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 St.,
St. Petersburg, Russia, 192148
skoltsov@hse.ru,
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 renormalization theory and 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 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. In addition, the results of TM significantly depend on the number of topics and inappropriate values of parameters 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 to 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

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 grid search for determining the best values of model parameters with respect to 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. An alternative approach is the Hierarchical Dirichlet Process model (HDP) [6] positioned by the authors as able to select the number of topics automatically. However, this class of models possesses a set of hidden parameters (α, γ, η), which, according to the authors themselves, can influence the results of TM and the number of topics found. Therefore, we do not consider HDP model in this work.

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 approximation of an optimal number of topics for a given collection. Here and further, the "optimal" number of topics for a dataset means the number of topics that corresponds to human judgment. 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 subsection 2.3. The author of [7] demonstrated that minimum Renyi entropy corresponds to the number of topics determined by human judgment. This approach also requires 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 to apply methods of statistical physics, we conclude that it is possible to apply 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 specially 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.

The rest of the paper proceeds as follows. 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. Subsections 2.3 and 2.4 review Renyi entropy approach which

was proposed in [7], [9]. Subsection 2.5 describes findings of work [8] which are necessary for 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 renormalization 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. The key idea of TM is based on an assumption that any large document collection contains a finite set of topics or semantic clusters, while each word and each text of such a collection belongs to each topic with a certain probability. This gives TM an important ability to co-cluster both words by topics and topics by documents simultaneously. Topics are defined as hidden distributions of both words and texts that are to be restored from the observed co-occurrences of words in texts. Mathematically, topic models are based on the following propositions [10]:

1. Let $\tilde{D}$ be a collection of textual documents with D documents and $\tilde{W}$ be a set (dictionary) of all unique words with W elements. Each document $d \in \tilde{D}$ is a sequence of terms $w_1, ..., w_n$ from dictionary $\tilde{W}$.
2. It is assumed that there is a finite number of topics, T, and each entry of a word w in document d is associated with some topic $t \in \tilde{T}$. A topic is understood as a set of words that often (in the statistical sense) appear together in a large number of documents.
3. A collection of documents is considered a random and independent sample of triples $(w_i; d_i; t_i)$, $i = 1, ..., n$, from the discrete distribution $p(w; d; t)$ on a finite probability space $\tilde{W} \times \tilde{D} \times \tilde{T}$. Words w and documents d are observable variables, and topic t is a latent (hidden) variable.
4. It is assumed that the order of words in documents is unimportant for topic identification (the "bag of words" model). The order of documents in the collection is also not important.

In TM, it is also assumed that the probability $p(w|d)$ of the occurrence of term w in document d can be expressed as a product of probabilities $p(w|t)$ and $p(t|d)$, where $p(w|t)$ is the probability of word w under topic t and $p(t|d)$ is the probability of topic t in document d. According to the formula of total probability and the hypothesis of conditional independence, one obtains the following expression [10], [11]: $p(w|d) = \sum_{t \in \tilde{T}} p(w|t)p(t|d) \equiv \sum_{t \in \tilde{T}} \phi_{wt}\theta_{td}$. Thus, constructing a topic model means finding the set of latent topics $\tilde{T}$, i.e., the set of one-dimensional conditional distributions $p(w|t) \equiv \phi_{wt}$ for each topic t, which constitute matrix Φ (distribution of words by topics), and the set of one-dimensional distributions $p(t|d) \equiv \theta_{td}$ for each document d, which form matrix Θ (distribution of documents by topics), based on the observable variables d and w. In fact, finding

hidden distributions in large text collections from the Internet is equivalent to understanding what people write about without reading millions of texts, that is, to identifying topics which are discussed in a large number of texts.

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 Ising model, only two possible directions are considered, in Potts model, the number of directions can be 3-5 [16]. Nodes with the same spin directions constitute clusters. 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. Block merge procedure is conducted on the whole lattice. Correspondingly, we obtain a new configuration of spins. 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, parameter of power-law, α, can be found by different algorithms, such as 'box counting' or others.

