Physica A 512 (2018) 1192–1204

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Application of Rényi and Tsallis entropies to topic modeling optimization

Sergei Koltcov

Laboratory for Internet Studies (LINIS), National Research University Higher School of Economics, ul. Soyuza Pechatnikov, d. 16, 190008 St. Petersburg, Russia

HIGHLIGHTS

- Rényi and Tsallis entropies are found to detect the optimal number of topics.
- Entropy approach reveals a meaningful difference in algorithm performance.
- The content of all topics is quasi-periodically related to the number of topics.

ARTICLE INFO

Article history:
Received 16 March 2018
Available online xxxx

Keywords:
Topic modeling
Renyi
Entropy
Free energy
Complex system

ABSTRACT

This study proposes to minimize Rényi and Tsallis entropies for finding the optimal number of topics T in topic modeling (TM). A promising tool to obtain knowledge about large text collections, TM is a method whose properties are underresearched; in particular, parameter optimization in such models has been hindered by the use of monotonous quality functions with no clear thresholds. In this research, topic models obtained from large text collections are viewed as nonequilibrium complex systems where the number of topics is regarded as an equivalent of temperature. This allows calculating free energy of such systems—a value through which both Rényi and Tsallis entropies are easily expressed. Numerical experiments with four TM algorithms and two text collections show that both entropies as functions of the number of topics yield clear minima in the middle area of the range of T. On the marked-up dataset the minima of three algorithms correspond to the value of T detected by humans. It is concluded that Tsallis and especially Rényi entropy can be used for T optimization instead of Shannon entropy that decreases even when T becomes obviously excessive. Additionally, some algorithms are found to be better suited for revealing local entropy minima. Finally, we test whether the overall content of all topics taken together is resistant to the change of T and find out that this dependence has a quasi-periodic structure which demands further research.

1. Introduction

Statistical physics is increasingly being used to describe objects and processes that go beyond physical phenomena. Thus, large arrays of textual data, which have been rapidly accumulating on the Internet in the last decade, require ever more complex methods for their automatic processing and modeling. A wide range of mathematical tools, including topic modeling, is used for this [1], but their properties and behavior remain underresearched. This makes parameter optimization for such models a difficult task. However, if the results of topic modeling are considered equivalent to nonequilibrium complex systems (since the former, as it will be shown below, possess some properties of such systems), this would make it possible to apply a whole range of approaches from statistical physics. First of all, these are models for analyzing the

E-mail address: skoltsov@hse.ru.

https://doi.org/10.1016/j.physa.2018.08.050
0378-4371/

processes of self-organization of large ensembles. The basis for such an analysis may be an approach in which behavior of a topic model of a textual collection as an ensemble would be determined by thermodynamic functions, such as entropy or free energy. It is known that complex systems can be characterized by exponential and power law distributions, which is especially true for social [2,3], biological [4,5] and economic systems [6,7]. However, for topic models of textual collections, where the units are documents, words and latent semantic variables (topics), Pareto-like distributions are more typical [8,9]. Proceeding from this, when applying the maximum entropy principle for such systems, we propose to use an approach based on deformed statistic with the underlying Rényi or Tsallis entropies [10,11]. In this case, the deformed statistic of complex systems, like its non-deformed equivalent in other cases, will describe the probabilistic features that characterize the topic model of a textual collection as a system that has a large number of "particles" and that can remain in thermodynamically equilibrium and nonequilibrium states. If the deformation parameter q is accounted for while modeling thermodynamically atypical systems with long-range interactions, we expect that behavior of such systems could be explained much better than with any standard statistic. Moreover, the search for optimal parameters describing the state of these systems can be achieved on the basis of an entropy maximization procedure [12].

Our attention in this work is focused on topic modeling [1], since it is the most effective and sometimes the only available method of obtaining knowledge about the topic structure of large textual collections of which nothing is known in advance. This task is often encountered in the studies of Internet content, including news, consumer reviews, and social network messages. At the same time, topic modeling (TM) as a mathematical approach is applicable not only to textual data [13], but also to mass spectra [14], images [15], and other objects. In essence, topic modeling is an expanded version of cluster analysis that allows simultaneous estimation of distributions of both words and documents over topics/clusters. Moreover, topic models also provide the opportunity to rank words and documents according to the probability within each topic/cluster, which is not typical for traditional cluster analysis. The major problem of this group of methods is the lack of ground truth, that is, of knowledge about the correct number and composition of clusters. This hinders investigation of the properties of these models and makes us seek solutions based on theories from other areas of science.

