National Research University Higher School of Economics

as a manuscript

Koltcov Sergei Nikolaevich

ENTROPIC TOPIC MODELS AND METHODS OF THEIR AGGREGATION

DISSERTATION SUMMARY
for the purpose of obtaining academic degree
Doctor of Computer Science

Saint-Petersburg - 2022

The dissertation work was performed at the federal state autonomous educational institution of higher education “National Research University “Higher School of Economics”

Scientific adviser: Mirkin Boris Grigorievich, Doctor of Sciences in System Analysis, Management and Information Processing, leading research fellow at International Centre of Decision Choice and Analysis, professor of the Faculty of Computer Science, School of Data Analysis and Artificial Intelligence, NRU HSE.

Introduction

Analysis of large textual data has become one of the most needed scientific directions in the modern world due to the development of electronic means of storing and transmitting such data. Such data, according to its volume, become comparable to physical mesoscopic systems. Therefore, machine learning models on the basis of mathematical formalism borrowed from statistical physics may be used to analyze such data. An example of such a model is topic modeling based on a sampling procedure, where the formalism of the Potts model is used to calculate the distribution of words by topics [15]. The task of topic modeling is to extract the distributions of observable variables (i.e., texts or images and their elements) on hidden variables called topics.

Currently, many topic models are developed [1] with different methods of determining hidden distributions and different measures for quality analysis. However, despite the broad usage of these models in different areas, a set of unsolved problems remains that, in turn, limits the application of TM.

One of the main problems is the problem of determining the number of components in a mixture of distributions since the parameter determining the dimension of the mixture in a model has to be set explicitly. Let us note that in the framework of topic modeling, an approach of automatic selection of the number of topics was developed according to the authors of the approach. However, such models possess a lot of hidden parameters, which significantly impact the modeling results. Moreover, such a model cannot correctly determine the number of topics in a dataset.

The second problem is the instability of topic modeling. This means that topic modeling results are not identical for different model runs on the same dataset and with the same parameter settings. On the one hand, this problem is related to the ambiguity of matrix decompositions (for topic models based on the E-M algorithm). On the other hand, this problem is associated with the presence of many local minima and maxima of the integrand (for topic models based on the Gibbs sampling procedure).

The third unsolved problem arising from the second one is related to the development of regularization procedures, which can be used to improve stability and other purposes [1]. Regularization means adding prior information to topic models in the form of different relations and limitations that reduce the number of possible solutions. Currently, many generative models with regularizers are proposed in the literature. However, there are no clear criteria for choosing a combination of regularizers and selecting regularization coefficients.

The above problems naturally impact the quality of topic modeling. Currently, the major measures of topic modeling quality are Shannon entropy, Kullback-Leibler divergence, log-likelihood, and perplexity. Moreover, it is known that distributions of words, at least in European languages, correspond to a power law, which is typical for complex statistical systems. Also, it is known that the behavior of complex systems can be investigated more efficiently with methods developed in the framework of mathematical formalism borrowed from the theory of complex systems.

Goals and objectives of the study

The goal of the dissertation is the development and investigation of a new class of computational topic models, namely, entropic topic models, aimed at advancing in solving the problems of determining optimal hyperparameters of topic models, including determining the presence of flat or hierarchical structures in datasets and developing stable clustering models of text collections.

Obtained results:

1. Entropic topic model based on one-parametric entropy (Renyi entropy and Tsallis entropy). This model is developed for the following generative algorithms: 1) LDA (Gibbs sampling algorithm), 2) pLSA (E-M algorithm), 3) VLDA (E-M algorithm), and 4) GLDA (Gibbs sampling algorithm).
2. Entropic topic model based on two-parametric entropy (Sharma-Mittal entropy). This model is implemented for the following generative models: 1) pLSA (E-M algorithm), 2) LDA (Gibbs sampling algorithm), and 3) ARTM with sparsing regularizers of matrices Φ and Θ (E-M algorithm).
3. Hierarchical entropic topic model. This model is implemented for the following generative hierarchical algorithms: 1) hLDA, 2) hPAM, 3) hARTM, 4) algorithm of cluster analysis HCA ('complete method').
4. Fractal model of estimation of the performance of generative topic models. This model is implemented for the following algorithms: 1) pLSA (E-M algorithm), 2) ARTM (E-M algorithm), and 3) LDA Gibbs sampling algorithm.
5. An aggregation method of topic models based on the renormalization procedure. The method is implemented for the following algorithms: 1) VLDA (E-M algorithm). 2) LDA (Gibbs sampling algorithm). 3) pLSA (E-M algorithm).
6. The aggregation method is implemented for three variants of merging topics: 1) merging based on minimum Renyi entropy, 2) merging of random topics, and 3) merging based on Kullback-Leibler divergence.
7. Granulated topic model based on Gibbs sampling procedure. This model is implemented for three variants of the function of local distribution of topics: 1) GLDA, 2) ELDA, and 3) TLDA.

The author 's personal contribution includes:

- General mathematical formulation of entropic model based on one-parametric Renyi entropy, published in two articles with one author.
- Organization and participation in large-scale computer experiments on analysis of entropic models' applicability for estimating different topic models' performance.
- Leading participation in the mathematical formulation of entropic model based on two-parametric Sharma-Mittal entropy and testing this model in a series of computer experiments.
- Formulation of the fractal model for estimating generative topic models performance and conducting computer experiments to test this model.
- Mathematical formulation of the aggregation method of topic models with renormalization procedure and conducting computer experiments on testing the effectiveness of the renormalization procedure.
- General mathematical formulation of the granulated sampling method.

On the topic of this dissertation, 8 articles were published in Q1-Q2 journals, according to WoS, and 11 articles indexed in Scopus.

Scientific novelty:

1. For the first time, the application of one-parametric Renyi entropy and two-parametric Sharma-Mittal entropy was proposed for optimization of topic models' performance.
2. For the first time, it was demonstrated that quality measures based on parameterized entropies outperform traditional quality measures such as log-likelihood or perplexity since they allow one to tune hyperparameters values of topic models and the number of distributions in the mixture simultaneously.
3. For the first time, the fractal model for estimating generative topic model performance was proposed. This model demonstrates the self-similar behavior of topic models that allows one to apply the renormalization procedure to them.
4. For the first time, the renormalization procedure of topic models was proposed, and its effectiveness for fast determining the optimal number of distributions in the mixture was demonstrated.
5. A granulated version of the topic model, which outperforms other topic models in terms of stability, was proposed.

Publications in high-impact journals (Q1-Q2 according to WOS and Scopus)

1. Koltcov, S., Ignatenko, V., Terpilovskii, M., Rosso, P. Analysis and tuning of hierarchical topic models based on Renyi entropy approach // PeerJ Computer Science, Vol. 7, 2021. Open access**:** https://peerj.com/articles/cs-608/
2. Koltsov S., Ignatenko V., Boukhers Z., Staab S. Analyzing the Influence of Hyper-parameters and Regularizers of Topic Modeling in Terms of Renyi entropy // *Entropy*. 2020. Vol. 22. No. 4. pp. 1-13.
3. Koltcov S, Ignatenko V. Renormalization Analysis of Topic Models // *Entropy*. 2020. Vol. 22. No. 5. pp. 1-23.
4. Koltsov S., Ignatenko V., Koltsova O. Estimating Topic Modeling Performance with Sharma–Mittal Entropy // *Entropy*. 2019. Vol. 21. No. 7. pp. 1-29.
5. Koltsov S. Application of Rényi and Tsallis entropies to topic modeling optimization // *Physica A: Statistical Mechanics and its Applications*. 2018. Vol. 512. pp. 1192-1204.
6. Koltcov, S.N. A thermodynamic approach to selecting a number of clusters based on topic modeling / Koltcov, S.N. // Technical Physics Letters. 2017. Vol. 43. No.12. pp. 90-95.
7. S. N. Koltsov, S. I. Nikolenko, and E. Yu. Koltsova Gibbs Sampler Optimization for Analysis of a Granulated Medium // Pis'ma v Zhurnal Tekhnicheskoi Fiziki, 2016, Vol. 42, No. 16, pp. 21–25.
8. Sergey Nikolenko, Sergei Koltcov, Olessia Koltsova. Topic modelling for qualitative studies // *Journal of Information Science*. 2017. Vol. 43. No. 1. pp. 88-102.

Standard level publications on the research topic (Scopus)

1. Koltsov S., Ignatenko V., Pashakhin S. How many clusters? An Entropic Approach to Hierarchical Cluster Analysis, in: *Intelligent Computing: SAI 2020: Volume 3* Vol. 1230. Book 3. Cham : Springer, 2020. pp. 560-569.
2. Koltsov S., Ignatenko V. Renormalization approach to the task of determining the number of topics in topic modeling, in: *Intelligent Computing: SAI 2020: Volume 1* Vol. 1228. Part 1. Switzerland : Springer, 2020. pp. 234-247.

3. Ignatenko V., Sergei Koltcov, Staab S., Boukhers Z. Fractal approach for determining the optimal number of topics in the field of topic modeling // *Journal of Physics: Conference Series*. 2019. Vol. 1163. No. 1. pp. 1-6.
4. Koltsov S., Pashakhin S., Dokuka S. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 10th International Conference on Social Informatics, SocInfo 2018; St.Petersburg*. Cham: Springer, 2018. pp. 308-321.
5. Mavrin A., Filchenkov A., Koltsov S. Four Keys to Topic Interpretability in Topic Modeling, in: *Artificial Intelligence and Natural Language, 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings* Issue 930. Switzerland : Springer, 2018. doi pp. 117-129.
6. Koltsov S., Nikolenko S. I., Koltsova O., Filippov V., Bodrunova S. Stable Topic Modeling with Local Density Regularization, in: *Internet Science, Proc. of 3d conf INSCI 2016, Lecture Notes in Computer Science series* Vol. 9934. Switzerland : Springer, 2016. doi pp. 176-188.
7. Koltsov S., Nikolenko S. I., Koltsova O., Bodrunova S. Stable topic modeling for web science: Granulated LDA, in: *WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference*. Elsevier, 2016. pp. 342-343.
8. Koltsov S., Koltsova O., Nikolenko S. I. Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated content, in: *Proceedings of WebSci '14 ACM Web Science Conference, Bloomington, IN, USA — June 23 - 26, 2014*. NY : ACM, 2014. pp. 161-165.
9. Nikolenko S. I., Koltsov S., Koltsova O. Measuring Topic Quality in Latent Dirichlet Allocation, in: *Proceedings of the Philosophy, Mathematics, Linguistics: Aspects of Interaction 2014 Conference*. St. Petersburg : The Euler International Mathematical Institute, 2014. pp. 149-157.
10. Koltsov S., Ignatenko V., Pashakhin S. Fast Tuning of Topic Models: An Application of Rényi Entropy and Renormalization Theory, in: *Proceedings of the 5th International Electronic Conference on Entropy and Its Applications* Vol. 46. Issue 1. MDPI AG, 2020. Ch. 5. pp. 1-8.
11. Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A. Interval Semi-supervised LDA: Classifying Needles in a Haystack, in: *Proceedings of the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013)* Part I: Advances in Artificial Intelligence and Its Applications. Berlin : Springer, 2013. pp. 265-274.

1. Analytical overview of scientific literature

1.1. Approaches to the problem of selecting the number of clusters

The overview considers the most interesting and valuable for this dissertation investigations. The main problem in searching for the optimal number of clusters in cluster analysis and topic modeling is the choice of the function based on which such searching is conducted. Discussion of many clustering quality measures, including functions for selecting the number of clusters, is presented in works [10, 11]. These and other works demonstrate that minimal intracluster distance is frequently used for these purposes in cluster analysis. However, the problem with this and similar measures is that dependence of such measures on the number of clusters is monotone increasing (or decreasing). Correspondingly, the development of transformation procedures for extracting peculiarities from these functions is needed. In work [3], an algorithm for determining the optimal number of topics based on ‘rate distortion theory’ was formulated. Modernization of this approach in the framework of non-extensive statistical physics for image clustering was implemented in the work [4].

Other approaches to solving this problem exist in cluster analysis [5, 6, 7]. In the work by Tibshirani [6], a method called ‘gap statistic’ was proposed. Its key idea is to measure the difference between null reference distribution and distribution obtained from clustering from the above distribution. This difference is calculated for different numbers of clusters. After that, a corresponding curve is plotted. In the framework of this approach, the authors assume that the optimal number of clusters corresponds to the situation when the logarithm of average intracluster distance becomes less than the analogous logarithm calculated for null reference distribution. In fact, this is an analog of measuring the dependence of entropy on the number of clusters with respect to initial entropy. In work [8], a clustering procedure is proposed based on searching maximum entropy (maximum entropy principle). However, the authors also rely on the classical variant of entropy (Shannon entropy). But already in work [9], a clustering method is implemented by applying the Tsallis entropy maximization principle through variation of parameter q.

Among all existing approaches in cluster analysis, the most interesting and informative is an approach based on free energy minimization [12]. Its main idea is as follows: each element of the statistical system is characterized by probabilities of belonging to different clusters. Correspondingly, for each element, one can formulate the notion of internal energy (expressed through the probability of belonging of the element to a cluster) and calculate the free energy of the entire system. The temperature in such a system turns into a free parameter, which is varied to minimize free energy. A disadvantage of this work is model testing only on clusters with Gaussian distributions. Moreover, the authors' calculations demonstrate that the free energy function looks like a monotone function without an explicit minimum

This dissertation is based on ideas of work [12]. However, in contrast to this work, the temperature is considered a number of clusters, and parameterized entropies (that possess a clear minimum) are considered instead of free energy. Theoretical statements of entropic topic models are described in chapter 2.

1.2. Overview of model types in topic modeling

Currently, more than forty different topic models are proposed in the literature on topic modeling, and the number of articles applying topic modeling exceeds several hundred. In general, one can distinguish three main types of models: 1. Flat topic models with different types of regularization [13, 14, 15, 16]. 2. Hierarchical topic models [17, 18, 19, 20]. 3. Topic models with elements of neural networks, where either different types of word embeddings or layers of neural networks are used [21, 22]. The most complete overviews of the various models and quality measures are presented in [1, 23]. In general, two main algorithms of determining distributions of words by topics and topics by documents dominate in the literature: 1. Expectation-Maximization algorithm. In the framework of this algorithm, the matrix of words in documents (F) is represented as the product of two matrices $F=\Phi\Theta$, where Φ is the matrix of words by topics distributions, and Θ is the matrix of distribution of topics by documents. 2. Algorithm of determining the probability of a word belonging to a topic in the form of a multidimensional integral. In this algorithm, the computation of probabilities is implemented with the Gibbs sampling procedure. Despite the different mathematical formalisms of the above algorithms, both lead to similar results [24]. Thus, the problems considered below are valid for different algorithms.

The problem of searching for the optimal number of topics/clusters in topic modeling is relevant and even more complicated to solve. This is due to the following reasons. First, this search is related to the topic's linguistic concept, which in turn causes considerable difficulties since it is difficult to formulate a linguistic criterion for separating two topics on a set of documents. Moreover, topic models often generate topics that are difficult to interpret and difficult to treat as topics. Second, in topic modeling as well as in cluster analysis, it is hard to formulate an appropriate functional dependence, which on the one hand, would characterize the topic model, and on the other hand, would be a function of the number of topics and hyperparameters. Nevertheless, there are several works, authors of which tried to solve the problem of selecting the number of topics in topic modeling. Based on ideas of cluster analysis, the authors of the work [25] considered the topic a semantic cluster (set of words), in the framework of which one can calculate intracluster distance. The authors used cosine measure as the function for minimization. Thus, according to the authors, the number of topics corresponding to the minimum average cosine measure calculated for all topics is optimum. Another approach to searching the optimal number of topics was proposed by Arun et al. in work [26] in the form of searching minimum Kullback-Leibler divergence under variation of the number of topics. The authors propose to implement SVD decomposition of matrices Φ and Θ and then to calculate Kullback-Leibler divergence based on two vectors containing singular values. In this case, the optimal number of topics corresponds to the situation when both matrices are described with the same number of singular values. The disadvantages of these two approaches are as follows. First, it is unclear how the minima of chosen functions are related to the entropic principle, widely used in information theory. Second, the addition of another calculation step, namely SVD decomposition and calculation of Kullback-Leibler divergence, significantly limits the application of Arun's approach to big data processing. Arun and his colleagues searched minimum Kullback-Leibler divergence on text collections that do not exceed 2500 texts. Third, the influence of the initial distributions on the results of topic modeling is not considered in both approaches. However, it is known that there is such an influence [27]. Fourth, the effect of semantic instability, which takes place in topic modeling [28], is not considered in the above approaches.

A topic model based on an additive regularization algorithm (ARTM), proposed in work [16], is worth discussing separately. This model is based on searching the maximum of a linear combination of log-likelihood and a set of regularizers. The values of coefficients determine the level of influence of regularizers on a topic model. Despite the broad usage of this model in Russian-language literature, it has one significant disadvantage: the principle of choosing values of regularization coefficients is not formulated in the theory of additive regularization. These values have to be set explicitly before topic modeling. In this work, a solution to this problem is proposed.

One of the leading quality measures in topic modeling is maximum log-likelihood [1] and perplexity, which is related to log-likelihood. In general, log-likelihood allows one to tune hyperparameters of flat topic models. However, it does not allow one to determine the optimal number of topic clusters. Moreover, log-likelihood is not suitable for tuning hierarchical topic models [2], where an additional unsolved problem of selecting the number of topics on each level of hierarchy exists as well as the traditional problem of selecting values of hyperparameters.

Moreover, coherence measure is widely used in topic modeling. This measure allows one to estimate the coherence of topics in a topic solution [67]. The essence of this measure is to calculate how often words with high probabilities co-occur in highly probabilistic documents.

Large coherence corresponds to the best solution. This measure does not allow one to determine the optimal number of components in the mixture distribution due to its monotone behavior.

Thus, the following unsolved problems for topic models of different types arise. 1. How to determine the optimal number of clusters in the topic solution is unclear. 2. The existing quality measures are not universal, i.e., unsuitable for all models. 3. no quality measure would allow one to tune several model parameters simultaneously (including hyperparameters, the number of clusters, and semantic coherence).

This work proposes a solution to the above problems using the application of parameterized entropies in topic models. Theoretical and experimental estimation of the applicability of parameterized entropies in topic models is presented in chapter 2.

1.3. Application of entropic principles in the field of topic modeling

The following part of the overview is devoted to applying the simulated annealing procedure for determining the hidden distributions in topic modeling. In works [29, 30], a classical version of the annealing algorithm based on the Markov process is used. In work by Tsallis [74], a modified annealing algorithm is proposed. However, in the field of machine learning, this algorithm was not applied.

In work by Zhu [31], the maximum entropy discrimination latent Dirichlet allocation (MedLDA) model was proposed. The essence of this model is the introduction of Kullback-Leibler divergence, which is an entropic regularizer, into log-likelihood. Let us also mention work [32], where a topic model was proposed, where minimum Shannon entropy calculated by words is used for determining the optimal values of the regularization coefficient. A significant disadvantage of this work is testing the proposed model on datasets marked-up only with two topics.

1.4. Stability of topic models

Despite a large number of works devoted to topic models, the number of works related to the estimation of their stability is very limited. The problem of topic model stability is related to topic model construction features.

The solution of the task of topic modeling is equivalent to stochastic matrix decomposition, where a large matrix F, containing documents d and words w, is approximated with a product of matrices Θ and Φ of lower dimensions. However, stochastic matrix decomposition is not uniquely determined, but with accuracy up to non-degenerate transformation [16]. If F=ΦΘ is a solution, then F=(ΦS)(S^(-1) Θ) is also a solution for all non-degenerate S, for which matrices Φ^'=ΦS and Θ^'=S^(-1) Θ are stochastic. In terms of TM algorithm, ambiguity in the reconstruction of the multidimensional density of a mixture of distributions is because the algorithm, starting from various initial points, will converge to different points from a set of solutions. It means that different runs of an algorithm on the same source data will lead to different matrices Θ and Φ. The tasks, solutions of which are not unique or are unstable, are called ill-posed. Regularization by Tikhonov [38] gives a general approach to solving such tasks. The essence of regularization is adding prior information that reduces the set of solutions. Regularization is implemented either by introducing limitations on matrices Θ and Φ [16] or modifying the sampling procedure [72, 73].