2.3 Entropy Based Approach

Entropic approach for tuning topic models [7], [9] is based on a set of assertions, which link TM to statistical physics and reformulate the problem of model parameter optimization in terms of thermodynamics, namely:

1. A collection of documents is considered an information system: a statistical system where the elements are words and documents amount to millions.

Correspondingly, behavior of such a system can be studied by application of models from statistical physics.
2. The total number of words and documents in the information system under consideration is constant (i.e., the system volume is not changed).
3. A topic is a state (an analogue of spin direction) that each word and document in the collection can take. Here, a word and a document can belong to different topics (spin states) with different probabilities.
4. A solution of topic modeling is a non-equilibrium state of the system.
5. Such information system is open and exchanges energy with the environment via changing the temperature. Here, the temperature of the information system is the number of topics that is a parameter and should be selected by searching for a minimum Kullback-Leibler (KL) divergence.
6. Since KL divergence is equivalent to the difference of free energies [18], to measure the degree to which a given system is non-equilibrium, one can use the following expression: $\Lambda_F = F(T) - F_0$, where F_0 is the free energy of the initial state (chaos) of the topic model and $F(T)$ is the free energy after TM for a fixed number of topics T [19].
7. The minimum of Λ_F depends on topic model parameters such as the number of topics and other hyper-parameters.
8. The optimal number of topics and the set of optimal hyper-parameters of the topic model correspond to the situation when the information maximum is reached. Note that free energy can be expressed through Renyi entropy and information maximum corresponds to Renyi entropy minimum.

It is known that in topic models, the sum of probabilities of all words equals the number of topics $T = \sum_{t \in \tilde{T}} \sum_{w \in \tilde{W}} \phi_{wt}$, where $\phi_{wt} \in [0, 1]$ for all $w \in \tilde{W}$, $t \in \tilde{T}$. In the framework of statistical physics, it is common to investigate the distribution of statistical systems by energy levels, where energy is expressed in terms of probability. We follow [9] and divide the range of probabilities $[0, 1]$ into two intervals, which means we are considering a two-level system, where the first level corresponds to words with high probabilities and the second level – to words with low probabilities close to zero. Consequently, we can introduce the density-of-states function for words with high probabilities under a fixed number of topics and a fixed set of parameters:

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

where N is the number of words with high probabilities. By high probability, we mean the probability satisfying: $p > 1/W$. The choice of such a level is informed by the fact that the values $1/W$ are the initial values of matrix Φ for topic models in LDA. The value $W \cdot T$ determines the total number of micro-states of the topic model (the size of matrix Φ), and normalizes the density-of-states function. During the process of TM, the probabilities of words get redistributed away from the flat distribution $1/W$. A small part of the words obtains probabilities higher than the threshold level, while the larger part of words gets probabilities lower than that. The energy of the upper level containing states with high probabilities

is expressed as follows:

$$E = -\ln(\tilde{P}) = -\ln\left(\frac{1}{T}\sum_{w,t}(\phi_{wt}\cdot\Omega(\phi_{wt}-1/W))\right), \quad (2)$$

where the step function $\Omega(\cdot)$ is defined by $\Omega(\phi_{wt}-1/W)=1$ if $\phi_{wt}\geq 1/W$ and $\Omega(\phi_{wt}-1/W)=0$ if $\phi_{wt}<1/W$. Therefore, in equation (2), we sum only the probabilities that are greater than $1/W$. The energy of the lower level is expressed in the same way, except that summing occurs for probabilities that are smaller than $1/W$. A level is characterized by two parameters: (1) the normalized sum of probabilities of micro-states, that lie in the corresponding interval, $\tilde{P}$, and (2) the normalized number of micro-states (density-of-states function), ρ, whose probabilities lie in this interval. Let us note that the density-of-states function is sometimes called the statistical weight of a complex system's level. For a two-level system, the main contribution to the entropy and the energy of the whole system is made by the states with high probabilities, that is mainly by the upper level. Respectively, the free energy of the whole system is almost entirely determined by the entropy and the energy of the upper level. The free energy of a statistical system can be expressed through Gibbs–Shannon entropy and the internal energy in the following way [20], [21]: $F = E - TS = E - S/q$, where $q = 1/T$. The entropy of such a system can be expressed through the number of micro-states belonging to the same level [22], [23]: $S = \ln(N)$. It follows that the difference of free energies of the topic model is expressed through $\tilde{P}$ and ρ in the following way: $\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0)T = -\ln(\tilde{P}) - T\ln(\rho)$, where E_0 and S_0 are the energy and the entropy of the initial state of the system, with $E_0 = -\ln(T)$ and $S_0 = \ln(WT)$. Hence, the degree to which a given system is non-equilibrium can be defined as the difference between the two free energies and expressed in terms of experimentally determined values ρ and $\tilde{P}$. Values ρ and $\tilde{P}$ are calculated for each topic model solution under variation of parameter T and hyper-parameters, i.e. Λ_F is a function of the number of topics T, hyper-parameters, and size of vocabulary W.