Thus, in this paper we use the concept of deformed entropy and a range of thermodynamic concepts to investigate behavior of topic models under conditions of the changing number of topics. The purpose of such a study is to find the optimal number of topics/clusters, first, based on the maximum information approach, and second, on the basis of the T-invariance principle introduced further in the work.

The rest of the paper proceeds as follows. Section 2 first briefly explains the logic of topic modeling, which is necessary to further describe the proposed solutions. Next, this section provides an overview of the available approaches for determining the optimal number of clusters in cluster analysis and topics in topic modeling, and their limitations are indicated. In Section 3, we propose our entropy approach to the analysis of topic models as complex nonequilibrium systems, an approach that allows finding the optimal number of topics. Sections 4 and 5 describe the data used and the results of numerical experiments performed to verify our approach. Section 4 shows that the minimum q-deformed entropy is reached with the 'correct' number of topics taken from the marked up textual data, and therefore can be used as a criterion for selecting the right number of topics. Section 5 shows experimentally that the overall lexical composition of all the topics taken together is nearly invariant, that is, it is resistant to changing the number of topics in the greater part of the range of variation, but this invariance is intermittent and is described by a quasiperiodic function. We conclude that the T-invariance parameter must be accounted for while choosing the number of topics and that it must be included in the general theory of parameter optimization for topic models in the future research.

2. Problems of topic modeling and cluster analysis

2.1. Introduction to topic modeling

Topic modeling as a version of cluster analysis is based on the following propositions [16]:

1. Let D be a collection of text documents, and W—a set (dictionary) of all unique words. Each document $d \in D$ is a set of terms $w_1, \ldots, w_n$ from dictionary W.
2. It is assumed that there is a finite number of topics T, and every entry of word w in document d is associated with some topic $t \in T$. A topic is taken to mean a set of words that are often found together in a large number of documents.
3. A collection of documents is considered a random and independent selection of triads (w_i, d_i, t_i), $i = 1, \ldots, n$ from the discrete distribution p(w,d,t) over the finite probabilistic space $W \times D \times T$. Words w and documents d are observable variables, topic $t \in T$ is a latent (hidden) variable.
4. It is assumed that the order of terms in documents is not important for identifying topics (the 'bag of words' approach), and neither is the order of documents in the collection.

In TM, it is also assumed that probability p(w|d) of the occurrence of terms w in documents d can be expressed as a product of distributions p(w|t) and p(t|d). According to the formula of total probability and the hypothesis of conditional independence, we have the following expression [17]:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td} \tag{1}$$

where p(w|t) is the distribution of words over topics, and p(t|d) is the distribution of documents over topics. Thus, to construct a topic model of a dataset means to solve the inverse problem of finding the set of latent topics T based on some observed data. This includes restoring the set of one-dimensional conditional distributions $p(w|t) \equiv \varphi(w,t)$ for each topic t (term-topic distributions that constitute matrix φ_{wt}) and the set of one-dimensional distributions $p(t|d) \equiv \theta(t,d)$ for each document d (document-topic distributions constituting matrix θ_{td}) based on the observed variables d and w.

In TM two approaches to inference of such distributions are being actively developed: 1. TM using maximum likelihood estimation. For the given approach, the most known models are probabilistic latent semantic analysis (pLSA) [17] and variational latent Dirichlet allocation (VLDA) [13], in which matrices p(w|t) and p(t|d) are found by the EM-algorithm. 2. TM based on Markov chain theory, or the model with Gibbs sampling (LDA/GS) [18], wherein p(w|t) and p(t|d) are computed as expected values with the Monte Carlo method. A brief description of the differences between these models is given in the supplementary material.

2.2. Nonequilibrium character and instability of topic modeling

The process of topic modeling can be regarded as the transition of the entire system to a nonequilibrium state. In LDA/GS, the initial distribution of words and documents in matrices φ_{wt} and θ_{td} is set to be flat, while in pLSA and VLDA it is obtained by means of a random number generator. For both types of algorithms, the initial distribution corresponds to the maximum entropy. However, regardless of the type of algorithm and the initialization procedure, word and document probabilities in the topic models are redistributed over topics during modeling in such a way that a considerable proportion of word probabilities (about 95% of all unique words) is close to zero, and only about 3–5% have relatively high values [19,20].

