 1/W$, $\tilde{P} = \sum_{w,t} p(w|t) \cdot \mathbb{1}_{\{p(w|t)>1/W\}}$ with $\mathbb{1}_{\{\cdot\}}$ being an indicator function. Thus, Renyi entropy of a topic solution can be expressed in the following form:

$$S_q^R = \frac{\ln(Z_q)}{q-1}. \tag{2}$$

We would like to notice that the above expression of Renyi entropy is in Beck notation [18].

Since the procedure of TM shifts the information system from a state of high entropy to a state of low entropy, the calculation of deformed Renyi entropy

after the TM allows estimating the effect of model hyperparameters and the number of topics on the results of TM. It was demonstrated [7,9] that minimum entropy corresponds to the number of topics which was selected by users in the process of dataset labeling. It allows us to link the procedure of searching for a minimum deformed entropy with the process of data labeling, which plays a crucial role in machine learning models. However, searching for the minimum Renyi entropy demands exhaustive search over the set of hyperparameters and numbers of topics which is a time-consuming process. A partial solution to this problem can be found through the analysis of self-similarity in topic solutions under variation of the number of topics.

2.4 Self-similar Behaviour in Topic Models

As it was shown in [8], topic models have the properties of self-similar behavior under variation of the number of topics. Such behavior is expressed in the fact that the 'density-of-states' function satisfies $\rho(\lambda\frac{1}{T}) = \beta(1/T)^{\alpha}$ with some β and α, and, therefore is linear in bi-logarithmic coordinates. However, such behavior is observed only in some ranges of the number of topics. Moreover, the inclination angles of linear pieces of the 'density-of-states' function are different in different regions that correspond to different fractal dimensions. The determination of the inclination angles was implemented according to the following steps: 1) Multidimensional space of words and topics is covered by a grid of fixed size (matrix $\Phi = \{\phi_{wt}\}$). 2) The number of cells satisfying $\phi_{wt} > 1/W$ is calculated. 3) The value of ρ for the fixed number of topics T is calculated according to Eq. (1). 4) Steps 1, 2, 3 are repeated with cell sizes (i.e. the number of topics) being changed. 5) A graph showing the dependence of ρ in bi-logarithmic coordinates is plotted. 6) Using the method of least squares, the slope of the curve on this plot is estimated, the value of the slope is equal to the value of fractal dimension calculated according to the following relation: $D = \frac{\ln(\rho)}{\ln(\frac{1}{T})}$.

In work [8], two datasets in different languages were tested under variation of the number of topics and it was demonstrated that there are large regions where the density-of-states function self-reproduces, i.e., fractal behavior is observed. Areas between such regions of self-similarity are transition regions. In such regions, change in the density-of-states function happens, i.e. the character of self-similarity changes. Work [8] demonstrates that transition regions correspond to human mark-up. Regions of self-similarity do not lead to changes in the structure of solutions of TM, therefore, it is sufficient to find transition regions in order to determine the optimal topic number in a collection. The disadvantage of this approach is in its computational complexity both in terms of time and computational resources: to find transition regions one needs to run topic modeling many times with multiple values of topic numbers. Since there are regions of self-similarity, we propose to apply renormalization theory to speed up the search for the topic number optimum.

3 Method

3.1 Application of Renormalization in Topic Modeling

In this subsection, we explain the main idea of renormalization for the task of topic modeling (its application for Latent Dirichlet Allocation model with Gibbs sampling procedure will be demonstrated in Subsect. 3.2). Recall that the output of TM contains matrix $\Phi = \{\phi_{wt}\}$ of size $W \times T$. Here, we consider a fixed vocabulary of unique words, therefore, the scale of renormalization depends only on parameter $q = 1/T$. Renormalization procedure is a procedure of merging two topics into one new topic. As a result of the merging procedure, we obtain a new topic $\tilde{t}$ with its topic-word distribution satisfying $\sum_w \phi_{w\tilde{t}} = 1$. Since the calculation of matrix Φ depends on a particular topic model, the mathematical formulation of renormalization procedure is model-dependent. Also, the results of merging depend on how topics for merging were selected. In this work, we consider three principles of selecting topics for merging:

- Similar topics. Similarity measure can be calculated according to Kullback-Leibler divergence [19]: $KL(t_1, t_2) = \sum_w \phi_{wt_1} \ln(\frac{\phi_{wt_1}}{\phi_{wt_2}}) = \sum_w \phi_{wt_1} \ln(\phi_{wt_1}) - - \sum_w \phi_{wt_1} \ln(\phi_{wt_2})$, where ϕ_{wt_1} and ϕ_{wt_2} are topic-word distributions, t_1 and t_2 are topics. Then two topics with the smallest value of KL divergence are chosen.
- Topics with the lowest Renyi entropy. Here, we calculate Renyi entropy for each topic individually according to Eq. (2), where only probabilities of words in one topic are used. Then we select a pair of topics with the smallest values of Renyi entropy. As large values of Renyi entropy correspond to the least informative topics, minimum values characterize the most informative topics. Thus, we choose informative topics for merging.
- Randomly chosen topics. Here, we generate two random numbers in the range $[1, \hat{T}]$, where $\hat{T}$ is the current number of topics, and merge topics with these numbers. This principle leads to the highest computational speed.

3.2 Renormalization for Latent Dirichlet Allocation Model

Let us consider Latent Dirichlet Allocation model with Gibbs sampling algorithm. This model assumes that word-topic and topic-document distributions are described by symmetric Dirichlet distributions with parameters α and β [20], correspondingly. Matrix Φ is estimated by means of Gibbs sampling algorithm. Here, values α and β are set by the user. Calculation of Φ consists of two phases. The first phase includes sampling and calculation of a counter c_{wt}, where c_{wt} is the number of times when word w is assigned to topic t. The second phase contains recalculation of Φ according to

$$\phi_{wt} = \frac{c_{wt} + \beta}{(\sum_w c_{wt}) + \beta W}. \tag{3}$$

For our task of renormalization, we use the values of counters c_{wt} and Eq. (3). Notice that counters c_{wt} form matrix $C = \{c_{wt}\}$, and this is the matrix which

undergoes renormalization. Based on matrix C, renormalized version of matrix Φ is then calculated. Algorithm of renormalization consists of the following steps:

1. We choose a pair of topics for merging according to one of the principles described in Subsect. 3.1. Let us denote the chosen pair of topics by t_1 and t_2.
2. Merging of selected topics. We aim to obtain the distribution of the new topic $\tilde{t}$ resulting from merging topics t_1 and t_2, which would satisfy Eq. (3). Merging for matrix C means summation of counters c_{wt_1} and c_{wt_2}, namely, $c_{w\tilde{t}} = c_{wt_1} + c_{wt_2}$. Then, based on new values of counters, we calculate $\phi_{w\tilde{t}}$ in the following way (analogous to Eq. (3)):

$$\phi_{w\tilde{t}} = \frac{c_{wt_1} + c_{wt_2} + \beta}{(\sum_w c_{wt_1} + c_{wt_2}) + \beta W}. \tag{4}$$

   One can easily see that new distribution $\phi_{.\tilde{t}}$ satisfies $\sum_w \phi_{w\tilde{t}} = 1$. Then, we replace column ϕ_{wt_1} by $\phi_{w\tilde{t}}$ and delete column ϕ_{wt_2} from matrix Φ. Note that this step leads to decreasing the number of topics by one topic, i.e., at the end of this step we have $T - 1$ topics.

Steps 1 and 2 are repeated until there are only two topics left. At the end of each step 2, we calculate Renyi entropy for the current matrix Φ according to Eq. (2). Then we plot Renyi entropy as a function of the number of topics and search for its minimum to determine the approximation of the optimal number of topics. Thus, our proposed method incorporates Renyi entropy-based approach and renormalization theory. Moreover, it does not require the calculation of many topic models with different topic numbers, but it only requires one topic solution with large enough T.