One can distinguish several works in research on the stability of topic solutions. In work by Griffiths and Steyvers [15], symmetric Kullback-Leibler divergence is proposed for estimation of similarity between two topics from different topic solutions. However, this work does not contain a detailed investigation of the applicability of this measure in practical experiments. A modified version of symmetric KLB divergence was proposed in work by Koltcov et al. [28], where a complete suitable algorithm of topic model stability estimation on three runs is presented. In work by Belford, the ‘Average Descriptor Set Difference (ADSD)’ measure was proposed [39]. This measure characterizes an average value of the number of similar words in two topic solutions. Moreover, this work also considers the ‘Average Term Stability’ measure based on average Jaccard distance. The authors of this work propose a new way to extract stable topics using the ‘K-Fold ensemble approach’. This approach is tested for LDA model with Gibbs sampling and NMF (Non-negative Matrix Factorization approach), which is closely related to TM. In the work by Greene et al. [40], ‘Average Jaccard (AJ) measure’ was also used to determine the optimal number of topics in marked-up datasets in English. In work [41], De Waal demonstrated that perplexity is not suitable for estimation of topic models stability since, first, it depends on the dataset size, that, in turn, complicates comparison between different datasets. Second, it has a monotone decreasing behavior.

The above works aim only at developing and testing a stability measure. However, several works propose a modification of a topic model itself to increase its stability. In the work of Koltcov et al. [42], it is demonstrated that the choice of regularization coefficients in LDA model with Gibbs sampling and ARTM model significantly impact topic model stability. Moreover, a granulated version of Gibbs sampling procedure (GLDA), which provides a very high level of topic model stability, is proposed in this work. A detailed description of GLDA is presented in chapter 5 of this work. It was demonstrated in this work that adding regularizers impacts topic model stability.

The most detailed overview of articles related to the problem of stability/instability of topic models is given in the work of Agarwal [44]. In general, the problem of instability of topic models is not solved completely.

Based on the literature overview, one can conclude the following. In English-language literature, the following models are most widely used: 1. pLSA (E-M algorithm). 2. LDA with Gibbs sampling. 3. Variation LDA (E-M algorithm). These models are most often used as a baseline for comparison with other topic models. In Russian-language literature, ARTM model with Gibbs sampling) is widely used. This model contains an alternative approach to the variational principle of topic model inference and inference based on the physical Potts model (an extended version of the Ising model).

Among quality measures of topic models, the most widely used are the following: 1. Maximum log-likelihood (tuning of topic models). 2. Kullback-Leibler divergence (determining the stability of topic models). 3. Coherence (determining coherence of topics in topic models).

In the field of topic modeling, the following problems are found: 1. The problem of determining the optimal number of topics. In the existing models, this number has to be set explicitly; however, the selection criteria are unclear. 2. The problem of estimation of hyperparameter values, including regularization coefficients, which significantly impact the results of topic modeling. The choice of such parameters can be partially solved using log-likelihood optimization; however, this approach is suitable only for several flat topic models. 3. The problem of stable topic model development. This problem is aggravated by the fact that stability significantly depends on the number of topics and hyperparameters values, the choice of which is not clear. 4. The problem of simultaneous estimation of a topic model in terms of hyperparameters tuning and in terms of semantic coherence of topics. The above problems remain unsolved since the development of topic modeling mainly proceeded in the direction of developing a large number of new models. Investigations related to the analysis of model tuning or solving the problem of stability are incomplete and very limited in their number. Thus, this dissertation aims to partially solve the above problems.

In the framework of this work, the following topic models are considered: 1. LDA (with Gibbs sampling), 2. pLSA (E-M algorithm), 3. VLDA (E-M algorithm), 4. GLDA (with Gibbs sampling). 5. ARTM with sparsing regularizers of matrices Φ and Θ (E-M algorithm). 6. hLDA. 7. hPAM. 8. hARTM. This choice is according to the following facts. First, these models are the most frequently used in the literature (especially as baselines when developing new models). Second, they are based on two different principles (E-M algorithm and Gibbs sampling procedure). Third, these models are designed to work with datasets having different topical structures. In this dissertation, datasets in different languages with mark-up and without mark-up and with the topical structure of different hierarchy depths and without it were used. This allows us to estimate the effectiveness of the developed models for determining different topical structures.

2. Entropic topic model based on parameterized Renyi entropy and Sharma-Mittal entropy

This chapter considers the theoretical formulation of an entropic topic model for one- and two-parametric entropies. Moreover, a series of computer experiments on marked-up datasets and collections without mark-up is presented. The experiments demonstrate the usefulness of parameterized entropies for topic model tuning and determining the presence of flat or hierarchical topical structure in the data.

The proposed entropic topic model is based on the ideas of work by Rose [12], where it was demonstrated that the clustering procedure might be considered in terms of the probability of belonging to a cluster. Such probability is expressed through the free energy of the entire statistical system (i.e., through the partition function of the system). In such a clustering model, the temperature is a parameter of the cluster model that can be tuned and found by means of an annealing procedure. In contrast to the model of Rose, the entropic topic model considers temperature as the number of clusters, and parameterized entropies are used as objective functions. This difference makes it possible to formulate an entropic model of hyperparameter tuning, including the number of topics, based on the search for the minimum of parametrized entropy. The entropic topic model generally is based on the following assertions [45, 46]. 1) A document collection is a mesoscopic information system containing elements (words and documents). Therefore, the behavior of such a system can be studied using methods from statistical physics. Moreover, such information systems are not close since information is exchanged with the surroundings: for example, a user can change the number of topics/clusters. Correspondingly, it is possible that such a system does not reach an equilibrium state in the sense of the maximum of Shannon entropy but may reach an intermediate equilibrium state, which is determined by a local minimum of the parametrized Renyi entropy or Tsallis entropy. 2) Topic is considered as a state (analogous to spin direction) that each word and document can take in the collection. Moreover, each word and document belong to all topics with different probabilities (matrices of these probabilities are usually denoted by Φ and Θ, correspondingly). The set of words and documents with high probabilities on a topic form what can be called a topic cluster. 3) Information system exchanges only energy with the surroundings by changing temperature. In this approach, the number of topics is considered as the temperature of the information system, which is set externally and is a parameter to be determined by means of searching for a minimum of non-symmetric Kullback-Leibler divergence (free energy is the physical analog of this measure). Since this measure is equivalent to the difference of free energies [47], where one part of free energy corresponds to the initial (equilibrium) state, and the second one characterizes the non-equilibrium state of the system [47], the following expression can be used as a measure of the degree to which a given information system is non-equilibrium: Λ_F=F(T) - F_0, where F_0 is free energy of the initial state (chaos) of the topic model, F(T) is the free energy under given number of topics T, obtained after topic modeling. 4) Minimum of Λ_F depends on different parameters of topic model. 5) The optimal number of topics and set of optimal parameters of topic model corresponds to the situation when information maximum $S = -I$ [48] is reached, i.e., minimum of Λ_F and minimum Renyi entropy, which can be expressed in terms of the difference of free energies.

In topic models, the sum of all word probabilities equals the number of topics: $T = \sum_{t=1}^{T} \sum_{n=1}^{W} p_{tn}$. In the framework of statistical physics, it is common to investigate the distribution of statistical system by energy levels, where the energy of a level is expressed through probability. According to this approach, in this work, the range of probabilities is divided into a fixed number of intervals, energies of these levels and the number of words on each level are determined. Let us note that the number of words in each interval depends on the number of topics and parameter values of the topic model. The division into intervals is conditional and convenient from a computational point of view. When the length of such an interval tends to zero, the distribution of words by intervals tends to density function ρ. However, to simplify the presentation, we will consider a two-level system in which words with a high probability will arise on one level, and words with a low probability (i.e., with a probability close to zero) will appear on the other level.

2.1. Entropic topic model based on Renyi entropy

Let us introduce a density-of-states function for the level of words with high probabilities under a fixed number of topics and a fixed set of parameters [46]:

$$\breve{\rho} = \frac{\sum_{t=1}^{T} \sum_{n=1}^{W} N_{tn}}{WT} \qquad (1),$$

N_{tn} is the number of words with high probabilities, T is the number of topics, n refers to summation by a list of unique words, t refers to summation by all topics. Probability is considered high if it satisfies $p_{tn} > 1/W$, where W is the number of unique words in the dataset. The choice of this threshold is due to the fact that the value $1/W$ is the initial value for the initialization of matrix Φ. Value WT determines the total number of all microstates in a topic model (under microstate we mean probability of one word in one topic), i.e., the size of matrix Φ is the normalization of the density-of-states function. During the process of topic modeling probabilities of words are redistributed with respect to the given threshold. A small part of words falls into the level with high probabilities $p_{tn} > 1/W$, and the larger part of words falls into another level, where $p_{tn} > 1/W$. The level of words with high probabilities in the topic model can be characterized by energy value, which can be expressed through the sum of probabilities of words residing on this level and normalized by the total number of topics:

$$E = -T \cdot \ln \breve{P} \qquad (2),$$

where $\breve{P} = \sum_{t=1}^{T} \sum_{n=1}^{W} p_{tn} / T$, the summation is for all words with high probabilities residing on this level, T is the number of topics. Thus, the level is determined by two experimentally measured values: 1. The sum of words probabilities on the given level $\breve{P}$. 2. The number of words residing on this level (density-of-states function $\breve{\rho}$).

For a two-level system, the main contribution to the entropy and energy of the entire system is given by words with high probabilities; therefore, the free energy of the entire system is approximately determined through entropy and energy of one level. The free energy of such a system is expressed through Gibbs entropy (Shannon entropy) and energy in the following way

[47]: $F = E - T \cdot \mathrm{S} = \mathrm{E} - \mathrm{S}/$, where $q = 1/T$. The entropy of the information system (Shannon entropy) is expressed through the number of words on one level as follows: $S = \ln(\breve{\rho}(T))$ [45]. The difference between free energy of the system is expressed through $\breve{P}$ and $\breve{\rho}$ as follows:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\breve{P}\right) - T \cdot \ln(\breve{\rho}) \qquad (3)\,,$$

where E_0, S_0 are energy and entropy of the system under initial distribution, which corresponds to maximum entropy, i.e., $S_0 = \ln(T)$ and $E_0 = -\ln(W \cdot T)$. Thus, the level of non-equilibrium of topic model is determined as the difference of free energies and is expressed through experimentally determined values $\breve{\rho}$ и $\breve{P}$. Normalization of these values, in its essence, is the entropy of the initial state, that is, chaos. Values $\breve{\rho}$ и $\breve{P}$ are calculated for each topic model under variation of free parameter T and other model parameters; thus, value Λ_F is a function of the number of topics T, size of the vocabulary W, i.e., of the dataset, and it depends on parameters values of the generative topic model.

2.2. Relation between free energy and Renyi entropy in topic models

Based on the partition function $\mathbf{Z_q} = \sum \breve{\boldsymbol{\rho}} \cdot \breve{\mathbf{P}} = \sum \breve{\boldsymbol{\rho}} \cdot \mathbf{e^{-q \cdot E}} = \sum \mathbf{e^{-q \cdot \Lambda_F}}$, q=1/T [49], one can express free energy of topic model through Renyi entropy and experimentally determined values $\breve{\boldsymbol{P}}$ и $\breve{\boldsymbol{\rho}}$ as follows:

$$S_q^R = \frac{\ln(Z_q)}{q-1} = \frac{\ln(e^{-q \cdot F})}{q-1} = \frac{-\mathrm{q} \cdot \Lambda_F}{q-1} = \frac{\Lambda_F}{T-1} \qquad (4).$$

Let us note that the relation between free energy and Renyi entropy can also be found with escort distribution [50, 51] since specifying the above partition function is equivalent to escort transformation.

Thus, Renyi entropy in topic models is expressed through free energy, parameter q, where $q = 1/T$, and experimentally determined values $\breve{P}$ и $\breve{\rho}$. In the framework of this approach, first, Renyi entropy characterizes the measure of the degree to which a given topic model is non-equilibrium since its calculation is based on the difference of free energies. Second, optimization of machine learning models can be implemented based on searching minimum Renyi entropy. Third, Renyi entropy in its formulation, in contrast to Shannon entropy, includes two processes in different directions; namely, an increasing the number of topics leads to decreasing Shannon entropy and increasing total energy, and to increasing the total sum of probabilities in the model. Thus, the difference between these two processes has a region of balance, where these two processes counterbalance each other. In this region, Renyi entropy is minimal. Moreover, entropy minimum corresponds to information maximum in a topic model. Therefore, topic model parameters can be tuned based on searching a minimum of one-parametric Renyi entropy.

2.3. Entropic model based on Sharma-Mittal entropy

Topic model based on Renyi entropy does not include a semantic component, which plays an important role in the practical application of clustering models on textual data. However, the entropic topic model may be extended using the application of two-parametric Sharma-Mittal entropy [52, 53]. It is expressed as follows: $S_{S,M} = \frac{1}{1-r}\left[\left(\sum_i p_i^q\right)^{\frac{1-r}{1-q}} - 1\right]$, where r and q are parameters that determine the type of entropy parameterization. Sharma-Mittal entropy includes Renyi and Tsallis entropies as special cases for certain r, q. For example, for r → 1 entropy $\mathrm{S}_{\mathrm{S,M}}$ equals Renyi entropy, and for r → q $\mathrm{S}_{\mathrm{S,M}}$ equals Tsallis entropy. Let us note that the limit of

Sharma-Mittal entropy when r → 0 equals the exponential function of Renyi entropy minus 1, that can be considered as parameterized perplexity. Namely, in this case, $\lim_{r\to 0} S_{S,M} = e^{S_q^R} - 1$. Let us demonstrate that $e^{S_q^R} - 1 > S_q^R$ if $S_q^R \neq 0$. Let us consider $f(x) = e^x - 1 - x$ for $x \neq 0$. We obtain that $f'(x) = e^x - 1$. Correspondingly, f increases for $x > 0$ and decreases for $x < 0$. Thus, $\min f(x) = f(0) = 0$. For example, if $S_q^R = 6$, $e^{S_q^R} - 1 \cong 402$; for $S_q^R = 1$, $e^{S_q^R} - 1 \cong 1.7$; for $S_q^R = 0.1$, $e^{S_q^R} - 1 \cong 1.005$, i.e., fluctuation of parameter r leads to very large values of entropy.

Since parameter q=1/T is related to the number of topics in a topic model, one has to define parameter r to apply Sharma-Mittal entropy for the analysis of topic models. The values of this parameter may vary in the region [0;1]. Moreover, if r=1, then Sharma-Mittal entropy turns into Renyi entropy, and, correspondingly, the quality of a topic model is determined only by Renyi entropy and parameter q. Based on this, one can conclude that parameter q for Sharma-Mittal entropy is the inverse number of topics. If $r = 0$, Sharma-Mittal entropy is as follows: $S_{S,M} = e^{S_q^R} - 1$, i.e., becomes very large. Since entropy maximum corresponds to information minimum, one can conclude that minimal values of parameter r, which lead to maximal values of $S_{S,M}$, correspond to minimal values of information.

In the literature, the concept of the Jaccard distance is used, which is defined as follows [54]: $J(X,Y) = 1 - \frac{X\cap Y}{X\cup Y}$. Jaccard distance measures the similarity of two sets (in our case, between two sets of words) and is determined as the size of the intersection of the sets divided by the size of the union of the sets. If two sets are identical, then this distance equals zero. Jaccard distance plays an important role especially in the field of computer science for the investigation of regular languages [55], and is related to entropy distance as follows:

$$D_H(X,Y) = 1 - \frac{I(X,Y)}{H(X,Y)} = J(X,Y) = 1 - J,$$

where $I(X,Y)$ is the mutual information of X and Y, $H(X,Y)$ is the joint entropy of X and Y. In information theory, mutual information corresponds to the intersection of sets X and Y, and joint entropy corresponds to the union of X and Y, and, correspondingly, entropy distance corresponds to Jaccard distance. If $J(X,Y) = 0$, then $D_H(X,Y) = 0$. Thus, we can define parameter r as follows. Parameter r in $S_{S,M}$ entropy will be responsible for semantic component of topic model, i.e., it will be measured by means of Jaccard distance. This parameter characterizes the value of variation of semantic composition under variation of the number of topics (and variation of hyperparameters values of the topic model). This is related to the fact that variation of model hyperparameters and the number of topics impact the composition of high-probability words in the topic model.

Thus, tuning the entropic topic model is implemented by selecting the number of topics (parameter q=1/T) and model hyperparameters under the condition of reaching a minimum of two-parametric Sharma-Mittal entropy, i.e., among the set of parameter values, one has to choose those values that correspond to the information maximum of the topic model for the chosen dataset.

2.4. Entropic topic model based on Sharma-Mittal entropy

Based on equation (4) and partition function $Z_q = \sum \tilde{\rho} \cdot e^{-q\cdot E}$, Sharma-Mittal entropy of a topic model can be expressed in terms of experimentally determined values $\tilde{P}$ и $\tilde{\rho}$ as follows [56]:

$$S_{S,M} = \frac{1}{1-r}\left[\left(Z_q\right)^{\frac{1-r}{q-1}} - 1\right] = \frac{1}{1-r}\left[\left(\check{\rho} \cdot \check{P}\ \right)^{\frac{1-r}{q-1}} - 1\right] =$$

$$= \frac{1}{1-r}\left[\left(\left(\frac{P(T)}{T}\right)^{q} \cdot \left(\frac{N_{tn}}{WT}\right)\right)^{\frac{1-r}{q-1}} - 1\right],\ \ (5),$$

where W is the number of words in vocabulary, T is the number of topics, $P(T)$ is the sum of probabilities on the second level, N_{tn} is the number of words with high probabilities, i.e., the number of words on the second level, n refers to summation by the list of unique words, t refers to summation by all topics. Correspondingly, equation (5) allows one to calculate two-parametric entropy of topic model based on experimentally observable values: normalized sum of probabilities of words on the given level $\check{P}$ and normalized density-of-states function $\check{\rho}$. Thus, on the one hand, $S_{S,M}$ allows one to estimate topic model parameters, for example, such as regularization parameters in LDA Gibbs sampling model and ARTM model, and the number of topics based on searching for a minimum of $S_{S,M}$, which, in turn, is characterized by the difference of entropies between the initial distribution and the distribution obtained in the result of modeling. On the other hand, it allows us to estimate what contribution to entropy is added by Jaccard distance between two different topic solutions with different parameter values and the number of topics. Correspondingly, the best values of topic model parameters correspond to the situation when entropy reaches its minimum, and the worst values correspond to entropy maximum.

2.5. Hierarchical entropic topic model

Textual collections may contain a flat topic structure or hierarchical structure. Currently, there are no methods for determining the type of structure except the entropic model proposed in work [2]. The general idea of determining the structure is as follows. As it was demonstrated earlier [46], a dataset may possess several local minima of parameterized entropy, which correspond to different numbers of topics. Correspondingly, these minima may be associated with different hierarchical levels. Thus, the number of minima may be a marker of a particular topic structure. If a dataset has only one minimum, it has only one level of topics; if a dataset has two minima, one can assert that it has two levels of hierarchy. Based on the above, the considered entropic topic model should be extended for hierarchical models in the following way [2]. Since the hierarchical structure in TM can be represented as a graph, where each node represents one topic, the procedure of hierarchical TM leads to the construction of a hierarchical tree with a fixed number of topics on each level. Each node-topic has a list of words and documents with probabilities of belonging to this topic. The total number of words on each level is a constant that equals the total number of elements W in the statistical system. The set of nodes-topics on one level is represented with matrix Φ (distribution of words by topics).