2.4 Application of Renyi Entropy in Topic Modeling

Using partition function:

$$Z_q = e^{-q\Lambda_F} = e^{-qE+S} = \rho(\tilde{P})^q, \quad (3)$$

$q = 1/T$ [23], one can express Renyi entropy in Beck notation through free energy [24] and through experimentally determined values ρ and $\tilde{P}$:

$$S_q^R = \frac{\ln(Z_q)}{q-1} = \frac{\ln(e^{-q\Lambda_F})}{q-1} = \frac{-q\Lambda_F}{q-1} = \frac{q\ln(\tilde{P}) + \ln(\rho)}{q-1}. \quad (4)$$

Summing up the advantages of Renyi entropy application to TM, the following can be said. First, since calculation of Renyi entropy is based on the difference of

free energies (i.e., on KL divergence), it is convenient to use Renyi entropy as a measure of the degree to which a given system (topic model) is non-equilibrium [7]. Second, Renyi entropy, in contrast to Gibbs–Shannon entropy, allows to account for two different processes: a decrease in Gibbs–Shannon entropy and an increase in internal energy, both of which occur with the growth of the number of topics. The difference between these two processes can have an area of balance when two processes counterbalance each other. In this area, Renyi entropy reaches its minimum. Entropy minimum corresponds to information maximum in topic models. Third, the search for the Renyi entropy minimum can be convenient for optimizing parameters of machine learning models.

2.5 Self-similar Behaviour in Topic Models

In this paper we build up on the findings from [8] which is the first work to apply multifractal approach to the analysis of statistical behaviour of topic models under variation of the number of topics. Recall that under a fixed number of topics, a topic solution is a matrix Φ, where the number of elements is $T \cdot W$. Each cell of the matrix contains probability ϕ_{wt} of belonging of a word w to a topic t. Therefore, the multidimensional space of words is covered by a grid of fixed size defined by matrix Φ. The size of each cell of this grid is $\epsilon = 1/(WT)$. Under fixed size of vocabulary W, the size of cells is defined by the number of topics and if $T \to \infty$ then the size of cells tends to zero. As it was mentioned above, the density-of-states function can be defined according to equation (1). This function depends on the size of cells and some degree $D(\epsilon)$ [25], [26]: $\rho(\epsilon) \approx \epsilon^{D(\epsilon)}$. The distribution of fractal dimensions $D(\epsilon)$ can be found using 'box counting' algorithm. Application of this algorithm to the calculation of fractal dimensions in topic models consists of the following steps: 1. Multidimensional space of words and topics is covered by a grid of fixed size (matrix Φ). 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 equation (1). 4. Steps 1, 2, 3 are repeated with cell sizes (i.e. the number of topics) being changed. 5. A graph showing 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 being equal to the value of fractal dimension calculated according to the following relation: $D(\epsilon) = \frac{\ln(\rho(\epsilon))}{\ln(\epsilon)}$.