4 Numerical Experiments

For our numerical experiments, the following datasets were used:

- Russian dataset (RD) from the Lenta.ru news agency [21]. Each document of the dataset was assigned to one of ten topic classes by dataset provider. We consider a subset of this dataset which contains 8,624 documents with a total number of 23,297 unique words (available at [22]).
- English dataset (ED) is the well-known '20 Newsgroups' dataset [23]. It contains 15,404 English documents with the total number of 50,948 unique words. Each of the documents was assigned to one or more of 20 topic groups. Moreover, it was demonstrated [24] that 14–20 topics can represent this dataset.

These datasets were used for topic modeling in the range $[2, 100]$ topics in the increments of one topic. Hyperparameters of LDA model were fixed at the values: $\alpha = 0.1$, $\beta = 0.1$. Research on the optimal values of hyperparameters for these datasets was presented in work [9], therefore, we do not vary hyperparameters in our work. For both datasets, the topic solution on 100 topics underwent

renormalization with successive reduction of the number of topics to one topic. Based on the results of consecutive renormalization, curves of Renyi entropy were plotted as functions of the number of topics. Further, the obtained Renyi entropy curves were compared to the original Renyi entropy curves [7] obtained without renormalization.

4.1 Russian Dataset

Figure 1 demonstrates curves of Renyi entropy, where the original Renyi entropy curve was obtained by successive topic modeling with different topic numbers (black line) and the other Renyi entropy curves were obtained from five different runs of the same 100-topic model by means of renormalization with a random selection of topics for merging. Here and further, the minima are denoted by circles in the figures. The minimum of the original Renyi entropy corresponds to 8 topics, minima of renormalized Renyi entropy correspond to 12, 11, 11, 17 and 8 topics, depending on the run. Accordingly, the average minimum of five runs corresponds to 12 topics. As it is demonstrated in Fig. 1, renormalization with merging of random topics, on one hand, provides correct values of Renyi entropy on the boundaries, i.e., for $T = 2$ and $T = 100$, on the other hand, the minimum can fluctuate in the region $[8, 17]$ topics. However, on average, random merging leads to the result which is quite similar to that obtained without renormalization. Figure 2 demonstrates renormalized Renyi entropy based on merging topics with the lowest Renyi entropy. It can be seen that for this principle of selecting topics for merging, renormalized Renyi entropy curve is flat around its minimum (unlike the original Renyi entropy curve) which complicates fining this minimum. The flat area around the global minimum is located in the region of 10–18 topics. At the same time, at the endpoints of the considered range of topics the renormalized Renyi entropy curve has values similar to those of the original Renyi entropy, i.e. for $T = 2$ and $T = 100$.

Figure 3 demonstrates the behavior of renormalized Renyi entropy when the principle of selecting topics for merging is based on KL divergence. It shows that this principle leads to the worst result: the renormalized Renyi entropy curve has a minimum that does not correspond to the optimal number of topics. However, just like all other versions of renormalized entropies, it behaves "correctly" on the boundaries, i.e. it has maxima for $T = 2$ and $T = 100$.

4.2 English Dataset

The results obtained on this dataset are similar to those based on the Russian dataset. Figure 4 demonstrates five runs of renormalization with randomly selected topics for merging on the English dataset. One can see that the curves are very similar to each other and to the original Renyi entropy curve. The minimum of the original Renyi entropy corresponds to 14 topics, minima of renormalized Renyi entropy correspond to 17, 11, 14, 23 and 12 topics, depending on the run of renormalization. Accordingly, the average minimum of five runs corresponds to 15 topics. Figure 5 demonstrates the renormalized Renyi entropy