The procedure of hierarchical topic modeling consists of the construction of a sequence of matrices Φ, in which the number of words is constant, but the number of topics is sequentially increasing (from one hierarchical level to another). Correspondingly, the portion of words with probabilities above $1/W$ is changes during the transition from one level to another in hierarchical topic modeling. Thus, each hierarchical level is characterized by the following parameters: 1. The number of topics T_i on level i. 2. The number of words with probabilities above the threshold $1/W$ on level i: $N_i = \sum_t N_{it}\left(\phi_{it} > \frac{1}{W}\right)$,where W is vocabulary size, t refers to summation for all topics. 3. Sum of probabilities of words $\tilde{P} = \sum_{t=1}^{T_i} \phi_{it} \left(\phi_{ti} > \frac{1}{W}\right)$. Based on the above values, one can determine internal energy and Shannon entropy (S) of the current level with respect to the equilibrium state of this level: $E_i = -ln(\tilde{P}/T_i)$, $S_i = ln(\frac{N_i}{WT_i})$, where i is the level number. Further, one can define free energy and Renyi entropy of i-th level by means of S_i and $E_i$as follows: $\Lambda_{F_i} = E_i - T_i \cdot S_i$. Renyi entropy of i-th level is expressed through the free energy of i-th level in the following way: $S_i^R = \frac{\Lambda_{F_i}}{T-1}$, where $q = 1/T_i$ is a parameter characterizing each hierarchical level.

Therefore, by measuring the entropy value on each hierarchical level under variation of model parameters (including the number of levels) for a given dataset, one can estimate the process of hierarchical model construction in terms of the behavior S_i^R under transition from one level to another, i.e., to estimate the dependence of entropy on the number of topics and parameter values. The process of clustering words by topics starts with entropy maximum, when all elements (words) of the statistical system are related to one or two topics, and also ends with entropy maximum, where all elements are related to all topics (for a large number of topics) with approximately the same probabilities. The location of a global minimum and a set of local minima of Renyi entropy in terms of the number of topics is determined by dataset features. Renyi entropy S_i^Rserves as a measure of the degree to which the given system is non-equilibrium, where entropy minimum corresponds to information maximum, and the number of Renyi entropy minima serves as a marker of topical structure.

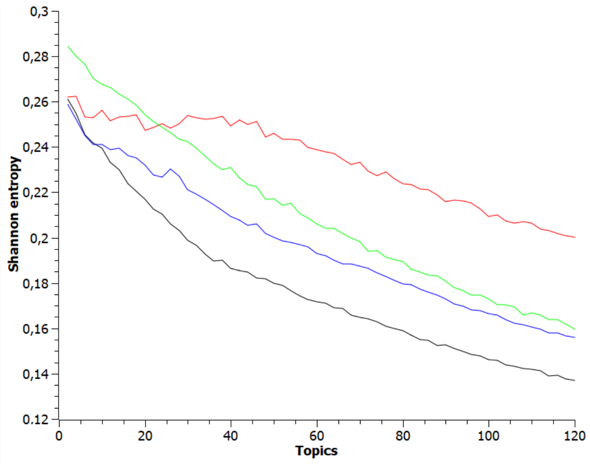
Let us note that this principle was used for tuning the hierarchical clustering procedure (based on the 'complete' method) [57] when clustering users of the social network VK.

2.6. Experimental testing of application of Renyi and Tsallis entropy in topic models

In this work, four topic models were investigated in terms of behavior of Renyi and Tsallis entropy as functions of the number of topics: 1. LDA GB. 2. Granulated LDA (GLDA GB). 3. PLSA (E-M algorithm). 4. Variational LDA (E-M algorithm). The choice of these models is due to the following reasons. First, these models are used as baselines in many articles in the field of topic modeling, Second, these models represent the main types of topic model inference algorithms. In each experiment, the number of microstates with probabilities $p_{tn} > \frac{1}{W}$ was computed for each model. Then, the density-of-states function, internal energy, entropy, and free energy were calculated for each model in dependence on the number of topics. Renyi and Tsallis entropies were calculated for each topic solution based on free energy.

Datasets: 1. 'Live Journal' dataset. This is a set of Russian-language posts from the social network 'Live Journal', size: 101481 posts; vocabulary size: 172939 unique words. The number of topics was varied in the region T = [2; 330] with the increments in two topics. 2. English-language dataset '20 newsgroups' [58]. Size: 15404 posts and N=50948 unique words, marked-up on 20 topics. The number of topics for the second dataset was varied in the region T = [2; 120] with the increments in two topics. The choice of these datasets is due to the following reasons. First, these datasets are in different languages, which allows us to demonstrate the cross-language applicability of entropic topic models and establish common model features for different languages. Second, different sizes of the collections show that changing size may lead to the appearance of additional local minima. Moreover, different clustering models were tested on the above English-language collection [59], which allows us to compare the results of topic modeling with cluster analysis results.

Figures (1) and (2) demonstrate Shannon and Renyi entropies for four models (’20 newsgroups’ dataset). Each model was run three times; then, the results were averaged. Entropies were calculated based on the averaged results. Averaging of modeling results is related to considering the instability of topic models.



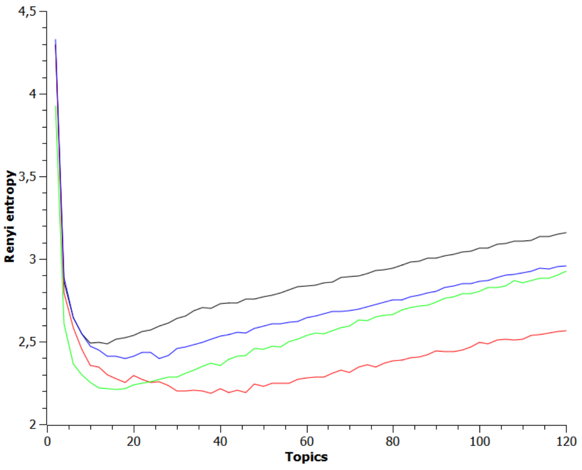


Fig. 1. Distribution of Shannon entropy.

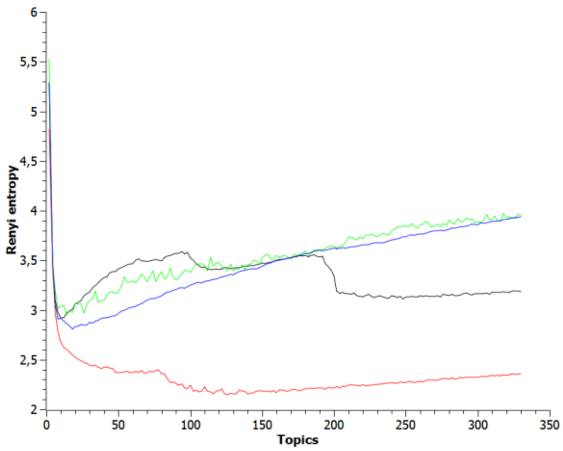
Fig. 2. Distribution of Renyi entropy.

Distributions of Shannon and Renyi entropies as functions of the number of topics for ’20 newsgroups’ dataset. LDA (Gibbs sampling): black, GLDA (Gibbs sampling): red, PLSA (E-M algorithm): blue, LDA (E-M algorithm): green.

Renyi entropy in contrast to Shannon entropy has a global minimum and demonstrates correct results for boundary values of the number of topics. For T→1 Renyi entropy has a maximum since topic modeling as any other clustering algorithm does not provide distribution of clusters, i.e. information is close to zero. At the same time, increasing the number of clusters/topics (i.e., T→∞) leads to uniform distribution of each word on topics, that corresponds to increasing entropy. However, different models provide slightly different locations of Renyi entropy minimum and different values of this minimum. To determine which of the above models provides more accurate results, one has to compare the results of TM with the cluster analysis results on the same collection. The authors of work [59] tested several clustering algorithms on ’20 newsgroups’ dataset and demonstrated that the optimal number of clusters varies in the region of 15-20 clusters for different algorithms due to the correlation of some topics.

Models LDA (Gibbs sampling, LDA GB) and LDA (E-M algorithm) demonstrate that the optimal number of topics is about 15, PLSA (E-M) model gives 20 topics. However, the GLDA model provides a significantly different number of topics, which is almost two times larger compared to LDA (Gibbs sampling) and LDA (E-M algorithm) models. This is related to the fact that in GLDA model, strong averaging is present in the sampling procedure, which leads to high stability, but shifts the global minimum of Renyi entropy.

The results of calculations of Renyi and Tsallis entropies for four topic models on the Russian-language dataset are presented in figures (3) and (4).



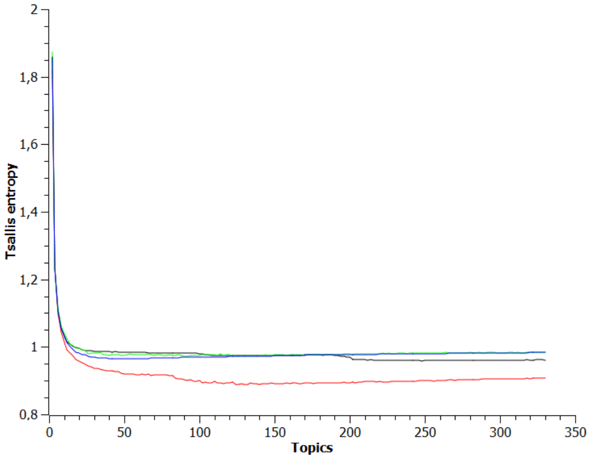


Fig. 3. Distribution of Renyi entropy. Fig. 4. Distribution of Tsallis entropy. Distribution of Renyi and Tsallis entropies as functions of the number of topics for 'Live Journal' dataset: LDA (Gibbs sampling): black, GLDA (Gibbs sampling): red, PLSA (E-M algorithm): blue, LDA (E-M algorithm): green.

The calculations demonstrate that models based on E-M algorithm demonstrate a substantial difference from models based on Gibbs sampling for the Russian-language dataset, especially for a large number of topics (over 100). LDA GB model demonstrates the presence of large jumps in Renyi entropy, which are related to significant fluctuations of the density-of-states function. However, LDA (E-M algorithm) model and PLSA (E-M algorithm) model do not see these jumps. Fluctuations in the density-of-states function in models with Gibbs sampling cannot be explained by features of the sampling procedure, since in work [45], investigation on the same dataset was implemented, where LDA (Gibbs sampling) was run three times for each number of topics, and the number of topics was varied in the range [105 - 120] with the increments in one topic and the range [120-600] with the increments in ten topics. A jump in the region of [110 – 120] topics was observed for all runs of the model. Thus, models based on Gibbs sampling are more sensible with respect to other models. Tsallis entropy calculated for LDA(Gibbs sampling) model also demonstrates a jump in the region of [110-120] topics and in the region of [190-200] topics, however, the amplitude of the jump is significantly smaller. This is because Tsallis entropy is more stable in terms of Leshe [60].

Based on the implemented calculations, one can conclude the following. First, parameterized Renyi entropy is suitable for determining the number of topics in textual datasets since its minimum corresponds to the results of human mark-up. The number of topics is the entropy parameter. Second, different topic models demonstrate different numbers of minima of parameterized entropy, but the location of a global minimum for different models is almost the same. Dataset features characterize the locations of minima and their number.

2.7. Numerical experiments on semantic stability of topic models

The indistinguishability of particles is an important factor when describing different physical statistical systems. It makes it possible to use a combinatorial approach for the calculation of the number of states and the estimation of the probability density function. In this case, it is not important which particles exactly reside in states with high probabilities. However, in the case of information systems consisting of many documents, a topic is formed from many different words and the semantic differences between words are important. Therefore, when investigating the behavior of textual systems, it is necessary to verify how reproducible are distributions of words in a semantic point of view under variation of the number of topics hyperparameters. In this work, semantic reproducibility in TM of two clouds of words T_1 and T_2 (corresponding to two different topics) was measured according to Jaccard distance.

Jaccard distance was calculated by pair-wise comparison of each topic solution with all other topic solutions and was stored in a matrix where each element contains a value of Jaccard distance $J_{t1,t2}$, where $t1, t2$ are topic numbers.

Figures (5) and (6) demonstrate curves of diagonal Jaccard distances for LDA (Gibbs sampling) and LDA (E-M algorithm) for the Russian-language dataset. The values of Jaccard distances are not provided for the English-language dataset since all the models demonstrated almost the same values, about 0.99.

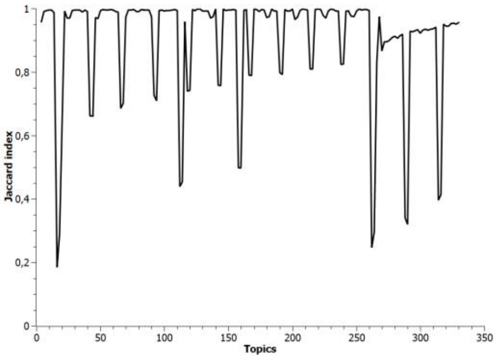


Fig. 5. Behavior of Jaccard distance for LDA Gibbs sampling model.

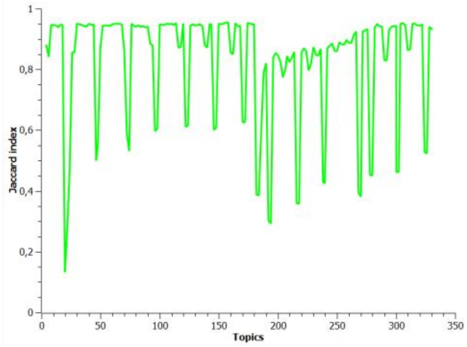


Fig. 6. Behavior of Jaccard distance for LDA (E-M algorithm) model.

The distribution of Jaccard distances demonstrates that models of both types have areas of semantic stability. Moreover, there are areas with a high level of the coefficient $J_{t1,t2} \cong 0.9$ and areas with a lower level $J_{t1,t2} \cong 0.5$. However, if a significantly large number of top words is used in each topic solution, for example, 1000 words, then such periodical structure almost disappears.

2.8. Experimental testing of Sharma-Mittal entropy and Renyi entropy as meausers of quality for estimating the number of topics and semantic coherence of topic models.

2.8.1. Experiments on application of Renyi entropy

In this part of the work, the possibility of determining the optimal parameter values in topic models was investigated. The investigation was implemented for the following models: 1. LDA Gibbs sampling (LDA GB) [20], 2. pLSA (E-M) [21], 3. ARTM with regularizers ‘sparse Φ’ and ‘sparse Θ’ [16]. Parameters of LDA GB are α, β, which characterize Dirichlet distribution, and T is the number of topics. Parameters of ARTM model are regularizer coefficients of sparsing matrices Φ, Θ, and the number of topics. PLSA model has only one parameter, which is the number of topics. Therefore, this model was compared to two other models.

Datasets: 1. ‘20 newsgroups dataset’ (human mark-up, the range is [15 - 20] topics). 2. Russian-language dataset (‘lenta_ru’) (custom mark-up for 10 topics). Analysis of the correlation of topics demonstrates that the ‘real’ number of topics is in the range [7-10]. The dataset size is 82852 documents, vocabulary size is 172939. All datasets were calculated for each model under variation of parameters. Then, the density-of-states function, distribution of Jaccard distances, Renyi entropy, and log-likelihood were calculated for each obtained topic solution. Moreover, two-parametric Sharma-Mittal ($S_{S,M}$) entropy was calculated for the estimation of semantic stability of topic solutions under variation of model hyperparameters, including the number of topics.

2.8.1.1. pLSA and LDA GB models: Renyi entropy

Renyi entropy curves for pLSA and LDA GB for two datasets are given in figures (7), (8).

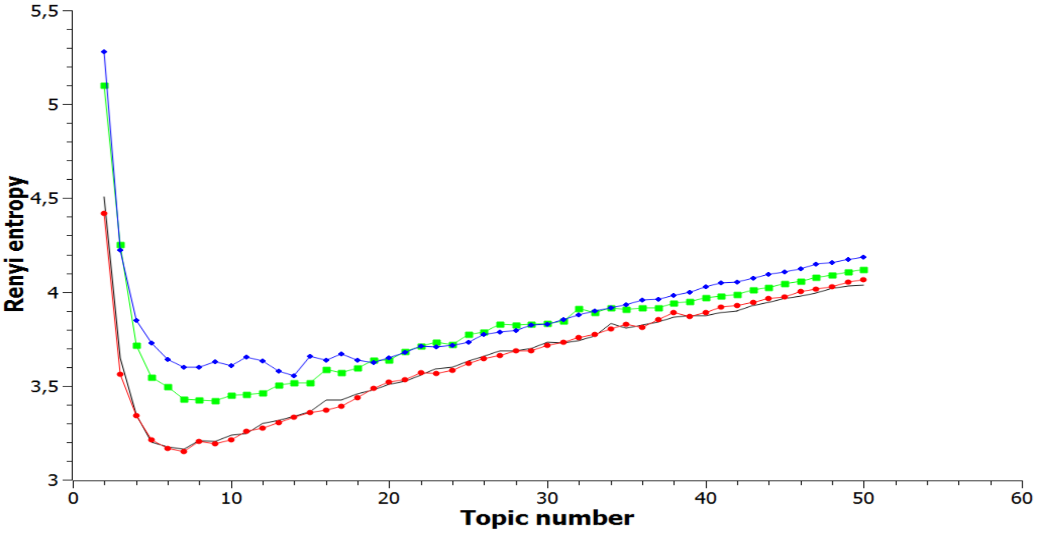


Fig. 7. Renyi entropy (‘Lenta_ru’). pLSA: black, LDA GB (α=0.1, β=0.1): red, LDA GB (α=0.5, β=0.1): green, LDA GB (α=1, β=1): blue.

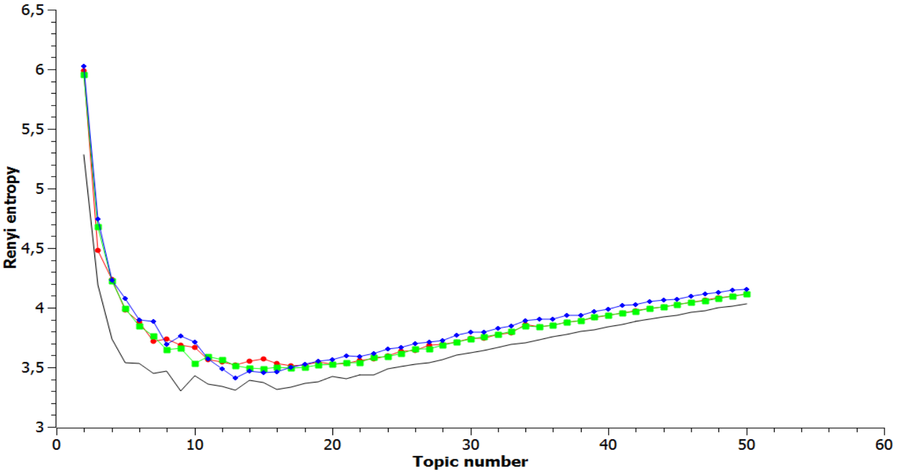


Fig. 8. Renyi entropy (’20 newsgroups’). pLSA: black, LDA GB (α=0.1, β=0.1): red, LDA GB (α=0.5, β=0.1): green, LDA GB (α=1, β=1): blue.

Figures (7) and (8) demonstrate that Renyi entropy of pLSA model and LDA Gibbs sampling model with parameters α=0.1, β=0.1 are very close to each other. Increasing regularization parameters α, β leads to increasing Renyi entropy. Moreover, minimum of parameterized entropy is shifted. Figure (9) demonstrates curves of log-likelihood as a function of the number of topics. One can see that increasing values α, β worsens log-likelihood, which is equivalent to increasing entropy. Thus, comparing the behavior of log-likelihood and Renyi entropy curves, one can conclude the following: 1. Renyi entropy is suitable for tuning topic model parameters, and Renyi entropy minimum corresponds to the optimal parameter values of the considered topic models. 2. Renyi entropy, based on its local minimum, allows us to determine the optimal number of topics in contrast to log-likelihood.