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 the such regions change in the density-of-states function happens, i.e. 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 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 subsection 3.2). Recall that the output of topic model contains matrix Φ 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 calculation of matrix Φ depends on a particular topic model, mathematical formulation of renormalization procedure is model-dependent. In addition, 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 [27]: $KL(t_1, t_2) = \sum_{w \in \tilde{W}} \phi_{wt_1} \ln(\frac{\phi_{wt_1}}{\phi_{wt_2}}) =$
  $= \sum_{w \in \tilde{W}} \phi_{wt_1} \ln(\phi_{wt_1}) - \sum_{w \in \tilde{W}} \phi_{wt_1} \ln(\phi_{wt_2})$, 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 equation (4), 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 [28], correspondingly. Matrix Φ is estimated by means of Gibbs sampling algorithm. Here, values α and β are set by 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

$$\phi_{wt} = \frac{c_{wt} + \beta}{(\sum_{w \in \tilde{W}} c_{wt}) + \beta W}. \tag{5}$$

For our task of renormalization, we use the values of counters c_{wt} and equation (5). Notice that counters c_{wt} form matrix $C = \{c_{wt}\}_{w \in \tilde{W}, t \in \tilde{T}}$, 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 subsection 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 equation (5). 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 equation (5)):

$$\phi_{w\tilde{t}} = \frac{c_{wt_1} + c_{wt_2} + \beta}{(\sum_{w \in \tilde{W}} c_{wt_1} + c_{wt_2}) + \beta W}. \tag{6}$$

   One can easily see that new distribution $\phi_{\cdot\tilde{t}}$ satisfies $\sum_{w \in \tilde{W}} \phi_{w\tilde{t}} = 1$. 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 equation (4). Then we plot Renyi entropy as a function of the number of topics and search for its minimum in order 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 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 (from the Lenta.ru news agency): a publicly available set of 699,746 news articles in Russian language dated between 1999 and 2018 (available at [29]). Each news item was manually assigned to one of ten topic classes by the dataset provider. We used a class-balanced subset of this dataset that consisted of 8,624 news texts (containing 23,297 unique

words). It is available at [30]. Some of these topics are strongly correlated with each other. Therefore, the documents in this dataset can be represented by 7–10 topics.
– English dataset (the well-known "20 Newsgroups" dataset http://qwone.com/jason/20Newsgroups/): 15,404 English news articles containing 50,948 unique words. Each of the news items belonged to one or more of 20 topic groups. Since some of these topics can be unified, 14–20 topics can represent the documents of this dataset [31]. This dataset is widely used to test machine learning models.

These datasets were used for topic modeling in the range $[2, 100]$ topics in the increments of one topic. Hyper-parameters of LDA model were fixed at the values: $\alpha = 0.1$, $\beta = 0.1$. Research on the optimal values of hyper-parameters for these datasets was presented in work [9], therefore, we do not vary hyperparameters in our work. For both datasets, 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 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 figure 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 end points 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 behavior of renormalized Renyi entropy when the principle of selecting topics for merging is based on KL divergence. It shows that

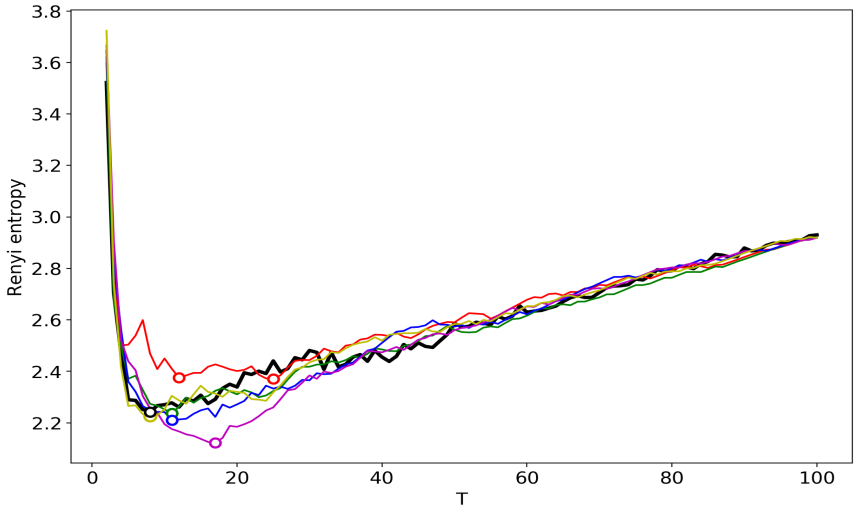

Fig. 1. Renyi entropy distribution over the number of topics T. Russian dataset. Original Renyi entropy – black. Renormalized Renyi entropy with random merging of topics: run 1 – red; run 2 – green; run 3 – blue; run 4 – magenta; run 5 – yellow.