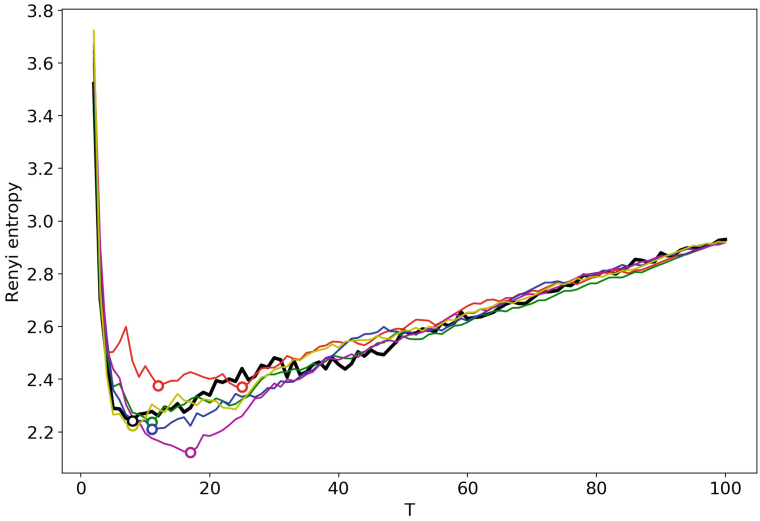

Fig. 1. Renyi entropy vs the number of topics T (RD). Original Renyi entropy – black. Renormalized Renyi entropy with random merging of topics: run 1 – red; run 2 – green; run 3 – blue; run 4 – magenta; run 5 – yellow. ()

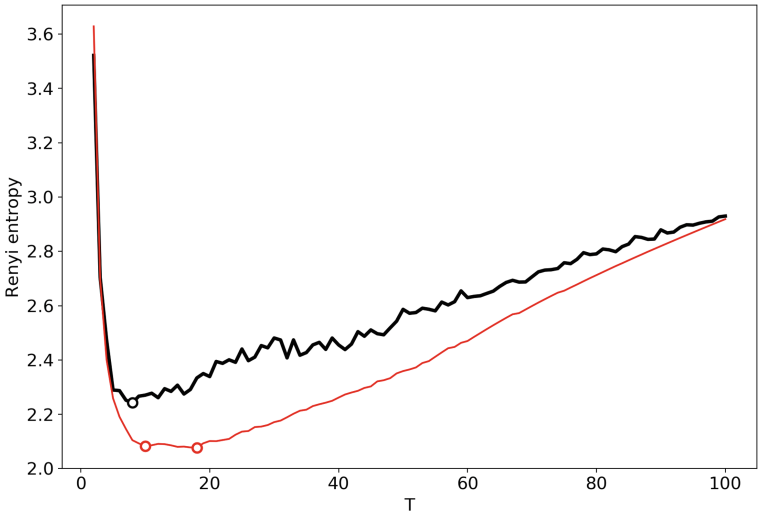

Fig. 2. Renyi entropy vs the number of topics T (RD). Original Renyi entropy – black; renormalized Renyi entropy (topics with the lowest Renyi entropy merged) – red.

curve, where topics with the lowest Renyi entropy were merged. The minimum of the renormalized entropy corresponds to 16 topics. On average, this type of renormalization leads to slightly lower values of Renyi entropy compared to the original Renyi entropy. Figure 6 demonstrates renormalized Renyi entropy,