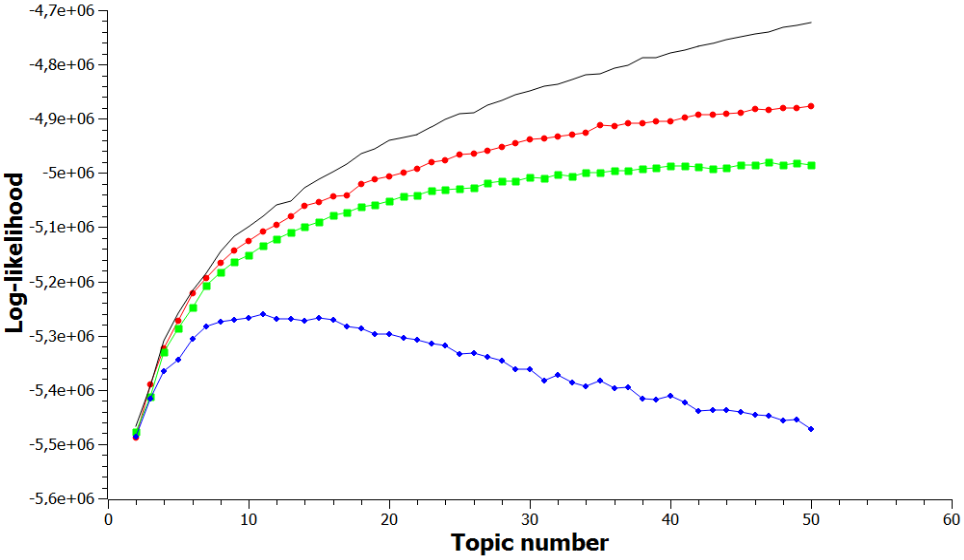


Fig. 8. Log-likelihood (‘lenta_ru’). pLSA: black, LDA GB (α=0.1, β=0.1): red, LDA GB (α=0.5, β=0.1): green, LDA GB (α=1, β=1): blue.

2.8.1.2. ARTM model with sparsing of matrix Φ: Renyi entropy

The result of topic modeling based on ARTM model significantly depends on regularization coefficients [16]. Increasing these values may lead to a significant change in the stability level of the topic model [43]. Based on the above, in this part of the work, the effect of the influence of regularization parameter Φ (τ_Φ) and the number of topics on the behavior of Renyi entropy under variation of ARTM model hyperparameters is analyzed. The number of topics was varied in the range [2-50], and the value of τ_Φ was varied in the range [-10, 10] when investigating this model. The set of Renyi entropy curves as functions of the number of topics is presented in figure (9). One can see that increasing parameter τ_Φ leads to shifting Renyi entropy minimum in the region of a small number of topics (about 2), which is significantly smaller than the ‘real’ number of topics (7-10). Thus, strong regularization of sparsing matrix Φ leads to an incorrect number of topics. Let us note that changing the sign of the regularization coefficient does not influence the modeling results. Renyi entropy curve of this model for the English-language dataset ’20 newsgroups’ is given in figure (10). Increasing regularization parameter leads to significant shifting Renyi entropy minimum. For τ_Φ = 1, the minimum is shifted to 10 topics, however, the real number of topics is 14-17. Further increment of τ_Φ leads to model deterioration. Thus, the best result of the topic model corresponds to the minimal value of the regularization coefficient, and Renyi entropy curve almost coincides with an analogous curve for pLSA model.

2.8.1.3. ARTM model with sparsing of matrix Θ: Renyi entropy

In this model, parameters are the regularization coefficient of matrix Θ (τ_Θ) and the number of topics. In contrast to the previous model, in this case, sparsing is implemented for the matrix of the distribution of topics in documents. In the experiments, the number of topics was varied in the range [2-50], and coefficient τ_Θ was varied in the range [-10, 10]. The set of Renyi entropy curves for this regularizer as functions of the number of topics is given in figure (11) ('lenta' dataset). Renyi entropy curves for regularization coefficients τ_Θ = [0.01, 0.1, 1] are almost indistinguishable. However, coefficient τ_Θ =10 does not allow us to calculate free energy and Renyi entropy since the model deteriorates (analogously to the previous one). Analogous result is obtained for '20 newsgroups' dataset. Thus, the best result is obtained with a small value of the regularization coefficient, since its increment leads to a significant decrease of log-likelihood and to an increment of Renyi entropy.

2.8.2. Experiments on application of quality measure based on Sharma-Mittal entropy to topic models.

In the framework of this set of experiments, an investigation of the behavior of two-parametric Sharma-Mittal entropy under variation of hyperparameters was implemented for pLSA, ARTM, and LDA Gibbs sampling models. The application of this type of parameterized entropy makes it possible to estimate changes in the semantic component of topic models to the level of entropy under variation of hyperparameters. The chosen topic models are most frequently used models in English-language and Russian-language scientific literature.

2.8.2.1. PLSA model: Sharma-Mittal entropy

For the calculation of two-parametric entropy $S_{S,M}$, first of all, pairwise values of Jaccard distances were calculated under variation of the number of topics. Examples of these calculations are visualized in the form of heat maps in figures (12), (13). Behavior of $S_{S,M}$ entropy curves for PLSA model (for two datasets) are given in figure (14). Large jumps of Sharma-Mittal entropy are due to small values of Jaccard distances.

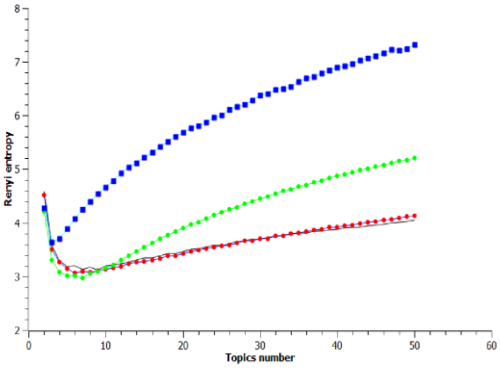


Fig. 9 ('Lenta' dataset).

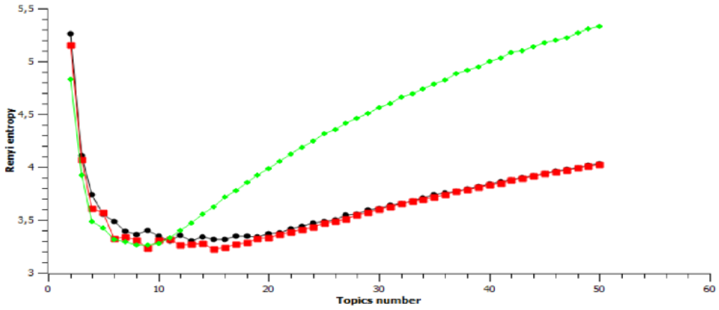


Fig. 10 ('20 newsgroups dataset).

Renyi entropy curves ('Lenta' dataset,'20 newsgroups' dataset) under variation of regularizers τ_Φ 'sparse Φ' (ARTM). Black: τ_Φ =0.01, red: τ_Φ =0.1, green: τ_Φ =1, blue: τ_Φ =10

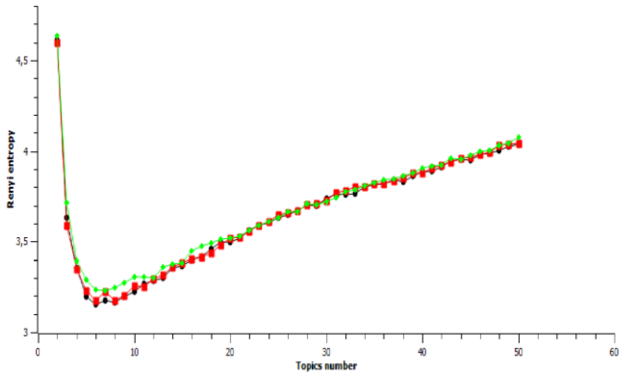
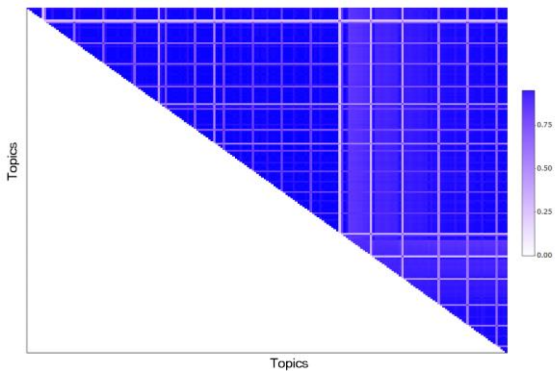


Fig. 11. Renyi entropy curves (‘Lenta’ dataset) under variation of regularizer τ ‘sparse Θ’ (ARTM). Black: τ_{Θ} =0.01, red: τ_{Θ} =0.1, green: τ_{Θ} =1.



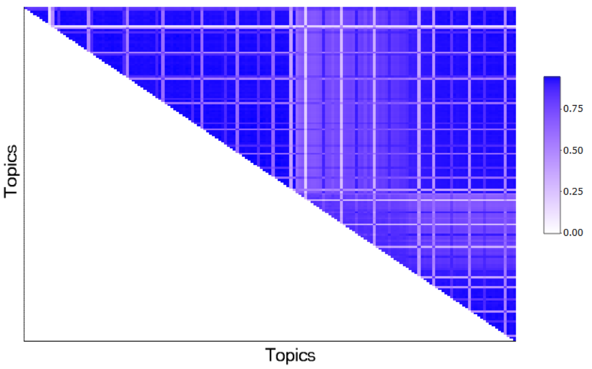


Fig. 12 (LDA Gibbs sampling). Fig. 13 (VLDA (E-Malgorithm)).

Jaccard distances for LDA Gibbs sampling model and VLDA (E-M algorithm) model.

However, two-parametric entropy also possesses a minimum, which allows us to find the optimal number of topics. Figures (15), (16) demonstrate entropy curves of $S_{S,M}$for PLSA model with truncated peaks (for the purposes of visualization of minima, since $S_{S,M}$ has large jumps for small values of Jaccard distance). These figures demonstrate that for the Russian-language dataset the minimum of two-parametric entropy lies in the region of [7-10] topics, and for the English-language dataset the minimum is in the region of [18-20] topics, which completely corresponds to the human mark-up.

2.8.2.2. LDA GB model: Sharma-Mittal entropy

The results of calucations of $\mathbf{S_{S,M}}$ entropy for LDA GB model in comparison to PLSA model are given in figures (17), (18). They demonstrate that two-parametric entropy also allows us to correctly estimate the ‘real’ number of topics for datasets in two different languages. Moreover, increasing the value of regularization coefficients α, β leads to increasing entropy and shiftinig minimum, that violates the possibility to correctly determine the number of topics in a dataset. Thus, one can conclude the following. First, $\boldsymbol{S_{S,M}}$ entropy allows us to correctly determine the optimal number of topics for datasets in different languages. Second, $\boldsymbol{S_{S,M}}$ entropy allows us to correctly choose hyperparameters of LDA GB model.

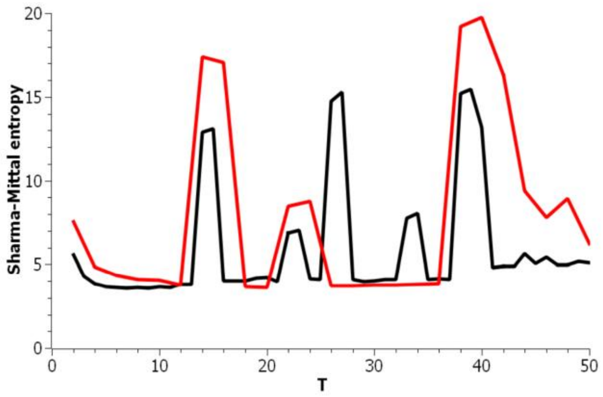
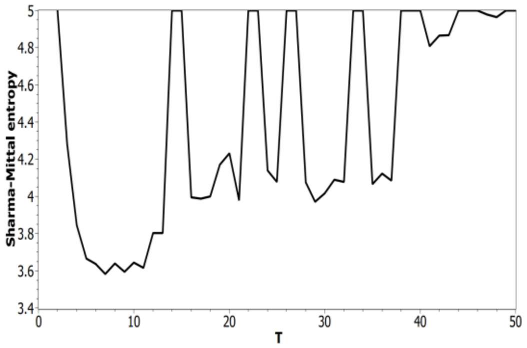


Fig. 14. Curve of $S_{S,M}$entropy for ‘Lenta’ dataset and ’20 newsgroups’ dataset (PLSA model) for diagonal elements of the matrix with Jaccard distances. Russian-language dataset: black; English-language dataset: red.



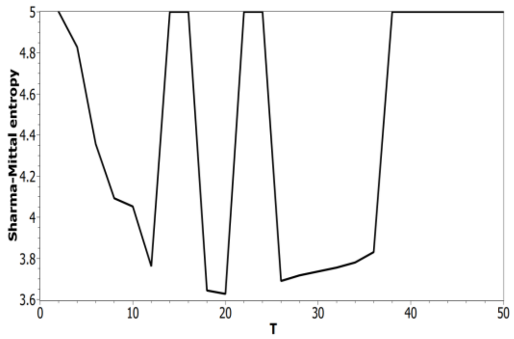
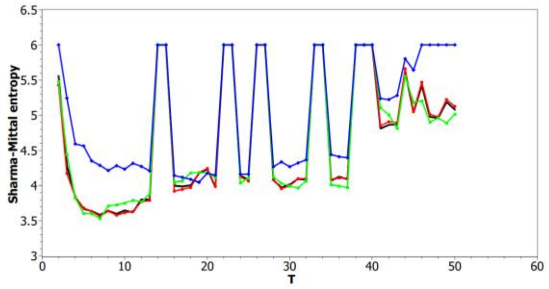


Fig. 15 $S_{S,M}$ (‘Lenta’ dataset). Fig. 16 $S_{S,M}$ (’20 newsgroups’ dataset).

Curves of $S_{S,M}$entropy for ‘Lenta’ and ’20 newsgroups’ datasets, PLSA model with truncated peaks.



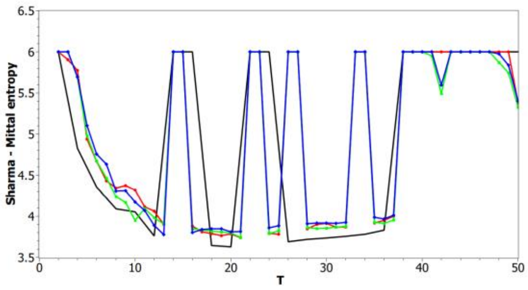


Fig. 17 $S_{S,M}$ (‘Lenta’ dataset). Fig. 18 $S_{S,M}$ (’20 newsgroups’).

Curves of $S_{S,M}$entropy (LDA GB vs pLSA) in dependence on the number of topics. (‘Lenta’, ’20 newsgroups’ datasets). PLSA: black; LDA (α = 0.1, β = 0.1): red; LDA (α = 0.5, β = 0.1): green; LDA (α = 1, β = 1): blue. Peaks are truncated.

2.8.2.3. ARTM model with sparsing of matrices Φ and Θ: Sharma-Mittal entropy

ARTM model is implemented based on the principle of additive regularization, where the regularization coefficient, which the user sets, determines the level of contribution of the specified regularizer to the result of topic modeling. Currently, there is no suitable method for determining the optimal coefficient value. Therefore, this work aims to demonstrate experimentally the possibility of application of parameterized entropy for tuning regularization coefficients in ARTM model.

In this model, parameters are values of regularization coefficients and the number of topics. Correspondingly, for the investigation of this model, the number of topics was varied in the range [2-50] and coefficients $\tau_{\Phi}, \tau_{\Theta}$ were varied in the range [-10, 10]. The set of entropy curves as functions of the number of topics is given in figures (19), (20). Increasing parameters $\tau_{\Phi}, \tau_{\Theta}$ in $S_{S,M}$ entropy as well as for Renyi entropy leads to increasing the total value of entropy, i.e., to worsening of topic model performance.

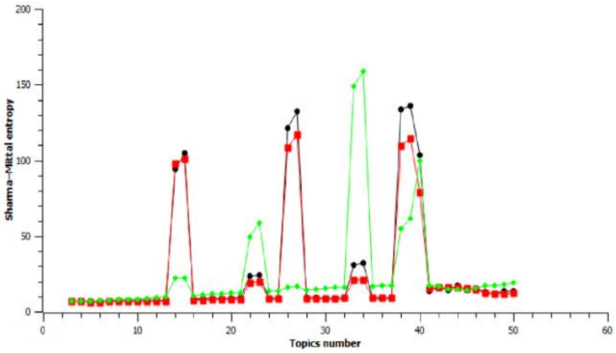


Fig. 19 $S_{S,M}$ entropy (sparse Φ).

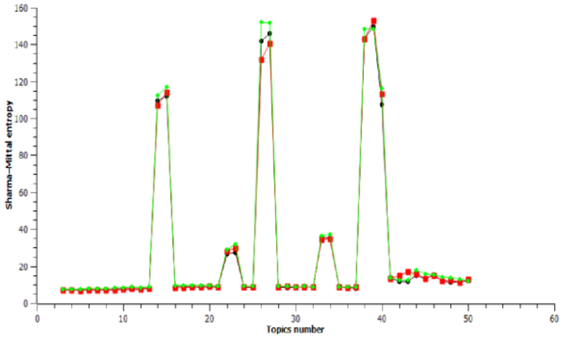


Fig. 20 $S_{S,M}$entropy (sparse Θ).

Curves of $S_{S,M}$ entropies for ARTM model with regularizers sparse Φ and sparse Θ. Black: τ_{Φ}, τ_{Θ}=0.01, red: $\tau_{\Phi}, \tau_{\Theta}$=0.1, green: $\tau_{\Phi}, \tau_{\Theta}$=1.

Thus, based on the analysis of the implemented computer experiments on marked-up datasets one can say the following: 1. Under variation of parameter $q = 1/T$, $S_{S,M}$ entropy and Renyi entropy allow us to determine the optimal number of topics and to choose the optimal value of regularization coefficient; 2. Variation of parameter r (Jaccard distance) in $S_{S,M}$ entropy leads to the appearance of areas of semantic stability, which are separated by peaks with large entropy values. However, the value of the jump depends on the number of words that are used for the calculation of Jaccard distance. 3. Minimum parameterized entropies for small values of parameterization coefficients correspond to the human mark-up of text collections.

2.9. Experiments on application of Renyi entropy to the analysis of hierarchical topic models

As it was noted, in the field of topics modeling, in addition to the problem of determining the optimal number of topics, the problem of determining ‘flat’ or hierarchical topical structure exists. This chapter presents the results of an experimental analysis of the behavior of three hierarchical models for marked-up datasets in different languages. In the experiments, the possibility of application of parameterized Renyi entropy as a marker of topical structure and for determining the optimal number of topics on different hierarchical levels is demonstrated.

To test the theoretical concept described in paragraph 2.5, the following experiments were conducted. First, the following models were used in computer experiments on the application of Renyi entropy for analysis of hierarchical topic models: 1. HLDA (model of hierarchical latent

Dirichlet allocation) [61]. 2. HPAM (model of hierarchical Pachinko allocation) [62]. 3. hARTM (hierarchical additive regularization of topic models) [63]. These models were tested by means of six marked-up datasets, two of which have a flat structure and the other four have a two-level structure.

Description of datasets: 1. Russian-language dataset ('Lenta_ru') (custom mark-up on 10 topics). 2. English-language dataset '20 newsgroups' [58] (custom mark-up on 20 topics). 3. 'WoS' has a hierarchical mark-up with two levels. It contains 46.985 annotations of published articles (*Web of Science*) and 80.337 unique words. The first level of mark-up contains 7 topics (computer science, electrical engineering, psychology, mechanical engineering, civil engineering, medical science, and biochemistry), and the second level contains 134 topics. Let us note that this dataset is highly unbalanced in terms of the distribution of documents by topics on the second level; therefore, in this work, we also consider its balanced subset. To balance this dataset, topics with less than 260 documents were removed. The balanced 'WoS' dataset contains 11.967 annotations of articles and 36.488 unique words, 7 topics on the first level and 33 topics on the second level. 4. 'Amazon' dataset (https://data.mendeley.com/datasets/9rw3vkcfy4/1) has a hierarchical mark-up with three levels, containing 6, 64, and 510 topics correspondingly. It contains 40.000 reviews on products from online shop *Amazon* and 31.486 unique words. The third level contains empty labels; therefore, in this work, only the first two levels of hierarchical mark-up are considered. Also, its balance version, which contains 6 topics on the first level and 27 topics on the second level, is considered. The total number of documents is 32.774, and the number of unique words is 28.422.