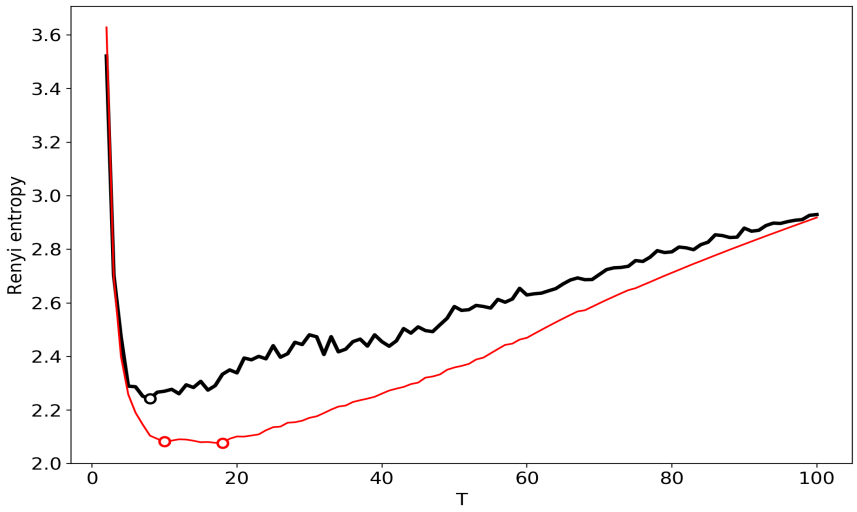

Fig. 2. Renyi entropy distribution over the number of topics T. Russian dataset. Original Renyi entropy – black; renormalized Renyi entropy (topics with the lowest Renyi entropy merged) – red.

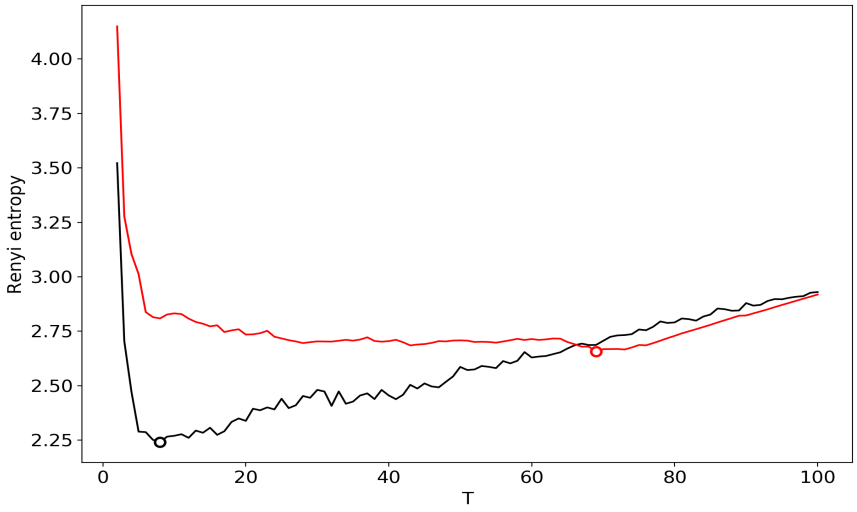

Fig. 3. Renyi entropy distribution over the number of topics T. Russian dataset. Original Renyi entropy – black; renormalized Renyi entropy (similar topics with the lowest KL divergence merged) – red.

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 renormalized Renyi entropy 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, 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 minimum at $T = 43$ that does not correspond either to the human mark-up or to the minimum of the original Renyi entropy.