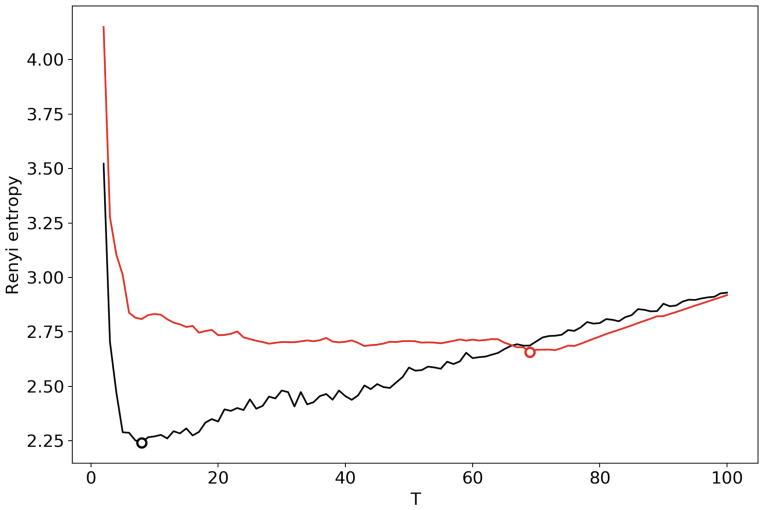

Fig. 3. Renyi entropy vs the number of topics T (RD). Original Renyi entropy – black; renormalized Renyi entropy (similar topics with the lowest KL divergence merged) – red.

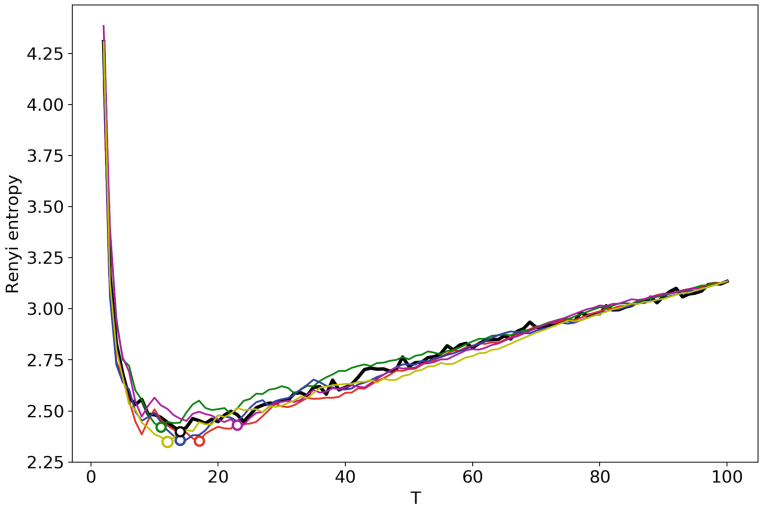

Fig. 4. Renyi entropy vs the number of topics T (ED). Original Renyi entropy – black. Renormalized Renyi entropy with random merging of topics: run 1 – red; run 2 – green; run 3 – blue; run 4 – magenta; run 5 – yellow.

where topics were merged based on KL divergence between them. Again, we can see that this type of merging leads to the worst result. The renormalized Renyi entropy has a minimum at $T = 43$ that does not correspond either to the human mark-up or to the minimum of the original Renyi entropy.

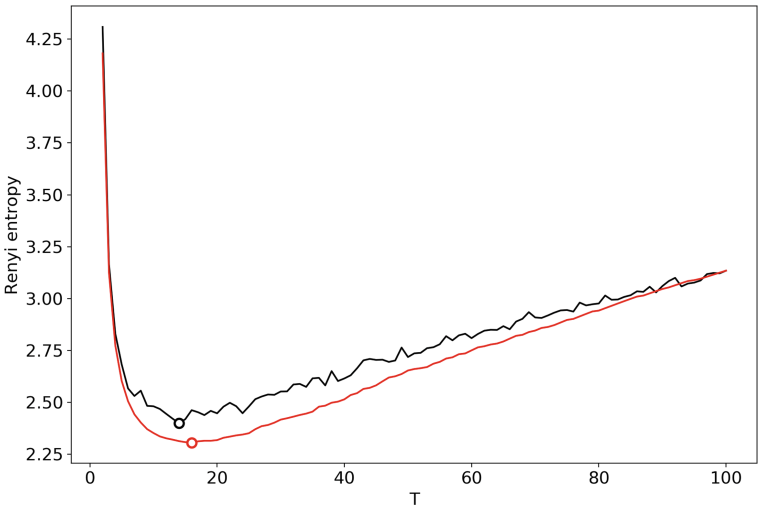

Fig. 5. Renyi entropy vs the number of topics T (ED). Original Renyi entropy – black; renormalized Renyi entropy (topics with the lowest Renyi entropy merged) – red.