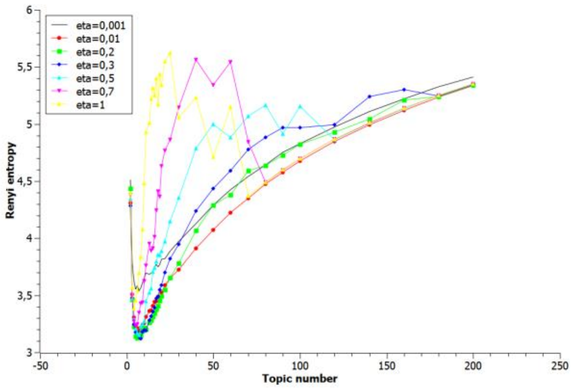
2.9.1. HPAM model

Hierarchical model HPAM depends on the following parameters: 1. The number of topics on the second level. 2. The number of topics on the third level. 3. Parameter 'eta' (η is the parameter characterizing the Dirichlet function). 4. Parameter 'alpha' (α). Let us note that the number of topics on the first level is always equal to one in HPAM model. Moreover, parameter α is set in the form of the initial value, which is further tuned by the algorithm. The investigation demonstrated that variation of the initial value of parameter α does not influence the modeling results; therefore, parameter α was not used in this work. Parameters of HPAM model for datasets with the flat topical structure were tuned in two stages. In the first stage, the number of topics on the third level was fixed; the number of topics on the second level and parameter η were varied. In the second stage, the number of topics on the second level and the value of η were chosen and fixed in such a way that led to minimum Renyi entropy at the first stage, and the number of topics on the third level was varied. According to the model's authors, the first level has one topic.

2.9.1.1. 'Lenta' dataset

For this dataset, the following experiments were conducted. In the first stage, the number of topics on the first and the third level were set to one, the number of topics on the second level was varied in the range [2-200]. The value of η was varied in the range [0.001-1]. Since topic modeling possesses a certain level of instability, all calculations were carried out 6 times (for a given combination of parameters), and Renyi entropy was averaged. Then for analysis of topic model behavior of the third hierarchical level, the best combinations of parameters that correspond to minimal values of Renyi entropy on the second level were selected. For these parameters, Renyi entropy on the third hierarchical level was calculated.

The results of Renyi entropy calculations as a function of the number of topics and parameter η on the second hierarchical level for HPAM model are given in figure (21).



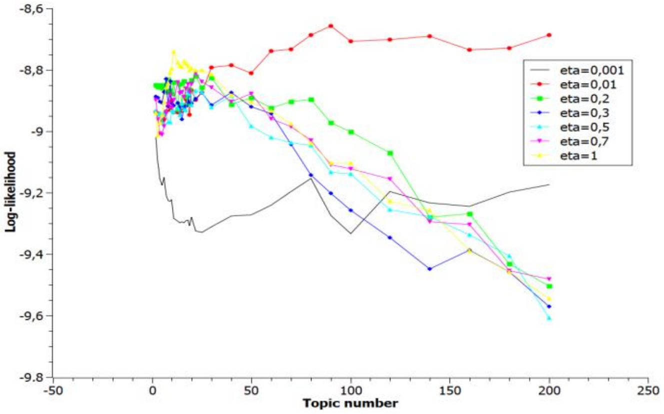


Fig. 21. Renyi entropy. Fig. 22 Log-likelihood.

Dependence of Renyi entropy minimum and log-likelihood on parameter η and the number of topics (Lenta) on the second hierarchical level of HPAM model.

Figure (22) demonstrates curves of log-likelihood for HPAM model. One can see that they are not suitable for analysis since this measure has very large fluctuations that do not allow us to determine the number of topics in the dataset nor to find the optimal value of parameter η. Moreover, since perplexity is an inverse value of log-likelihood, it is also unsuitable for the actual tuning of HPAM model.

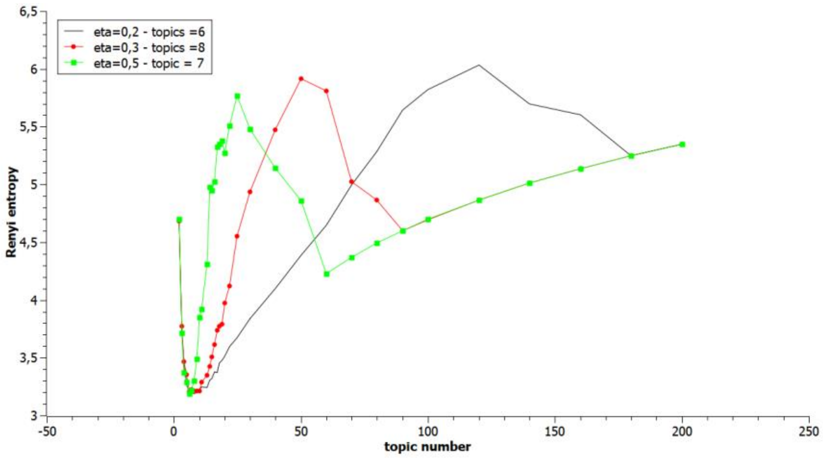
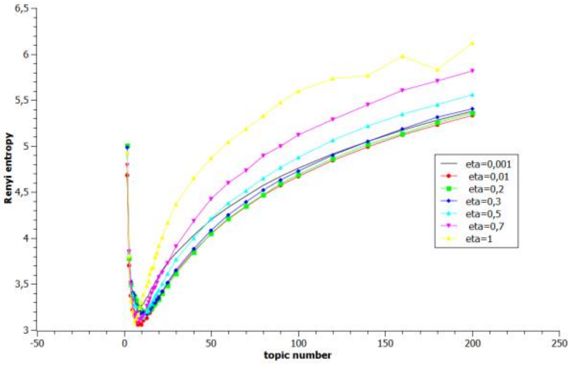


Fig. 23. Dependence of Renyi entropy on the number of topics on the third level under fixed number of topics on the second level and specified parameter η. HPAM model (Lenta).

Calculation of entropy on the third level demonstrates that variation of the number of topics leads to the presence of one global minimum in the region of 6 topics and sharp fluctuations of the entropy when increasing the number of topics above 50. Sharp changes in entropy are replaced by almost straight lines. This is related to the fact that in the region of strong fluctuations the model deteriorates: the number of words with high probabilities and the sum of probabilities becomes constant, and entropy increasing is explained only by the fact that the formula of entropy calculation contains the number of topics. Thus, the number of topics is increasing, but the statistical features of the model do not change. Therefore, HPAM model can see one global minimum for a small number of topics.

2.9.1.2. ’20 Newsgroups’ dataset

HPAM model for ‘20Newsgroups’ was investigated in the same way as for the Russian-language dataset. Renyi entropy curves as functions of the number of topics on the second hierarchical level for different values of parameter η are presented in figure (24). In general, their behavior under variation of the number of topics and parameter η is analogous to Renyi entropy curves for ‘Lenta’ dataset.



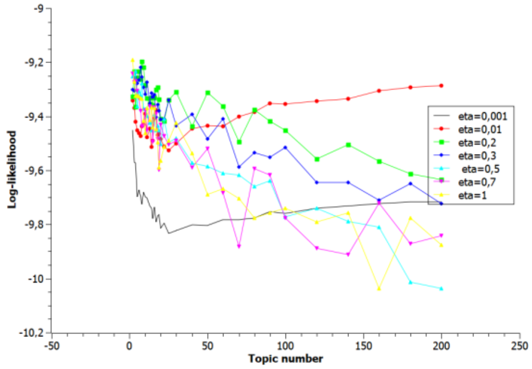


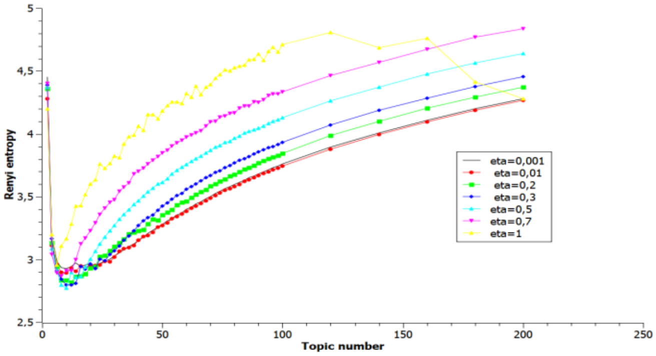
Fig. 24. Renyi entropy. Fig. 25 Log-likelihood

Dependence of Renyi entropy minimum and log-likelihood on parameter η and the number of topics (’20 Newsgroups’) on the second hierarchical level of HPAM model.

Figure (25) demonstrates that log-likelihood is also unsuitable for tuning HPAM model for the English-language dataset ’20 Newsgroups’.

2.9.1.3. Balanced and unbalanced datasets WoS

For these datasets, at the first stage, calculations of HPAM model were implemented for the following range of parameters: 1. The number of topics was varied in the region [2-60] with the increments in two topics, 2. Parameter η was varied as follows: [0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1]. The number of topics in the first and third hierarchical levels was fixed equal to one. HPAM model was run 6 times for each combination of parameters. After that, the average value of Renyi entropy was calculated. Figures (26) and (27) demonstrate averaged curves of Renyi entropy for different values of parameter η under variation of the number of topics.



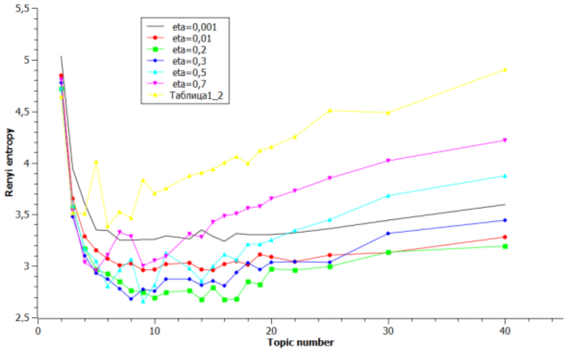
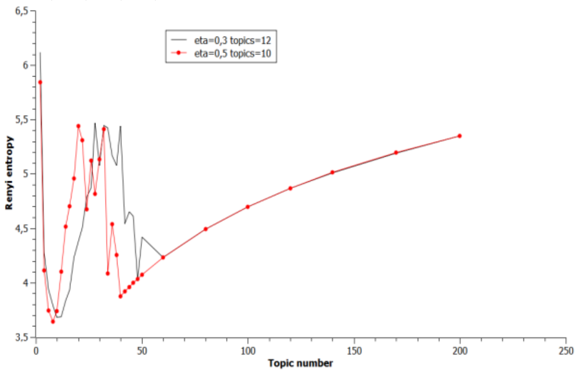


Fig. 26. Renyi entropy. Fig. 27 Log-likelihood

Renyi entropy on the second hierarchical level (balanced and unbalanced ‘WoS’).

The last two figures demonstrate that dataset balancing leads to the appearance of brightly expressed Renyi entropy minimum, i.e, dataset balancing improves topic modeling. Moreover, in this case, the accuracy of determining the number of topics us significantly higher.

Calculations on the third level were implemented under the fixed number of topics and the corresponding value of η from the second hierarchical level. Then, on the third level, the number of topics was varied for several values of η. The results of Renyi entropy calculations are given in figures (28), (29).



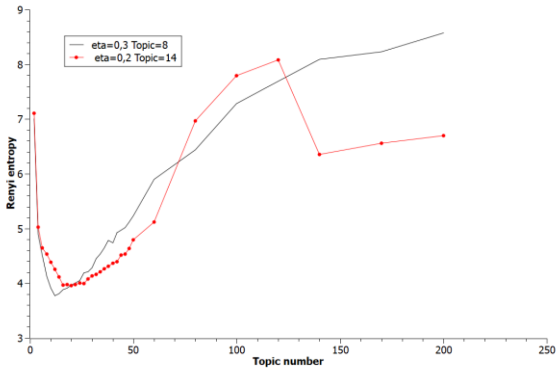


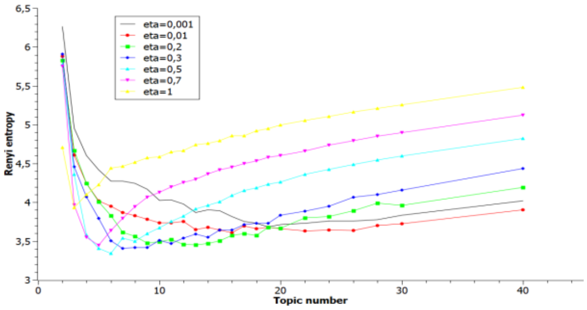
Fig. 28. Renyi entropy curves (balanced 'WoS' dataset).

Fig. 29. Renyi entropy curves (unbalanced 'WoS' dataset)

Renyi entropy curves under variation of the number of topics and parameter η on the third hierarchical level (balanced and unbalanced 'WoS' dataset)

2.9.1.4. Balanced and unbalanced 'Amazon' datasets

For these two datasets, the calculations were implemented analogously to calculations for 'WoS' datasets. The results are presented in figures (30), (31) (variation of the number of topics and parameter η on the second level). The results of Renyi entropy calculations on the third level are demonstrated in figures (32), (33).



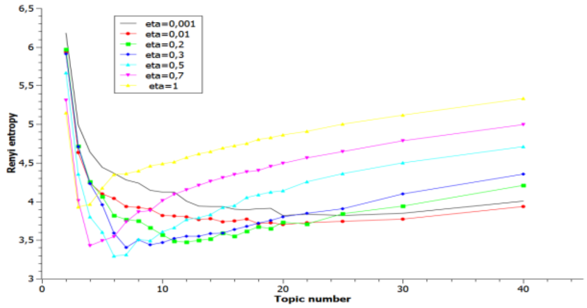
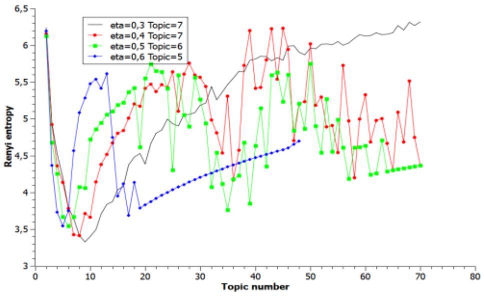


Fig. 30. Renyi entropy curves (balanced 'Amazon' dataset).

Fig 31. Renyi entropy curves (unbalanced 'Amazon' dataset)

Renyi entropy curves on the second hierarchical level (balanced and unbalanced 'Amazon' datasets) under variation of the number of topics and parameter η.



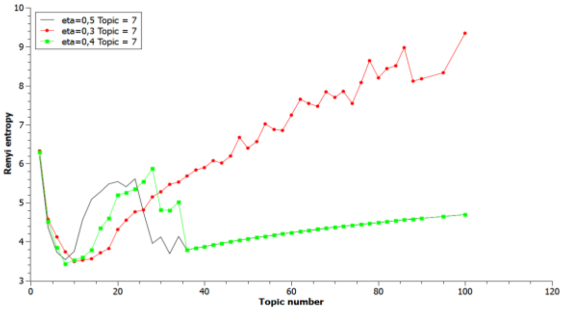


Fig. 32. Renyi entropy curves (balanced ‘Amazon’ dataset). Fig. 33. Renyi entropy curves (unbalanced ‘Amazon’ dataset).

Renyi entropy on the third level of HPAM model (balanced and unbalanced ‘Amazon’ datasetsc) under variation of the number of topics and parameter η.

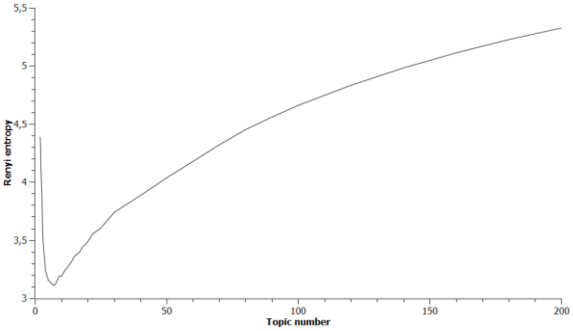
One can see that HPAM model gives sharp jumps of entropy both for ‘flat’ datasets and for datasets with hierarchical mark-up for a large number of topics. Therefore, one can conclude that HPAM model cannot differentiate between flat and hierarchical structures of datasets and can be used only for determining one level of hierarchy.

2.9.2. HLDA model

The authors of this model claim that their model finds the number of topics for a dataset automatically based on a hierarchical Chinese restaurant process [61]. However, as investigations demonstrate, this model significantly depends on the concentration parameter and leads to a large spread in the number of topics when this parameter is varied [2]. Since there is no possibility of correctly determining the concentration parameter, this model was not considered in this work. A complete investigation of this model is given in work [2].

2.9.3. hARTM model

hARTM model, proposed by the authors of work [63], has the following parameters: 1. The number of topics on each level. 2. Seed is a parameter characterizing the initialization procedure (setting random numbers generator). This model was investigated for four considered datasets. The results of Renyi entropy calculations for ‘flat’ datasets are presented in figures (34), and (35).



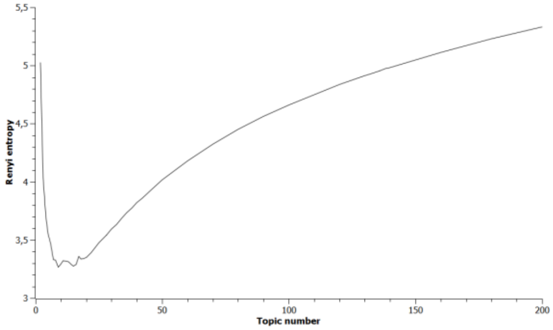


Fig. 34. Renyi entropy curves (‘Lenta’ dataset). Fig. 35. Renyi entropy curves (’20 Newsgroups’ dataset).

Dependence of Renyi entropy on the number of topics on the first hierarchical level in hARTM model (‘Lenta’, ’20 Newsgroups’).

The calculations demonstrate that hARTM model determines well the flat structure and does not possess fluctuations for large numbers of topics. The results of calculations for ‘WoS’ dataset are demonstrated in figures (36), (37). These curves demonstrate that, first, balancing the dataset leads to entropy decreasing in topic model; second, balancing leads to changing the location of the second entropy minimum. Moreover, the first minimum is almost not changed. It means that removing documents, which compose small topics, does not influence the set of words with high probabilities on the first level. The existence of the second hierarchical level is demonstrated by the presence of the second local minimum. Moreover, the balancing procedure impacts the minimum location for the second hierarchical level.

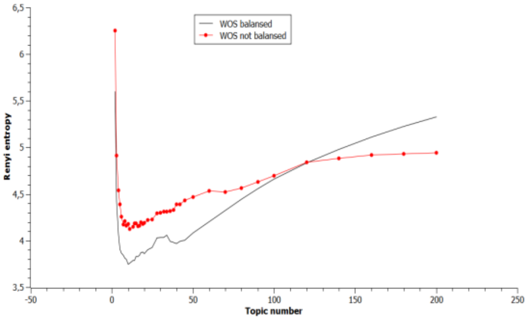


Fig. 36. Renyi entropy curves (‘WoS’ dataset).

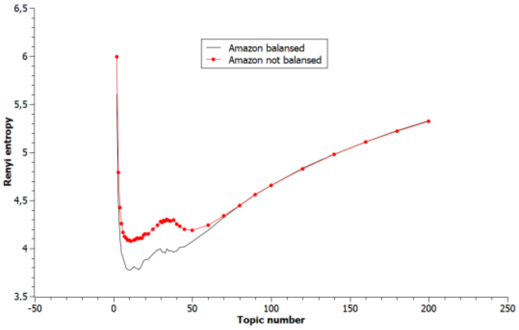


Fig. 37. Renyi entropy curves (‘Amazon’ dataset).

Renyi entropy curves for balanced and unbalanced ‘WoS’ and ‘Amazon’ datasets on the first hierarchical level. Black: balanced dataset, red: unbalanced dataset.

However, let us note that the second hierarchical level is determined with a smaller accuracy than the first one. This is related to the fact that words on the second level have smaller values of probabilities; therefore, determining the difference between words on the second level and words below the threshold $\frac{1}{W}$ is complicated due to the instability of topic models.

Thus, one can conclude the following based on the implemented investigations of entropic topic models. First, models based on Renyi and Sharma-Mittal entropy, namely, LDA (Gibbs sampling algorithm), pLSA (E-M algorithm), VLDA (E-M algorithm), GLDA (Gibbs sampling algorithm), and ARTM with sparsing regularizers for matrices Φ and Θ (E-M algorithm) allow us to determine optimal hyperparameters of topics models. The optimal number of topics in topic models is determined by searching the minimum of parameterized entropies. Variation of regularization parameters leads to shifting the whole entropy curve. Along with that, the best value of the regularization parameter corresponds to the lowest entropy curve (among all curves obtained under variation of the parameter). Second, the application of two-parametric entropy allows us to estimate the semantic stability of topics models under variation of model hyperparameters including the number of topics. Third, a hierarchical topic model based on hARTM makes it possible to determine the presence of a hierarchical or ‘flat’ structure in datasets in different languages and correctly set the optimal number of topics at two levels of the hierarchy.