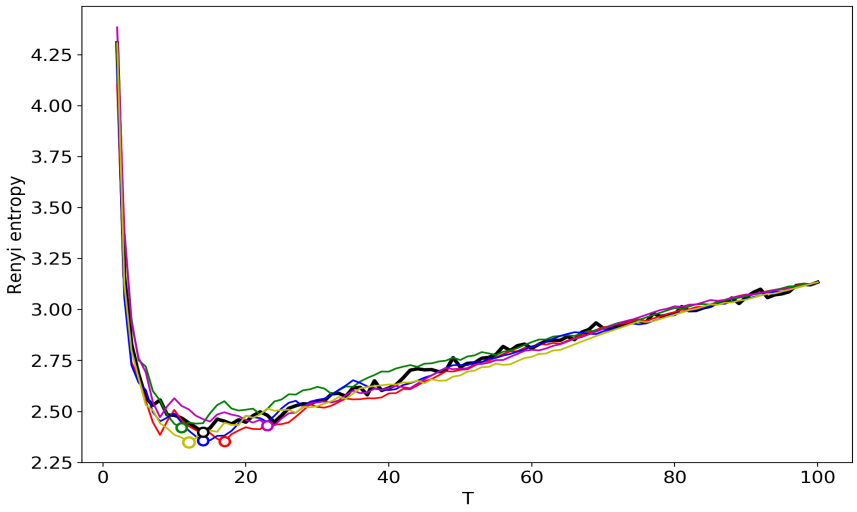

Fig. 4. Renyi entropy distribution over the number of topics T. English dataset. Original Renyi entropy – black. Renormalized Renyi entropy with random merging of topics: run 1 – red; run 2 – green; run 3 – blue; run 4 – magenta; run 5 – yellow.

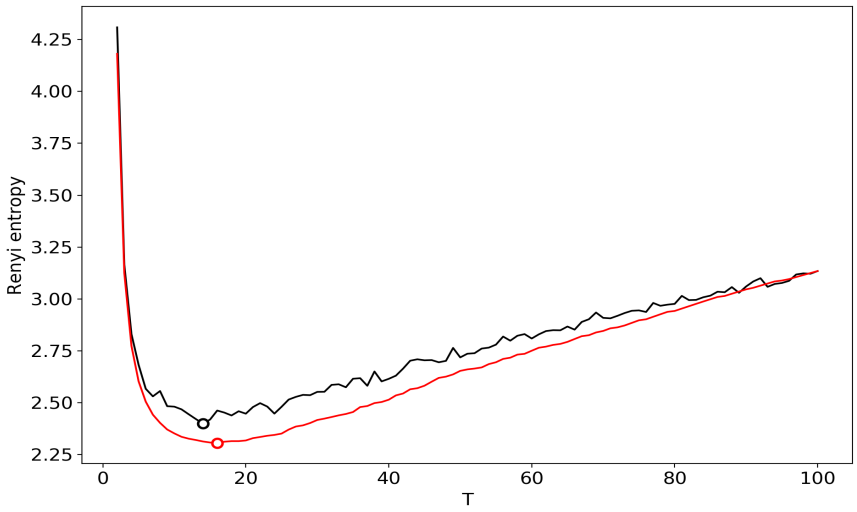

Fig. 5. Renyi entropy distribution over the number of topics T. English dataset. Original Renyi entropy – black; renormalized Renyi entropy (topics with the lowest Renyi entropy merged) – red.