The similarity of solutions obtained in the course of topic modeling to the nonequilibrium state of physical systems makes it possible to enrich TM with concepts from statistical physics; however, this similarity is not absolute. Physical systems are characterized by the indistinguishability of particles: that is, if the topic solution was a physical system, it would not matter which particles (words) populated the states with high probability values. However, for textual content researchers, the specific composition of the most probable words of each topic and all the topics together is precisely the main informative result. This circumstance has two consequences.

First: the composition of each topic is important, but the nondeterministic nature of the TM algorithms leads to the fact that the same algorithm run on the same data with the same parameters yields slightly different topics. This does not allow the text researcher to answer the main question: what are the topics contained in the given collection? The problem of such TM instability is very little studied; one of the few solutions is the previously proposed extension for the LDA/GS algorithm–Granulated LDA (GLDA) [20,21], which demonstrates almost 100% stability (for more details, see the supplementary material). Although GLDA has a number of serious shortcomings, such as weak interpretability and high topic correlation, we use this algorithm among others in our experiments in this study to test the applicability of the proposed approach to both low-deterministic and high-deterministic LDA algorithms.

Second: it is also important which words turn out to be most probable in all topics as a whole when changing the number of topics. If the topic solutions for different numbers of topics T give radically different compositions of top words, the algorithm as a whole is not suitable for use. If, however, most values of T produce similar top word compositions, and only a few outlying ranges of T yield diverging word compositions, then cutting off such abnormal ranges, together with searching for the minimum entropy, can become a valuable tool for selecting the optimal number of topics. This problem has not yet been investigated at all, and in this paper we are just approaching the first results of studying it. Stability of the compositions of top words across models with the different number of topics will be further called T-invariance, where T is the number of topics/clusters.

2.3. Approaches to estimation of T

The main problem in finding the optimal number of clusters in cluster analysis and topics in topic modeling is the choice of a function for optimization. In cluster analysis, minimization of intracluster distance and maximization of intercluster distance are most often used. The problem with these approaches is, however, that the increase in the number of clusters leads to quite a smooth monotonic dependence of this kind of functions on the number of clusters. Consequently, various transformation procedures are used to find extrema for such functions [22]. For this task, several solutions are proposed in the cluster analysis [23–25]. More clustering quality measures are discussed in works [26,27]. There are also models for determining the number of clusters based on the entropy maximization principle [28,29], but they use the Gibbs–Shannon entropy only.

For our purposes, the most interesting approach employed in cluster analysis is the one based on the ideas of statistical physics, namely, on free energy minimization [30]. Its main idea is as follows: each element of a system is characterized by probabilities of belonging to different clusters. Accordingly, for each element, we can formulate the concept of internal energy and thus calculate the free energy of the entire system. The temperature in such a system is considered a free parameter, which is varied in order to minimize the free energy. Such a thermodynamic approach is successfully used in the theory of dynamical systems [31] and in the analysis of images [32] and neural networks [33]. Further development of this approach occurs in the framework of nonextensive statistics. The discussion of the application of q-deformed statistic for machine learning is presented in [34,35], and the generalized version of the ‘rate distortion theory’ is discussed in [36]. The possibility

of applying deformed statistic for image segmentation is discussed in [37]. However, in none of these studies is q-deformed statistic used to determine the number of topics in topic modeling, although the problem of finding the optimal number of topics is just as relevant for it and even more complex.

This happens due to the following reasons. First, in TM it is difficult to formulate both a semantic concept of a topic and linguistic criteria for differentiating between topics and, consequently, to develop quality measures for topics and topic solutions. Second, in TM, as well as in cluster analysis, a hard task is a justifiable choice of a function that would link the quality of a topic model to the number of topics in a way that would allow optimization of the latter. Nevertheless, there are several works in which the authors have attempted to solve the problem of choosing the number of topics specifically in TM. The authors of [38] develop the ideas borrowed from cluster analysis and show that topic solutions with minimal correlation between the topics (as determined by the cosine similarity measure) correspond to solutions with a minimum value of another quality measure, perplexity. This is an interesting work, but we have never come across data on which the function of the perplexity dependence on the number of topics would have a minimum (as it has in [38]), instead of monotonically decreasing. This might indicate that the authors of [38] have worked with very specific datasets. In [39] it is proposed to perform the singular-value decomposition (SVD) of matrices φ and θ, then to select two vectors containing singular quantities and finally to calculate the distance between them. This distance estimated with the Kullback–Leibler divergence is what the authors suggest to minimize. According to them, the optimal number of topics corresponds to the situation where both matrices are described by the same number of singular values. Unfortunately, this approach is not verified with any alternative measures of TM quality, and in addition the operation of SVD and calculation of the Kullback–Leibler divergence severely restrict the application of this approach to big data. The collections used in [39] do not exceed 2500 texts.