3. Fractal model for estimating results of topic models

The behavior of an information statistical system can be investigated using a fractal model. This is because Renyi entropy describes fractal statistical systems well [50]. This mathematical formalism is based on the scaling procedure, that is, changing the scale. The fractal model can be described as follows [64]. Topic solution for a fixed number of topics is represented with matrix Φ, where the total number of cells is $T * W$, where T is the number of topics (columns in the matrix), W is the number of unique words (number of rows). Each cell of the matrix contains probability p_{ij} of belonging of word w_i to topic T_j, and the size of a cell equals $\varepsilon \sim 1/(WT)$. For a fixed vocabulary size ($W = const$), the size of the cell is determined only by the number of topics, and for $T \to \infty$, cell size tends to zero. Density-of-states function is $\breve{\rho} = \frac{N_i}{W\mathrm{T}}$, where N_i is the number of cells in the topic solution with probabilities (p_{ij}) above $\frac{1}{W}$, i.e., this function estimates the cloud of highly probable words and is a function of the number of topics. During the process of topic modeling, this function changes from 0 up to some value $\breve{\rho}_i(E)$<1, which depends on the number of topics. Correspondingly, density $\breve{\rho}(E)$depends on cell size and degree $D(\varepsilon)$ [64]: $\breve{\rho}(E) \cong \varepsilon^{-D(\varepsilon)}$. The distribution of fractal dimensions D(ε) was determined by means of 'box counting' algorithm [65]. Its application for calculations of fractal dimensions in TM consists of the following steps. 1. The space of words is covered with a grid of fixed size, that is matrix $\Phi = \phi_{wt}$. 2. The number of cells containing probabilities above the threshold $p_{wt} > 1/W$ is calculated. 3. Value $\rho_{wt}$is calculated for the given number of topics T_t. 4. Steps 1, 2, 3 are repeated under variation of cell size, i.e., under variation of the number of topics. 5. Figure of function $\breve{\rho}(E)$ is plotted in bi-logarithmic coordinates. 6. Function slope is estimated by means of least square method, and this slope represents fractal dimension taken with the opposite sign: $D(E) = -\frac{\ln(p(E))}{\ln(\varepsilon)}$. Linear parts of function $\breve{\rho}(E$in bi-logarithmic coordinates characterize the process of self-reproduction of the density-of-states function in topic models.

3.1. Experiments on determining the fractal dimension in topic models

In the investigation of fractal properties of topic models, a set of computer experiments was conducted. In the calculations, the following datasets were used: 1. 'Lenta'. 2. '20 newsgroups' dataset. For both collections, a series of calculations were implemented, where the number of topics was changed in the range [2-50] with increments in one topic. All models were run three times, and the modeling results were averaged. For each averaged solution, value $\breve{\rho}(E)$ was calculated. The obtained curves were analyzed in bi-logarithmic coordinates. In experiments, the following topic models were used: 1. pLSA (E-M algorithm); 2. ARTM (E-M algorithm); 3. LDA Gibbs sampling. Examples of modeling and calculations of fractal dimensions are given in figures (38), (39), (40), (41).

Fractal analysis of the behavior of topic models demonstrates that text collections possess self-similar regions and a transition region between them. Moreover, this transition region between liner regions corresponds to Renyi entropy minimum [64]. Thus, the problem of analyzing the evolution of the topic model under variation of the number of topics can be reduced to the problem of locating the area that separates regions of self-similarity. The last problem is considered in the next chapter and is solved by means of the application of renormalization theory to topic modeling.

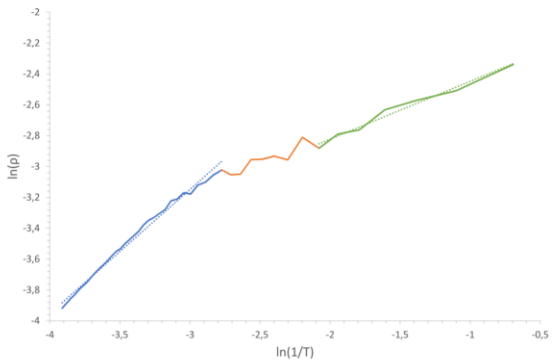


Fig. 38. Distribution of fractal dimensions (pLSA model)

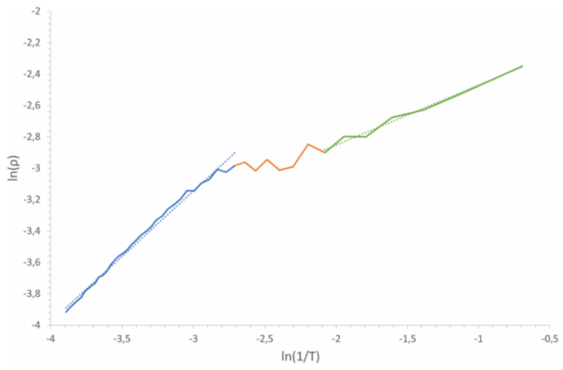


Fig. 39. Distribution of fractal dimensions (LDA GB model (α=0.4, β=0.5))

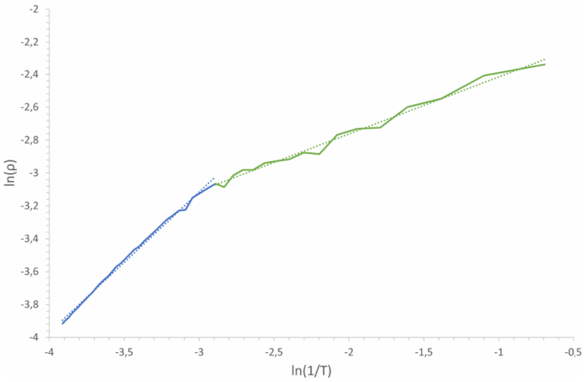


Fig. 40. Distribution of fractal dimensions (ARTM (sparse Φ = 0.01)).

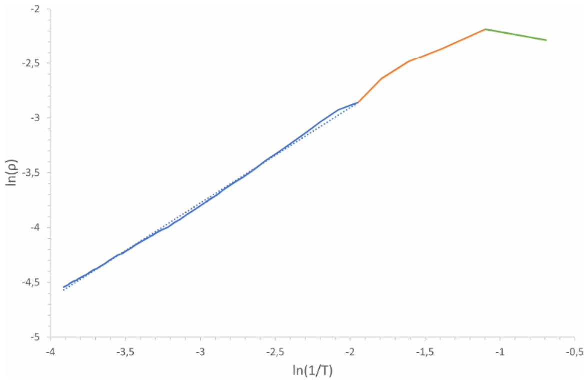


Fig. 41. Distribution of fractal dimensions (ARTM (sparse Φ = -10)

4. Aggregation method of topic models based on renormalization procedure.

4.1. Introduction into renormalization theory

Renormalization is a mathematical formalism that is widely used in different fields of physics such as percolation analysis and analysis of phase transitions. Renormalization consists of constructing a procedure for changing the scale of the system, in which the system's behavior remains the same. Theoretical foundations of the renormalization procedure were laid in the works of Kadanoff [33] and Wilson [34]. Renormalization procedures have been widely developed in the theory of fractals since fractal behavior has the property of self-similarity [35, 36].

The essence of the renormalization procedure is as follows. Let us consider a grid consisting of a set of nodes. We do not consider the physical features of these nodes and give only the formulation of the renormalization procedure. Each node is characterized by spin direction. In turn, a spin can take a particular direction, the number of which depends on the task. For example, in the Ising model, only two positions of spin are considered. In the Potts model, the number of positions can be 3-5 [37].

Nodes with the same spin constitute clusters. The scaling procedure or renormalization takes place according to the principle of block unification, in which several nearest nodes are replaced by one node. The direction of the new spin is taken as the direction of the majority of spins in the selected block. The procedure of block unification is carried out over the entire surface.

Accordingly, a new configuration of spins appears as a result. The scaling procedure can be implemented several times. Based on the principle that the new spin configuration must be equivalent to the old one, the possibility of building a procedure for estimating field parameters and critical exponent values appears. Let us note that subsequent application of the renormalization procedure or coarse-graining of the initial system gives approximate results. However, despite this fact, this method is widely used since it allows one to obtain critical exponent values in phase transitions, where standard mathematical models are not applicable. The renormalization procedure is successfully applied there, where scale invariance is observed. Scale invariance is characterized by power-law distributions. The mathematical expression of self-similarity is expressed as follows. Let $f(x)=cx^{\alpha}$, where c, α are constant. In the case of scale transformation in the form of $x \rightarrow \lambda x$ we obtain the same type of functional dependence but with another coefficient, i.e., $f(\lambda x)=\beta x^{\alpha}$. Thus, power-law distribution possesses scale invariance. The power parameter can be determined using different algorithms, for example, such as ‘box counting'.

4.2. General statement of the aggregation problem in the form of renormalization procedure in topic modeling

The general task of aggregation of topic models under variation of mixture distribution size is to apply renormalization technology. The definition of renormalization is borrowed from quantum field theory. It is an iterative renormalization method in which the transition from regions with lower energy to regions with higher energy is related to a change in the scale of the system. Renormalization is closely related to scale and conformal invariance and symmetries in which the system appears to be the same on all scales (so-called self-similarity). Renormalization of topic models is implemented as follows [65]. The result of TM is matrix $\Phi = \phi_{wt}$, which consists of a set of one-dimensional distributions of words by topics. Matrix size is determined by the number of words W and the number of topics T. In this work, a fixed vocabulary of unique words is considered. Therefore, scale change of topic model depends only on parameter $q = 1/T$. Renormalization procedure is procedure of merging topic pair into one topic. After merging two topics, a new topic is normalized since the sum of probabilities of all words in a topic always equals one. Since calculation of the element of matrix ϕ_{wt} depends on the model type, the mathematical formulation of the renormalization procedure is specific for each model. Moreover, the result of merging depends on the choice of topic pairs to be merged. In this work, three principles of merging topics are considered:

1) Principle of pairwise topic merging based on minimum Kullback-Leibler divergence. It assumes that topics with similar probability distributions should be merged. The calculation for topic pairs is as follows:

$$D_{KL}(p \mid q) = \sum_{i=1} p(x_i) \cdot ln\left(\frac{p(x_i)}{q(x_i)}\right) =$$
$$= -\sum_{i=1} p(x_i) \cdot ln\big(q(x_i)\big) + \sum_{i=1} p(x_i) \cdot ln\big(p(x_i)\big) \qquad (6).$$

Topics with the smallest Kullback-Leibler divergence are merged.

2) Principle of topic merging based on minimum Renyi entropy, calculated for each topic. The calculation is according to formula (4), but for summation, only probabilities of words in one topic are used. Further, two topics are merged if they have the smallest entropy values.

3) Merging randomly chosen topics.

Below we consider three renormalization procedures of topic models based on different algorithms of restoring hidden distributions. The first and the third model are based on E-M algorithm (VLDA, pLSA), and the second one is based on Gibbs sampling procedure (LDA GB).

4.3. Renormalization procedure for VLDA model based on E-M algorithm

In VLDA (variational Latent Dirichlet Allocation) [66], the model parameter is the number of topics. The results of the model implementation are T-dimensional vector вектор α, where each value α_i characterizes Dirichlet distribution for each topic, and matrix of distribution of words w by topics $t:\ \Phi = (\phi_{wt})_{w\in W, t\in T}$. Variational E-M algorithm is used for estimating values of matrix Φ and Newton-Rapson method is used for estimating values of vector α. Counter in this algorithm is calculated according to the following expression [65]:

$$\mu_{nt} = \phi_{w_n t} \exp\left(\psi\left(\alpha_t + \frac{L}{T}\right)\right), \qquad (7)$$

where L is the document length, n is the number of the current words, $w_n$is the word from the list of unique words, corresponding to the current term, ψ is digamma function, μ_{nt} is an auxiliary variable, playing role of a counter, and ϕ_{wt} is expressed through this auxiliary variable taking into account normalization during variational E-M algorithm.

For renormalization task, the sum of counters was used. An output of this algorithm is matrix $(\phi_{wt})_{w\in W, t\in T}$ and vector α. Renormalization algorithm consists of the following steps:

1) Selection of a pair of topics for merging according to one of the methods listed in 4.1. Let us denote the chosen topics by t_1 и t_2.
2) Merging of the selected topics. Values of the distribution of the 'new' topic $\phi_{\cdot t_1}$obtained in the result of merging t_1 и t_2 are calculated as follows [29]:
$$\phi_{wt_1} := \phi_{wt_1} \cdot \exp\left(\psi(\alpha_{t_1})\right) + \phi_{wt_2} \cdot \exp(\psi(\alpha_{t_2})). \qquad (8)$$
Further, new column $\phi_{\cdot t_1}$is normalized so that $\sum_{w\in W} \phi_{wt_1} = 1$. New value of $\alpha_{t_1} = \alpha_{t_1} + \alpha_{t_2}$corresponding to the 'new' topic is also recorded. After that, column $\phi_{\cdot t_2}$is removed from matrix Φ, and $\alpha_{t_2}$is removed from vector α. At this step, the number of topics is reduced by one, i.e., we obtain $T - 1$ topics. Then, the new values of vector α are normalized so that the sum of vector components equals 1.
3) Calculation of the overall Renyi entropy for the reduced number of topics. After that, when the new topic solution is formed, Renyi entropy is calculated for this new solution according to equation (4).

Steps 1, 2, 3 are iteratively repeated until there are only two topics left. Based on the results of renormalization, Renyi entropy curve is plotted as a function of the renormalization parameter, that is, of the number of topics. Then, Renyi entropy curve obtained in the result of renormalization is compared to the Renyi entropy curve obtained with the successive implementation of topic models under variation of the number of topics. Comparing two curves, one can estimate the effect of renormalization for this model. The region of Renyi entropy minimum corresponds to the region of the optimal number of topics.

4.4. Renormalization procedure for LDA model based on Gibbs sampling procedure

LDA model (Latent Dirichlet allocation) with Gibbs sampling is based on symmetric Dirichlet distributions, where the distribution of words in a topic is characterized by parameter β, and the distribution of topics in documents is characterized by parameter α. Matrix $\Phi = (\phi_{wt})$ is calculated by means of Gibbs sampling procedure. Values of β, α and the number of topics are set by a user. Calculation of matrix Φ consists of two stages. In the first stage, the sampling procedure is carried out, during which the counter $c_{wt}$is formed. In the second stage, elements of matrix ϕ_{wt} are calculated according to the following relation:

$$\phi_{wt} = \frac{c_{wt} + \beta}{(\sum_{w \in W} c_{wt}) + \beta W}, \qquad (9)$$

where c_{wt} equals the number of times word w was associated with topic t. For the task of renormalization of this model, we use the counter c_{wt} and relation (9). The algorithm's output is matrix $(\phi_{wt})_{w \in W, t \in T}$ and counters c_{wt}. The input of renormalization procedure is matrix c_{wt}, which undergoes the renormalization procedure, and based on this matrix, the final renormalization matrix ϕ_{wt} is calculated.

Renormalization algorithm for Gibbs sampling procedure consists of the following steps:

1) Selection of a pair of topics for merging according to one of the three described methods. Let us denote the selected topics by t_1 and t_2.
2) Merging of the selected topics. The new topic is obtained by summation of word frequencies of the two selected topics. Then, based on the new values of counters, the elements of matrix ϕ_{wt} are calculated. Renormalization equation is as follows [68]:

$$\phi_{wt_1} := \frac{c_{wt_1} + c_{wt_2} + \beta}{(\sum_{w \in W} c_{wt\_1} + c_{wt_2}) + \beta W}. \qquad (10)$$

New topic $\phi_{\cdot t_1}$already satisfies the property: $\sum_{w \in W} \phi_{wt_1} = 1$. Then, column $\phi_{\cdot t_2}$is removed from matrix Φ, i.e., the size of topic solution is reduced.

Steps 1 and 2 are iteratively repeated until there are only two topics left. Based on the results of renormalization, Renyi entropy curve is plotted as a function of the renormalization parameter, that is, of the number of topics.

4.5. Renormalization procedure for pLSA model

pLSA model is the simplest since it does not have regularizers and parameters. The only parameter is the number of topics [13]. The renormalization algorithm in this case consists of the following steps:

1) Selection of a pair of topics for merging according to one of the three methods. Let us denote the selected topics by t_1 и t_2.
2) Merging of the selected topics. In this model, the new topic is expressed through a simple summation of probabilities:

$$\phi_{wt_1} := \phi_{wt_1} + \phi_{wt_2} \qquad (11).$$

3) Normalization of the new topic. After summation, normalization of the new topic is carried out so that the sum of probabilities in the new topic equals 1. Then, column $\phi_{\cdot t_2}$is removed from matrix Φ.

Steps 1, 2, 3 are iteratively repeated until there are only two topics left. Based on the results of renormalization, Renyi entropy curve is plotted as a function of the renormalization parameter, that is, of the number of topics.

4.6. Experiments on renormalization

In the framework of investigation of renormalization, three datasets were used: 1. Russian-language dataset (Lenta.ru). 2. Dataset ’20 newsgroups’. 3. French-language dataset containing 25000 documents in French and 18749 unique words. The french-language dataset does not have a mark-up on topics. Topic modeling was implemented for these datasets with a variation of the number of topics 2-100 with increments in one topic. The following parameters were used for LDA model with Gibbs sampling: α=0.1, β=0.1. Investigation on the optimal hyperparameters for these datasets was conducted in work [15]; therefore, parameters were not varied in this work. Then, for each dataset, a topic solution on 100 topics underwent the renormalization procedure. Based on renormalization procedure, Renyi entropy curves as functions of the number of topics were plotted. Finally, Renyi entropy curves obtained with renormalization were compared to Renyi entropy curves that were obtained by the successive topic modeling (without renormalization).

4.6.1. Renormalization of LDA GB model (‘Lenta’ dataset)

The results of calculations for this model are presented in figures (43)-(45).

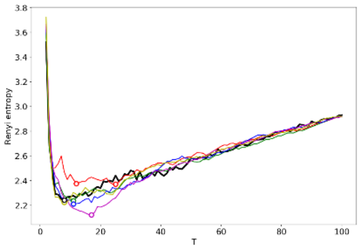


Fig. 43.

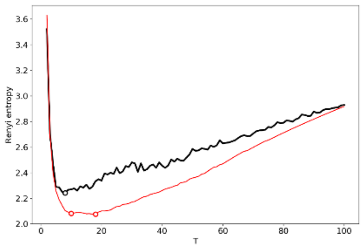


Fig. 44.

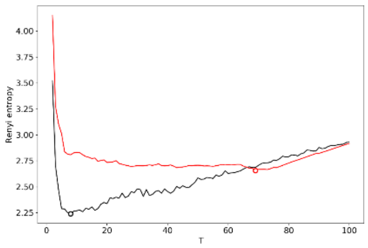


Fig. 45.

Renyi entropy curves as functions of the number of topics for the Russian-language dataset. Black: successive topic modeling. Fig. (43): merging random topics. Fig. (44): merging based on minimum Renyi entropy. Fig. (45): merging based on Kullback-Leibler divergence.

The last figure demonstrates that the renormalization of the topic model based on minimum Kullback-Leibler divergence produces the worst result among the three types of renormalization. The best result in terms of determining the optimal number of topics is demonstrated by the procedure of merging based on minimum Renyi entropy.

4.6.2. Renormalization of LDA GB model (’20 newsgroups’ dataset)

The results of calculations for this model are given in figures (46)-(48).

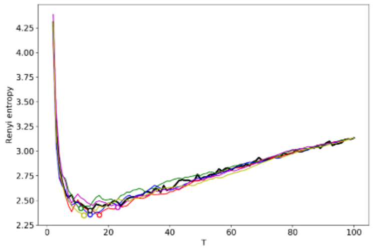


Fig. 46.