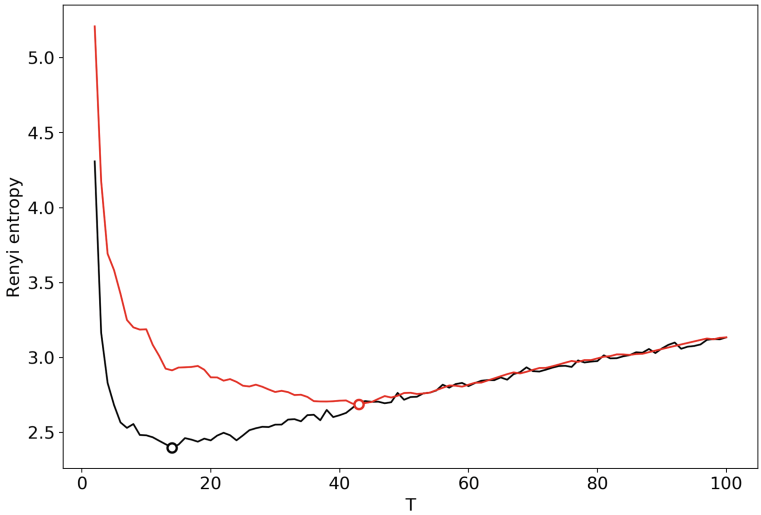

Fig. 6. Renyi entropy vs the number of topics T (ED). Original Renyi entropy – black; renormalized Renyi entropy (similar topics with the lowest KL divergence merged) – red.

4.3 Comparison of Computational Speed of Original and Renormalized Models

Table 1 demonstrates computational speed for a sequence of topic models and for renormalization. All calculations were performed on the following equipment:

Table 1. Computational speed.

Dataset	TM simulation and calculation of Renyi entropy	Renormalization (random)	Renormalization (minimum Renyi entropy)	Renormalization (minimum KL divergence)
Russian dataset	90 min	8 min	16 min	140 min
English dataset	240 min	23 min	42 min	480 min

notebook Asus, Intel Core I7 - 4720 HQ CPU 2.6 GHz, Ram 12 Gb, Operation system: Windows 10 (64 bits). Calculations on both datasets demonstrate that renormalization with randomly selected topics for merging is the fastest. Moreover, this type of renormalization leads to the most similar behavior of the renormalized Renyi entropy curve to the original Renyi entropy. Also, the computational speed for this type of renormalization is almost 11 times higher than that of the original Renyi entropy. Renormalization based on merging topics with the lowest KL divergence is the slowest: such calculation is even more time-consuming than regular grid-search calculation with a reasonable number of iterations. Renormalization in which topics with the lowest Renyi entropy are merged takes the second place: its computation is five times faster than that of the original Renyi entropy.

Summarizing the obtained results, we conclude that renormalization with randomly selected topics for merging could be an efficient instrument for the approximation of the optimal number of topics in document collections. However, it is worth mentioning that one should run such renormalization several times and average the obtained number of topics.

5 Conclusion

In this work, we have introduced renormalization of topic models as a method of fast approximate search for the optimal range of T in text collections, where T is the number of topics into which a topic modeling algorithm is supposed to cluster a given collection. This approach is introduced as an alternative to computationally intensive grid search technique which has to obtain solutions for all possible values of T in order to find the optimum of any metric being optimized (e.g. entropy). We have shown that, indeed, our approach allows to estimate the range of the optimal values of T for large collections faster than grid search and without substantial deviation from the "true" values of T, as determined by human mark-up.

We have also found out that some variants of our approach yield better results than others. Renormalization involves a procedure of merging groups of

topics, initially obtained with the excessive T, and the principle of selection of topics for merge has turned out to significantly affect the final results. In this work, we considered three different merge principles that selected: 1) topics with minimum Kullback-Leibler divergence, 2) topics with the lowest Renyi entropy, or 3) random topics. We have shown that the latter approach yielded the best results both in terms of computational speed and accuracy, while Renyi-based selection produced an inconvenient wide flat region around the minimum, and the KL-based approach worked slower than non-renormalized calculation. Since on our collections, random merge produced speed gain of more than one hour, corpora with millions of documents are expected to benefit much more, in the numbers amounting to hundreds of hours.