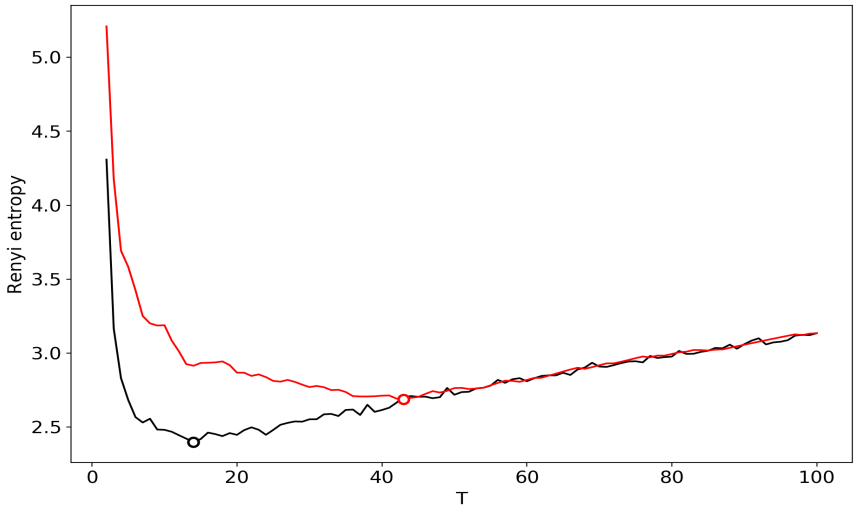

Fig. 6. Renyi entropy distribution over the number of topics T. English dataset. Original Renyi entropy – black; renormalized Renyi entropy (similar topics with the lowest KL divergence merged) – red.

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: 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 renormalized Renyi entropy curve to the original Renyi entropy. In addition, 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 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.

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

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 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 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 Expectation-Maximization algorithm. This appears to be a promising direction for future research deserving a separate paper.

Acknowledgements. 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. Journal of the American Statistical Association 101 (476) (2006).
7. Koltsov S.: Application of Rényi and Tsallis entropies to topic modeling optimization. Physica A: Statistical Mechanics and its Applications 512, 1192–1204 (2018). doi: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. Journal of Physics: Conference Series 1163, 012025 (2019). doi:10.1088/1742-6596/1163/1/012025
9. Koltsov S., Ignatenko V., Koltsova O.: Estimating Topic Modeling Performance with SharmaMittal Entropy. Entropy 21 (7), 1–29 (2019). doi: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. Machine Learning 101, 303–323 (2015). doi: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). doi: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. Materials and Structures 28(8), 435–443 (1995). doi:10.1007/BF02473162
16. Essam, J. W.: Potts models, percolation, and duality. Journal of Mathematical Physics 20(8), 1769–1773 (1979). doi:10.1063/1.524264
17. Wilson, K.G., Kogut, J.: The renormalization group and the expansion. Physics Reports 12 (2), 75–199 (1974). doi:10.1016/0370-1573(74)90023-4
18. Akturk, E., Bagci, G.B., Sever, R.: Is Sharma-Mittal entropy really a step beyond Tsallis and Renyi entropies?. https://arxiv.org/abs/cond-mat/0703277

19. Koltcov, S.N.: A thermodynamic approach to selecting a number of clusters based on topic modeling. Tech. Phys. Lett. 43, 90–95 (2017). doi:10.1134/S1063785017060207
20. Tsallis, C.: Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer, New York (2009).
21. Klimontovich, Yu.L.: Problems in the statistical theory of open systems: Criteria for the relative degree of order in self-organization processes. Sov. Phys. Usp. 32, 416–433 (1989).
22. Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S.E., Berry, M.J., Bialek, W.: Thermodynamics and signatures of criticality in a network of neurons. PNAS 112 (37), 11508–11513 (2015).
23. Mora, T., Walczak, A.M.: Renyi entropy, abundance distribution and the equivalence of ensembles. arXiv:abs/1603.05458
24. Beck, C.: Generalised information and entropy measures in physics. Contemp. Phys. 50, 495-510 (2009). doi:10.1080/00107510902823517
25. Feder, J: Fractals. 1st edn. Plenum Press, New-York (1988).
26. Sornette, D.: Critical Phenomena in Natural Sciences. 1st edn. Springer, Heidelberg (2006).
27. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. In: Handbook of Latent Semantic Analysis. 1st edn. Lawrence Erlbaum Associates, Mahwah (2007).
28. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
29. News dataset from Lenta.ru. https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta
30. Balanced subset of news dataset from Lenta.ru. https://yadi.sk/i/RgBMt7lJLK9gfg
31. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications. 1st edn. Chapman and Hall, New York (2008).