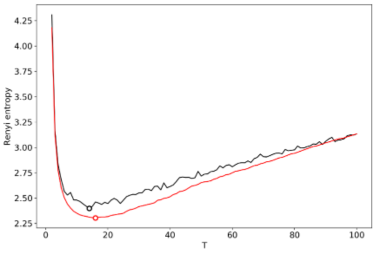


Fig. 47.

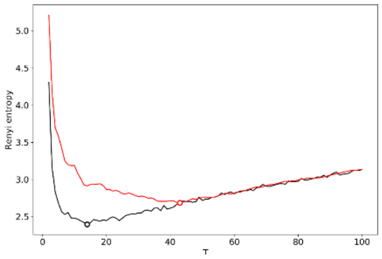


Fig. 48.

Renyi entropy curves as functions of the number of topics for the English-language dataset. Black: successive topic modeling. Fig. 46: merging random topics. Fig. 47: merging based on minimum Renyi entropy. Fig. 48: merging based on Kullback-Leibler divergence.

These calculations also demonstrate that the best result is demonstrated by the procedure of merging based on minimum Renyi entropy.

4.6.3. Renormalization of LDA GB model (French-language dataset)

The results of renormalization for the French-language dataset are given in the following three figures below.

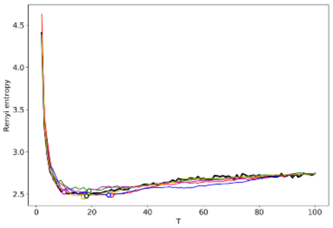


Fig. 49.

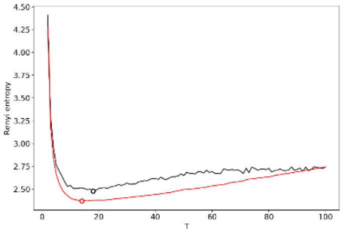


Fig. 50.

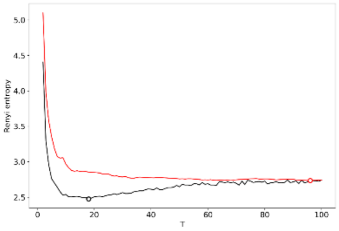


Fig. 51.

Renyi entropy curves as functions of the number of topics for the French-language dataset. Black: successive topic modeling. Fig. 49: merging random topics. Fig. 50: merging based on minimum Renyi entropy. Fig. 51: merging based on Kullback-Leibler divergence.

4.6.4. Renormalization of VLDA model ('Lenta' dataset)

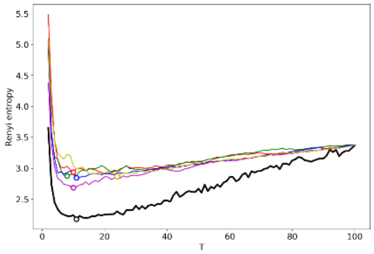


Fig. 52.

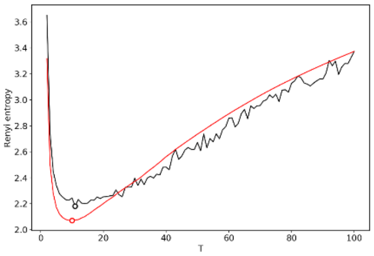


Fig. 53.

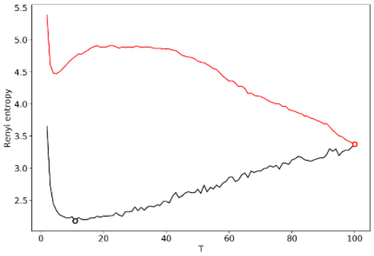


Fig. 54.

Renyi entropy curves as functions of the number of topics for the Russian-language dataset. Black: successive topic modeling. Fig. 52: merging random topics. Fig. 53: merging based on minimum Renyi entropy. Fig. 54: merging based on Kullback-Leibler divergence.

4.6.5. Renormalization of VLDA model ('20 newsgroups' dataset)

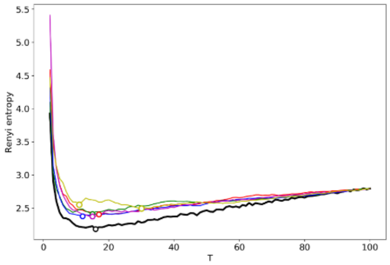


Fig. 55.

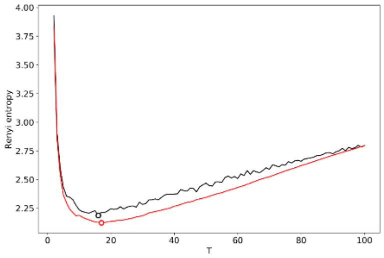


Fig. 56.

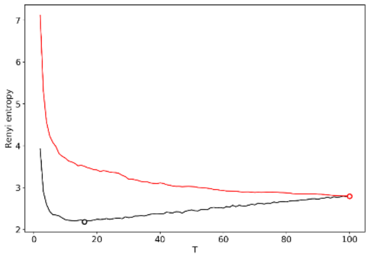


Fig. 57.

Renyi entropy curves as functions of the number of topics for the English-language dataset. Black: successive topic modeling. Fig. 55: merging random topics. Fig. 56: merging based on minimum Renyi entropy. Fig. 57: merging based on Kullback-Leibler divergence.

4.6.6. Renormalization of VLDA model (French-language dataset)

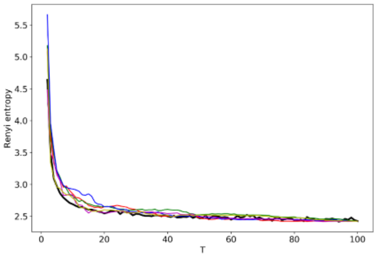


Fig. 58.

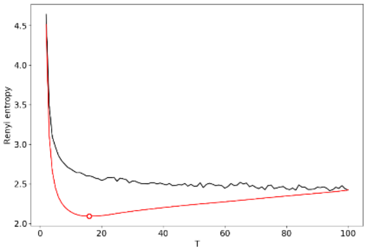


Fig. 59.

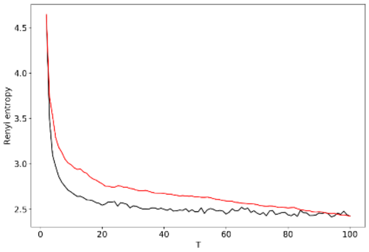


Fig. 60.

Renyi entropy curves as functions of the number of topics for the French-language dataset. Black: successive topic modeling. Fig. 58: merging random topics. Fig. 59: merging based on minimum Renyi entropy. Fig. 60: merging based on Kullback-Leibler divergence.

Renormalization of VLDA model also demonstrates that the best result in terms of determining the optimal number of topics is achieved by merging procedure based on minimum Renyi entropy, and the worst one corresponds to merging based on minimum Kullback-Leibler divergence. However, the renormalization model of VLDA performs worse than LDA GB model.

4.6.7. Renormalization of pLSA model (‘Lenta’ dataset)

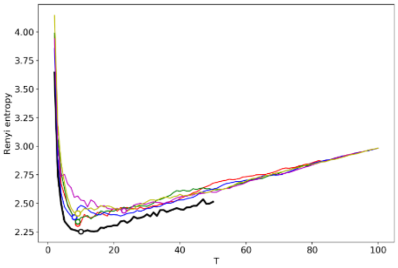


Fig. 61.

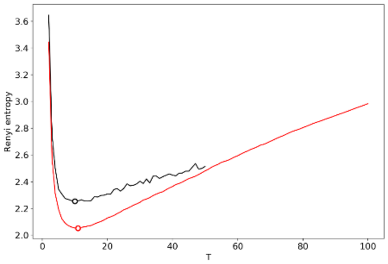


Fig. 62.

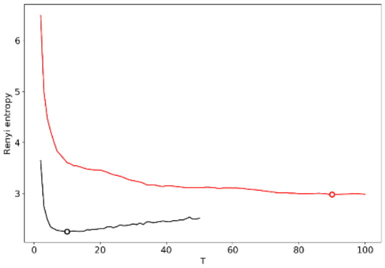


Fig. 63.

Renyi entropy curves as functions of the number of topics for the Russian-language dataset. Black: successive topic modeling. Fig. 61: merging random topics. Fig. 62: merging based on minimum Renyi entropy. Fig. 63: merging based on Kullback-Leibler divergence.

4.6.8. Renormalization of pLSA model (’20 newsgroups’ dataset)

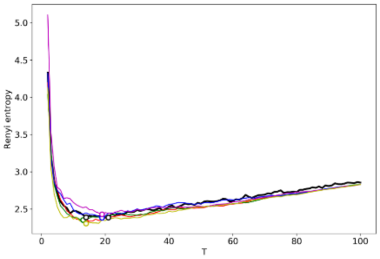


Fig. 64.

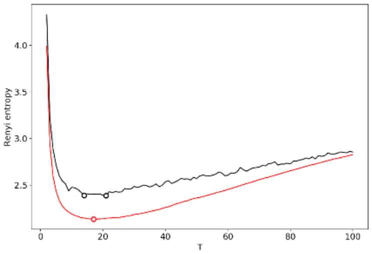


Fig. 65.

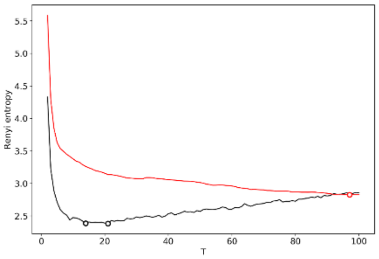


Fig. 66.

Renyi entropy curves as functions of the number of topics for the English-language dataset. Black: successive topic modeling. Fig. 64: merging random topics. Fig. 65: merging based in minimum Renyi entropy.Fig. 66: merging based on Kullback-Leibler divergence.

4.6.9. Renormalization of pLSA model (French-language dataset)

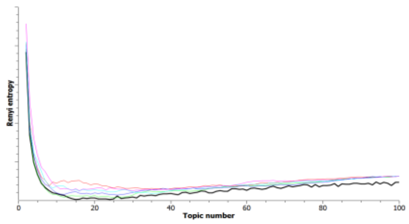


Fig. 67.

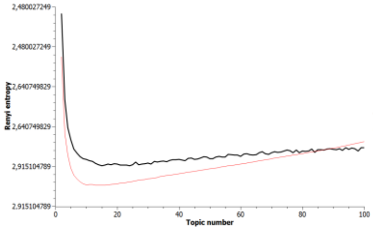


Fig. 68.

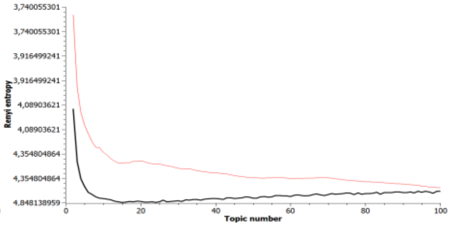


Fig. 69.

Renyi entropy curves as functions of the number of topics for the French-language dataset. Black: successive topic modeling. Fig. 67: merging random topics. Fig. 68: merging based on minimum Renyi entropy.Fig. 69: merging based in minimum Kullback-Leibler divergence.

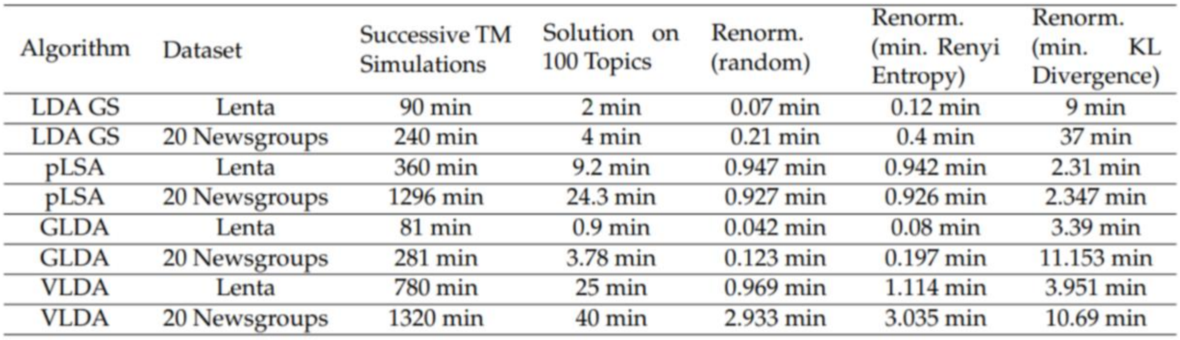
Thus, based on the conducted research, one can conclude the following: 1. The best performance is achieved by the renormalization procedure based on minimum Renyi entropy. 2. Renormalization based on minimum Kullback-Leibler divergence is not suitable for determining the optimal number of topics in text collections. 3. Renormalization is applicable to different European languages.

4.7. Comparison of aggregation methods according to the speed of models based on renormalization procedure

Table (4) demonstrates the time costs required for successive runs of topic modeling and the time costs of renormalization for marked-up datasets.

This table demonstrates that the renormalization procedure is performed 10-800 times faster than the successive calculation of topic models under variation of the number of topics. Calculations according to three renormalization procedures for datasets in different languages demonstrate that, first, the fastest procedures are renormalizations based on merging random topics and merging topics with minimal values of local Renyi entropy. Renormalization based on Kullback-Leibler divergence is the slowest. Moreover, this type of renormalization provides the worst result in terms of similarity of Renyi entropy curve obtained with successive topic modeling and Renyi entropy curve obtained with renormalization. The best result is achieved by renormalization based on merging topics according to local Renyi entropy. Merging random topics leads to significant fluctuations of Renyi entropy minimum. However, if renormalization curves are averaged for several runs, then averaged Renyi entropy curve also allows us to determine the optimal value of the number of topics. Thus, the most convenient in terms of speed and accuracy of estimation of the optimal number of distributions in the mixture of topics is the renormalization procedure based on minimum local Renyi entropy.

Table 4: Time costs of different models.



Algorithm	Dataset	Successive TM Simulations	Solution on 100 Topics	Renorm. (random)	Renorm. (min. Renyi Entropy)	Renorm. (min. KL Divergence)
LDA GS	Lenta	90 min	2 min	0.07 min	0.12 min	9 min
LDA GS	20 Newsgroups	240 min	4 min	0.21 min	0.4 min	37 min
pLSA	Lenta	360 min	9.2 min	0.947 min	0.942 min	2.31 min
pLSA	20 Newsgroups	1296 min	24.3 min	0.927 min	0.926 min	2.347 min
GLDA	Lenta	81 min	0.9 min	0.042 min	0.08 min	3.39 min
GLDA	20 Newsgroups	281 min	3.78 min	0.123 min	0.197 min	11.153 min
VLDA	Lenta	780 min	25 min	0.969 min	1.114 min	3.951 min
VLDA	20 Newsgroups	1320 min	40 min	2.933 min	3.035 min	10.69 min

5. Granulated variant of a topic model

As was mentioned, in the field of topics modeling, one of the main problems is the problem of stability. At the same time, the central part of scientific works is aimed at measuring stability using different quality measures. In this chapter, in contrast to other works, a new model that allows us to significantly improve the stability of topic models based on Gibbs sampling procedure is proposed. In this part of the work, we consider a model that is a modification of LDA model based on Gibbs sampling, where an explicit form of the local density function of the distribution of words by topics within a window of a given size is specified. Parzen–Rosenblatt window is set as [69]:

$$p(r) = \frac{1}{mh}\sum_{i=1}^{m} K(\frac{r-r_i}{h}), \quad (12),$$

where K(w) is an arbitrary even function called a kernel. The kernel K(w) must satisfy the normalization condition: $\int K(r)dr = 1$. In practice, the following kernels are frequently used: 1. Rectangular kernel. $K_r = const$ with a given window size h. 2. Epanechnikov kernel [70] $K_r = const \cdot (1 - r^2)$. 3. Triangular kernel $K_r = const \cdot (1 - |r|)$. These kernel functions were used in this work for the regularization of topic modeling.

5.1. Regularization of a topic model by specifying local density of distribution of words by topics

In this work, the regularization of a topic model is based on the idea of the existence of a topical dependence between a pair of unique words, i.e., on assumption about the existence of a local density of distribution of topics, which can be set by the kernel function. We assume that a topic consists of words that are not only described by Dirichlet distribution but also often co-occur together in a text. Specifying a type of distribution function of words by topics inside a window (local density) and window size, one can influence the character of model regularization.

In general, Gibbs sampling algorithm, taking into account the local density of the distribution of words by topic, is as follows:

```
• Initialization of matrices $\Phi = \phi_{wt}$ and $\Theta = \theta_{td}$.
• Outer loop on the number of iterations
    o Loop on documents
        ▪ Loop on words in a current document.
```

In the internal loop, random sampling according to Dirichlet distribution [21] is implemented. For a randomly selected anchor word (central word in a window), belonging to a topic is calculated, and topics of the other words inside the window are determined by local density function: $T(w_i) = T_0 \cdot K(w_0)$, where T_0 is the topic of the anchor word, obtained from Dirichlet distribution, $K(w_0)$ is the local density function, w_i are words inside the window.

  - ▪ End of the loop on words
- ○ End of the loop on documents
- End of the outer loop (on the number of iterations)

At the final stage of topic modeling, after the end of sampling, the final calculation of matrices ϕ_{wt}, θ_{td} of distributions of words and documents by topics is implemented based on the counters. Thus, specifying the type of local density function of distribution of words by topics and window size, we perform regularization of a topic model [72].

5.1.1. Rectangular kernel of regularization (granulated sampling, GLDA)

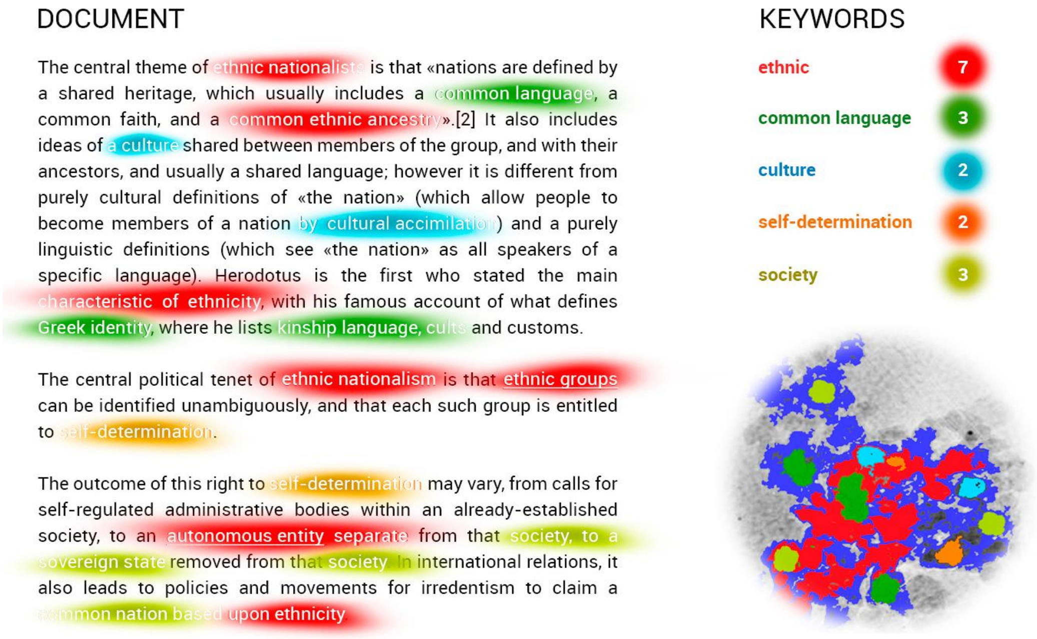
The stepwise function is considered the first kernel in this work. Its essence is that all words inside a given window have the same topic *K(T)=T(anchor word)*, i.e., the topic of the anchor word. The second regularization parameter is the window size. Thus, each document is considered a granulated surface consisting of granules (topics). An example of granulated text is given in figure (70). Since initially a combination of words that often occur inside one granule is not known, a granulated sampling variant forms a statistical dependence between closely located words. The results of the calculation of the stability of the topic model with a rectangular kernel are presented in table 5 (GLDA model).