A limitation of the renormalization approach is that it is model-dependent, i.e. the procedure of merge of selected topics depends on the model with which the initial topic solution was obtained. However, although we have tested our approach on topic models with Gibbs sampling procedure only, there seem to be no theoretical obstacles for applying it to other topic models, including the Expectation-Maximization algorithm. This appears to be a promising direction for future research deserving a separate paper.

Acknowledgments. The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2019.

References

1. Wallach, H.M., Mimno, D., McCallum, A.: Rethinking LDA: why priors matter. In: Proceedings of the 22nd International Conference on Neural Information Processing Systems, pp. 1973–1981. Curran Associates Inc., USA (2009)
2. Manning, C.D., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
3. 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, Stroudsburg (2011)
4. Röder, M., Both, A., Hinneburg, A.: Exploring the space of topic coherence measures. In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 399–408. ACM, New York (2015)
5. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 952–961. Association for Computational Linguistics, Stroudsburg (2012)
6. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006). https://doi.org/10.1198/016214506000000302
7. Koltsov, S.: Application of Rényi and Tsallis entropies to topic modeling optimization. Phys. A **512**, 1192–1204 (2018). https://doi.org/10.1016/j.physa.2018.08.050

8. Ignatenko, V., Koltcov, S., Staab, S., Boukhers, Z.: Fractal approach for determining the optimal number of topics in the field of topic modeling. J. Phys: Conf. Ser. **1163**, 012025 (2019). https://doi.org/10.1088/1742-6596/1163/1/012025
9. Koltsov, S., Ignatenko, V., Koltsova, O.: Estimating topic modeling performance with Sharma-Mittal entropy. Entropy **21**(7), 1–29 (2019). https://doi.org/10.3390/e21070660
10. 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, New York (1999)
11. Vorontsov, K., Potapenko, A.: Additive regularization of topic models. Mach. Learn. **101**, 303–323 (2015). https://doi.org/10.1007/s10994-014-5476-6
12. Kadanoff, L.P.: Statistical Physics: Statics. Dynamics and Renormalization. World Scientific, Singapore (2000)
13. Wilson, K.G.: Renormalization group and critical phenomena. I renormalization group and the Kadanoff scaling picture. Phys. Rev. B **4**(9), 3174–3183 (1971). https://doi.org/10.1103/PhysRevB.4.3174
14. Olemskoi, A.I.: Synergetics of Complex Systems: Phenomenology and Statistical Theory. Krasand, Moscow (2009)
15. Carpinteri, A., Chiaia, B.: Multifractal nature of concrete fracture surfaces and size effects on nominal fracture energy. Mater. Struct. **28**(8), 435–443 (1995). https://doi.org/10.1007/BF02473162
16. Essam, J.W.: Potts models, percolation, and duality. J. Math. Phys. **20**(8), 1769–1773 (1979). https://doi.org/10.1063/1.524264
17. Wilson, K.G., Kogut, J.: The renormalization group and the $\in$ expansion. Phys. Rep. **12**(2), 75–199 (1974). https://doi.org/10.1016/0370-1573(74)90023-4
18. Beck, C.: Generalised information and entropy measures in physics. Contemp. Phys. **50**, 495–510 (2009). https://doi.org/10.1080/00107510902823517
19. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Handbook of Latent Semantic Analysis. 1st edn. Lawrence Erlbaum Associates, Mahwah (2007)
20. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
21. News dataset from Lenta.ru. https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta
22. Balanced subset of news dataset from Lenta.ru. https://yadi.sk/i/RgBMt7lJLK9gfg
23. 20 Newsgroups dataset. http://qwone.com/jason/20Newsgroups/
24. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications, 1st edn. Chapman and Hall, New York (2008)