5.1.2. Epanechnikov kermel (ELDA)

Epanechnikov kernel is a symmetric function that demonstrates that topics inside a given window are distributed as follows:

$$K(w) = T(anchor\ word) \cdot (1 - r^2), \quad (13),$$

where $r = 1$ corresponds to the rightmost word in the window, $r = -1$ corresponds to the leftmost word in the window. This means that the farther the word is from the anchor word, the more the topic of the word differs from the topic of the central word (the difference goes in the direction of decreasing the topic number). The results of the calculation of the stability of the topic model with Epanechnikov kernel are given in table 5 (ELDA model).



DOCUMENT

The central theme of ethnic nationalists is that «nations are defined by a shared heritage, which usually includes a common language, a common faith, and a common ethnic ancestry».[2] It also includes ideas of a culture shared between members of the group, and with their ancestors, and usually a shared language; however it is different from purely cultural definitions of «the nation» (which allow people to become members of a nation by cultural accimilation) and a purely linguistic definitions (which see «the nation» as all speakers of a specific language). Herodotus is the first who stated the main characteristic of ethnicity, with his famous account of what defines Greek identity, where he lists kinship language, cults and customs.

The central political tenet of ethnic nationalism is that ethnic groups can be identified unambiguously, and that each such group is entitled to self-determination.

The outcome of this right to self-determination may vary, from calls for self-regulated administrative bodies within an already-established society, to an autonomous entity separate from that society, to a sovereign state removed from that society. In international relations, it also leads to policies and movements for irredentism to claim a common nation based upon ethnicity.

KEYWORDS

ethnic	7
common language	3
culture	2
self-determination	2
society	3

Fig. 70. Example of a granulated surface in physics, and text presented in form of a granulated surface.

5.1.3. Triangular kernel (TLDA)

In this case, the local density function is set in form of a triangle:

$$K(w) = T(anchor\ word) \cdot (1 - |r|) \qquad 14.$$

This model is almost analogous to ELDA model. The difference is that decreasing the topic number for words at the edges of the window is faster than in ELDA model. The results on the stability of the topic model with the triangular kernel are presented in table 5 (LDA model).

5.2. Investigation of stability of ARTM, GLDA, pLSA, SLDA models

In general, in the framework of topic modeling, models based on LDA dominate. However, the addition of regularizers in LDA model is related to the complexity of Bayesian inference, which in turn, complicates the construction of multi-purpose topic models satisfying simultaneously a large number of regularizers. In work [16], the authors propose an alternative to the Bayesian approach, namely, additive regularization of topic models, ARTM. It has several fundamental differences from the Bayesian approach.

In this case, the construction of multi-purpose topic models is significantly simplified due to additive regularizers. The addition of a regularizer demands a slight modification of M-step in a ready E-M-like algorithm. In the framework of this work, the stability of ARTM model with two regularizers was investigated: 1. Sparsing regularizer of matrix Φ. 2. Sparsing regularizer of matrix Θ.

5.2.1. Experiments on stability of topic models

Nine topic models were investigated: 1. pLSA. 2. ARTM sparse Φ. 3. ARTM sparse Θ. 4. VLDA, 5. LDA GB. 6. Semi-supervised LDA GB. 7. GLDA. 8. ELDA 9. TLDA. For testing models, documents from the social network ‘’Live Journal’ were used. The total number of documents is 101481. In each modeling, 200 topics were used. The results of model calculations are presented in table (7). Each model was run three times under variation of the number of topics. The stability of topic models was calculated for three runs by means of Kullback-Leibler divergence.

Table 7.

Topic model	The number of stable topics	Average value of Jaccard distance
PLSA	54	0.47
PLSA+sparsing regularizer of matrix φ(w,t), regularization coefficient α=0.5	9	0.44
PLSA+ sparsing regularizer of matrix Θ(t,d), regularization coefficient β=0.2	87	0.47
Variational Latent Dirichlet Allocation (VLDA)	111	0.53
LDA (Gibbs sampling)	77	0.56
SLDA (Gibbs sampling)	84	0.62
GLDA (window size: ±1)	195	0.64
GLDA (window size: ±2)	195	0.71
GLDA (window size: ±3)	197	0.73
ELDA (window size: ±1)	184	0.23
ELDA (window size: ±2)	192	0.33
ELDA (window size: ±3)	199	0.20
TLDA (window size: ±1)	162	0.63
TLDA (window size: ±2)	200	0.3
TLDA (window size: ±3)	200	0.68

The calculations demonstrate that setting regularizers may both increase and decrease model stability. At the same time, adding information about local relations between words can significantly improve the stability of the topic model The proposed variant of the sampling procedure is significantly more stable than such models as pLSA, VLDA, ARTM.

Conclusion.

As it was noted, there are three significant problems in topic modeling: 1. Determining the number of components in a mixture of distributions, including determining the presence of flat and hierarchical structures in datasets. 2. Problem of tuning hyperparameters and regularization coefficients. 3. Problem of stability (reproducibility of topic solution). Correspondingly, in the framework of the research of this dissertation, ways to solve these problems have been proposed.

First, the entropic topic model (based on Renyi entropy) was implemented to determine the optimal number of distributions in the mixture for generative topic models. This model allows us to estimate the number of topics in datasets in European languages and to determine the optimal hyperparameters of topic models. Second, the hierarchical entropic topic model, which allows us to estimate the number of hierarchical levels and to determine a type of hierarchy in a dataset, was implemented. The number of Renyi entropy minima corresponds to the number of hierarchical levels in datasets. Third, an entropic topic model based on two-parametric Sharma-Mittal entropy was implemented. In this model, one entropy parameter is expressed in terms of the inverse number of topics, the second parameter is expressed through Jaccard distance, which allows us to take into account semantic similarity between topic solutions under variation of the size of the mixture of distributions. Fourth, the fractal model for estimating the performance of generative topic models was implemented. This model allows us to identify the linear regions of the word distribution density function and the transition regions corresponding to the minima of Renyi entropy. Fifth, the method of aggregation of topic models based on the renormalization procedure, which allows us to speed up searching the optimal size of distribution mixture for a dataset in hundreds of times, was implemented. In this part of the work, it was demonstrated that the most effective renormalization procedure in terms of searching the correct number of topics and in terms of calculation speed is renormalization based on merging topics with minimal values of Renyi entropy. Six, a granulated topic model based on Gibbs sampling procedure, where the regularization procedure of the topic model is set with a local density function of topic distribution, was implemented. The proposed model demonstrates a high level of stability in comparison to other topic models.

Although this work is complete, the proposed models can be used for further development of the field of machine learning as follows: 1. Renyi and Sharma-Mittal entropies are variants of the parameterized logarithmic function, where parameters significantly change its behavior. Based on the above, one can formulate a class of mathematical models in the field of machine learning on the basis of searching for the maximum of parameterized log-likelihood (one or two-parameterized variants of logarithm). 2. Renormalization procedure can be incorporated inside the existing algorithms of topic models, that in turn, may significantly speed up the performance of topic models. 3. Principle of searching for a minimum of parameterized entropy can be used for the optimization of the existing clustering algorithms (including the hierarchical clustering procedure). 4. The principle of searching for parameterized entropy minimum can be used to determine the optimal number of layers in neural networks. Preliminary experiments on networks of bounded and deep Boltzmann machines show that the behavior of Renyi entropy as a function of the number of layers in such networks is similar to the behavior of Renyi entropy in topic models as a function of the number of topics. 5. Granulated variant of the sampling procedure can be used for the development of topic models, where the granulated sampling procedure will take into account not the nearest words in a document, but the nearest words according to their word embeddings.

Funding

The research included in the dissertation, as well as the preparation of the dissertation, were supported by HSE grants in 2014-2021:

- 2014 Social and Political Processes Online: The Structure and Content of Social Interactions
- 2015 Interdisciplinary internet studies
- 2016 Internet use and Internet users: cross-country and cross-regional comparisons
- 2017 Internet as sociotechnical phenomenon
- 2018 Social and textual measurement of social network user profiles
- 2019 Social networks as a socio-psychological and textual phenomenon
- 2020 Online communication: cognitive limits and methods of automatic analysis
- 2021 Modeling individuals behavior and socio-psychological characteristics based on multimodal digital traces

References

1. Chauhan, Uttam and Apurva Shah. “Topic Modeling Using Latent Dirichlet allocation.” *ACM Computing Surveys (CSUR)* 54 (2022): 1 - 35.
2. Koltsov, Sergei, Vera Ignatenko, Maxim Terpilovskii and Paolo Rosso. “Analysis and tuning of hierarchical topic models based on Renyi entropy approach.” *PeerJ Computer Science* 7 (2021): n. pag.
3. Catherine, Ş A. and Ugar. “Finding the number of clusters in a data set : An information theoretic approach C.” (2003).
4. Stephens, Greg J., Thierry Mora, Gašper Tkačik and William Bialek. “Statistical thermodynamics of natural images.” *Physical review letters* 110 1 (2013): 018701 .
5. Mirkin, Boris G.. “Clustering for data mining - a data recovery approach.” *Computer science and data analysis series* (2005).
6. Tibshirani, Robert, Guenther Walther and Trevor J. Hastie. “Estimating the number of clusters in a data set via the gap statistic.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2000): n. pag.
7. Fujita, André, Daniel Y. Takahashi and Alexandre Galvão Patriota. “A non-parametric method to estimate the number of clusters.” *Comput. Stat. Data Anal.* 73 (2014): 27-39.
8. Aldana-Bobadilla, Edwin and Ángel Fernando Kuri Morales. “A Clustering Method Based on the Maximum Entropy Principle.” *Entropy* 17 (2015): 151-180.
9. Ramírez-Reyes, Abdiel, Alejandro Raúl Hernández-Montoya, Gerardo Herrera-Corral and Ismael Domínguez-Jiménez. “Determining the Entropic Index q of Tsallis Entropy in Images through Redundancy.” *Entropy* 18 (2016): 299.
10. Milligan, Glenn W. and Martha Cooper. “An examination of procedures for determining the number of clusters in a data set.” *Psychometrika* 50 (1985): 159-179.
11. SH Cha, Taxonomy of nominal type histogram distance measures, Proceedings of the American conference on applied mathematics, 325-330, 2008.
12. Rose, Gurewitz and Fox. “Statistical mechanics and phase transitions in clustering.” *Physical review letters* 65 8 (1990): 945-948 .
13. Hofmann, Thomas. “Probabilistic Latent Semantic Indexing.” *ACM SIGIR Forum* 51 (2017): 211 - 218.

14. Blei, David M., A. Ng and Michael I. Jordan. “Latent Dirichlet Allocation.” *J. Mach. Learn. Res.* 3 (2003): 993-1022.
15. Griffiths, Thomas L. and Mark Steyvers. “Finding scientific topics.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004): 5228 - 5235.
16. Vorontsov, Konstantin V., Anna Potapenko and Alexander Plavin. “Additive Regularization of Topic Models for Topic Selection and Sparse Factorization.” *SLDS* (2015).
17. Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. “Hierarchical Dirichlet Processes.” *Journal of the American Statistical Association* 101 (2006): 1566 - 1581.
18. Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. “Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes.” *NIPS* (2004).
19. Mimno, David, Wei Li and Andrew McCallum. “Mixtures of hierarchical topics with Pachinko allocation.” *ICML '07* (2007).
20. Belyy A. V., Seleznova M. S., Sholokhov A. K., Vorontsov K. V. Quality Evaluation and Improvement for Hierarchical Topic Modeling, Computational Linguistics and Intellectual Technologies. Dialogue 2018. pp. 110-123
21. Dieng, Adji B., Francisco J. R. Ruiz and David M. Blei. “Topic Modeling in Embedding Spaces.” *Transactions of the Association for Computational Linguistics* 8 (2020): 439-453.
22. Miao, Yishu, Edward Grefenstette and Phil Blunsom. “Discovering Discrete Latent Topics with Neural Variational Inference.” *ArXiv* abs/1706.00359 (2017): n. pag.
23. Daud, Ali, Juan-Zi Li, Lizhu Zhou and Faqir Muhammad. “Knowledge discovery through directed probabilistic topic models: a survey.” *Frontiers of Computer Science in China* 4 (2009): 280-301.
24. Asuncion, Arthur U., Max Welling, Padhraic Smyth and Yee Whye Teh. “On Smoothing and Inference for Topic Models.” *UAI* (2009).
25. Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang and Sheng Tang. “A density-based method for adaptive LDA model selection.” *Neurocomputing* 72 (2009): 1775-1781.
26. Arun, R., V. Suresh, C. E. Veni Madhavan and M. Narasimha Murty. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.” *PAKDD* (2010).
27. Roberts, Margaret E., Brandon M Stewart and Dustin Tingley. “Navigating the Local Modes of Big Data: The Case of Topic Models.” *Computational Social Science* (2016).
28. Koltsov, Sergei, Sergei I. Nikolenko, Olessia Koltsova and Svetlana Bodrunova. “Stable topic modeling for web science: granulated LDA.” *Proceedings of the 8th ACM Conference on Web Science* (2016): n. pag.
29. Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. “Evaluation methods for topic models.” *ICML '09* (2009).
30. Foulds, James R. and Padhraic Smyth. “Annealing Paths for the Evaluation of Topic Models.” *UAI* (2014).
31. Zhu, Jun, Amr Ahmed and Eric P. Xing. “MedLDA: maximum margin supervised topic models.” *J. Mach. Learn. Res.* 13 (2012): 2237-2278.
32. Sristy, Nagesh Bhattu and Durvasula V. L. N. Somayajulu. “Entropy Regularization for Topic Modelling.” *I-CARE 2014* (2014).
33. Kadanoff, Leo P.. “Statistical Physics: Statics, Dynamics and Renormalization.” (2000).

34. Wilson, Kenneth G.. “Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture.” *Physical Review B* 4 (1971): 3174-3183.
35. Olemskoi, A. Synergetics of Complex Systems: Phenomenology and Statistical Theory, KRASAND Publ. House, Moscow, 2009, 384 p. (in Russian).
36. Carpinteri, Alberto, Bernardino Chiaia and Giuseppe Andrea Ferro. “Size effects on nominal tensile strength of concrete structures: multifractality of material ligaments and dimensional transition from order to disorder.” *Materials and Structures* 28 (1995): 311-317.
37. Essam, John W.. “Potts models, percolation, and duality.” *Journal of Mathematical Physics* 20 (1979): 1769-1773.
38. Tikhonov, A.N. and Arsenin, V.Y. Solutions of Ill-Posed Problems. Winston, New York, (1977).
39. Belford, Mark, Brian Mac Namee and Derek Greene. “Stability of topic modeling via matrix factorization.” *Expert Syst. Appl.* 91 (2018): 159-169.
40. Greene, Derek, Derek O'Callaghan and Pádraig Cunningham. “How Many Topics? Stability Analysis for Topic Models.” *ECML/PKDD* (2014).
41. De Waal, A., Barnard, E.: Evaluating topic models with stability. In: 19th Annual Symposium of the Pattern Recognition Association of South Africa (2008).
42. Koltsov, Sergei, Sergei I. Nikolenko, Olessia Koltsova, Vladimir Filippov and Svetlana Bodrunova. “Stable Topic Modeling with Local Density Regularization.” *INSCI* (2016). *Lecture Notes in Computer Science series* Vol. 9934. Switzerland : Springer, (2016)
43. Derbanosov, R. Stability of topic modeling via modality regularization [Текст] / R. Derbanosov, M. Bakhanova // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. (2020)
44. Agrawal, Amritanshu, Wei Fu and Tim Menzies. “What is wrong with topic modeling? And how to fix it using search-based software engineering.” *Inf. Softw. Technol.* 98 (2018): 74-88.
45. Koltsov, Sergei. “A thermodynamic approach to selecting a number of clusters based on topic modeling.” *Technical Physics Letters* 43 (2017): 584-586.
46. Koltcov, Sergei. “Application of Rényi and Tsallis entropies to topic modeling optimization.” *Physica A: Statistical Mechanics and its Applications* (2018): n. pag.
47. Tsallis, Constantino. “Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World.” (2009).
48. Beck, Christian. “Generalised information and entropy measures in physics.” *Contemporary Physics* 50 (2009): 495 - 510.
49. Mora, Thierry and Aleksandra M. Walczak. “Rényi entropy, abundance distribution, and the equivalence of ensembles.” *Physical review. E* 93 5 (2016): 052418 .
50. Beck, Christian and Friedrich Schögl. “Thermodynamics of chaotic systems.” (1993).
51. Klimontovich, Yu. L. Statistical Theory of Open Systems (Yanus, Moscow, 1995; Springer, Dordrecht, 1995).
52. Sharma, Bhu Dev and Asha Garg. “Nonadditive Measures of Average Charge for Heterogeneous Questionnaires.” *Inf. Control.* 41 (1979): 232-242.
53. Nielsen, Frank and Richard Nock. “A closed-form expression for the Sharma–Mittal entropy of exponential families.” *Journal of Physics A: Mathematical and Theoretical* 45 (2011): n. pag.
54. Jaccard, P.. “The distribution of the flora in the alpine zone 1.” *New Phytologist* 11: 37-50.

55. Parker, Austin J., Kelly B. Yancey and Matthew P. Yancey. “Regular Language Distance and Entropy.” *MFCS* (2017).
56. Koltsov, Sergei, Vera Ignatenko and Olessia Koltsova. “Estimating Topic Modeling Performance with Sharma–Mittal Entropy.” *Entropy* 21 (2019): n. pag.
57. Koltsov, Sergei, Vera Ignatenko and Sergei Pashakhin. “How Many Clusters? An Entropic Approach to Hierarchical Cluster Analysis.” *SAI* (2020).
58. News Dataset from Usenet. Available online: http://qwone.com/~jason/20Newsgroups/ (accessed on 31 October 2019).
59. Basu, Sugato, Ian Davidson and Kiri L. Wagstaff. “Constrained Clustering: Advances in Algorithms, Theory, and Applications.” (2008).
60. Lesche, Bernhard. “Instabilities of Rényi entropies.” *Journal of Statistical Physics* 27 (1982): 419-422.
61. Blei, David M., Thomas L. Griffiths, Michael I. Jordan and Joshua B. Tenenbaum. “Hierarchical Topic Models and the Nested Chinese Restaurant Process.” *NIPS* (2003).
62. Mimno, David, Wei Li and Andrew McCallum. “Mixtures of hierarchical topics with Pachinko allocation.” *ICML '07* (2007).
63. Chirkova, Nadezhda. “Additive Regularization for Hierarchical Multimodal Topic Modeling.”. Machine Learning and Data Analysis, 2:187–200. (2016)
64. Ignatenko, Vera, Sergei Koltsov, Steffen Staab and Zeyd Boukhers. “Fractal approach for determining the optimal number of topics in the field of topic modeling.” *Journal of Physics: Conference Series* (2019): n. pag.
65. Schroeder, Manfred. “Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise.” (1991).
66. Koltcov, Sergei and Vera Ignatenko. “Renormalization Analysis of Topic Models.” *Entropy* 22 (2020): n. pag.
67. Mimno, David, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders and Andrew McCallum. “Optimizing Semantic Coherence in Topic Models.” *EMNLP* (2011).
68. Koltsov, Sergei and Vera Ignatenko. 2020. "Renormalization Analysis of Topic Models" *Entropy* 22, no. 5: 556. https://doi.org/10.3390/e22050556.
69. Rosenblatt, Murray. “Remarks on Some Nonparametric Estimates of a Density Function.” *Annals of Mathematical Statistics* 27 (1956): 832-837.
70. Epanechnikov, V. A. Nonparametric estimation of multidimensional probability density. Theory Probab. Appl. 14, 153–158, 1973.
71. Koltsov, Sergei, Sergei I. Nikolenko and E. Y. Koltsova. “Gibbs sampler optimization for analysis of a granulated medium.” *Technical Physics Letters* 42 (2016): 837-839.
72. Newman, David, Edwin V. Bonilla and Wray L. Buntine. “Improving Topic Coherence with Regularized Topic Models.” *NIPS* (2011).
73. Andrzejewski, David and Xiaojin Zhu. “Latent Dirichlet Allocation with Topic-in-Set Knowledge.” *HLT-NAACL 2009* (2009).
74. Tsallis, Constantino and Daniel A. Stariolo. “Generalized simulated annealing.” *Physica A-statistical Mechanics and Its Applications* 233 (1996): 395-406.