One of the most well-known approaches to the problem of determining the number of topics in TM is the Hierarchical Dirichlet Process (HDP) model [40,41]. It allows constructing a hierarchy of topics in the form of a tree while initially assuming the existence of an infinite number of topics. The choice between tree branches and the selection of the number of levels in the tree are determined by the user, as well as by the features of the task and the dataset. However, the algorithm contains some built-in parameters that limit the structure of the tree and, therefore, affect the total number of topics. Those are the concentration parameter γ which significantly affects the size of the tree [40], and the predefined constant determining the number of topics that describe each document [41]. These parameters must be set by the user on bases that are not completely clear, and their modification can lead to a change in the number of topics.

Finally, among the approaches to determining the number of topics, it is worth mentioning the principle of calculating the nonequilibrium free energy formulated in [42]. However, [42] tests this principle only for one TM algorithm with Gibbs sampling, and only on one dataset. Unlike the present work, [42] does not investigate the relationship between the number of topics and the composition of the most probable words (T-invariance).

In this paper, we propose an extension of the thermodynamic approach for the analysis of all types of topic models by using Rényi and Tsallis entropies as the main measures of TM quality. Additionally, we investigate T-invariance of TM measuring it with the Jaccard index.

3. Application of the entropy approach to the analysis of complex textual systems

Number of topics can be considered a parameter characterizing the algorithm's 'resolution'. In this context, we suggest to go beyond comparing topic solutions based on pairwise comparisons of separate topics or eigenvectors, as it was previously done. Instead, we propose to consider a set of topics as a statistical ensemble of highly probable words. This gives us a possibility to investigate the behavior of the distribution of this ensemble while modifying the extensive parameter T (number of topics). In accordance with the maximum entropy principle [12], we consider entropy to be negative information; thus, the maximum of entropy corresponds to the minimum of information. We assume that the 'true number of topics' (the best resolution of the algorithm) corresponds to the maximum of the information received (or to the minimum of nonextensive entropy of the topic model).

Proceeding from this, the collection of documents can be considered a mesoscopic information system consisting of millions of elements (words and documents) with an initially unknown number of topics. If we regard the change in the number of topics set by the researcher as a process in which the system exchanges information with the environment, then such a system will be an 'information thermostat' [11]. The latter, by definition, differs from a physical thermostat by being an open system. Accordingly, with a change in the number of topics, the information system may not reach an equilibrium state in the sense of the Gibbs–Shannon entropy maximum, but it may stabilize in an intermediate equilibrium state, which is determined by the local minimum of Rényi or Tsallis entropy.

The totality of words that are statistically frequently found together in a large number of documents forms what can be called a topic. If a similar topic is fairly consistently reproduced from solution to solution on the same dataset, then such a topic can be considered a dissipative structure in the sense of Prigogine [43]. A collection of documents can contain only a finite number of such structures. Therefore, the cumulative set of words with a probability above a certain threshold (which gives these words a capacity of characterizing all dissipative structures as a whole) presumably should be constant. It is these stable dissipative structures that should be revealed through topic modeling.

As was partly mentioned in Section 2.2, the share of words with high probability is extremely small, and the value of probabilities in top words differs sharply from that of all other words. That is, the distribution of words and documents in

such information systems is extremely non-homogeneous, which means that calculation of entropy in such systems requires accounting for their nonequilibrium character.

Based on this, our approach can be formulated as follows [42]: **1.** In the information thermodynamic system under consideration, the total number of words and documents is a constant, that is, volume change is absent. **2.** A topic is a state (equivalent to the direction of a spin) that each word and document in the collection can take. Both words and documents can simultaneously belong to different topics with different probabilities. **3.** The information thermodynamic system is open and exchanges energy with the external environment by changing the 'temperature', which is understood as the number of topics (or clusters) T. This value is set from the outside and is a parameter that must be determined by searching for the minimum nonextensive entropy of the system. To measure the degree to which a given system is nonequilibrium we propose to use the entropy difference $\Lambda_s = S - S_0$ (also known as Lyapunov function or relative entropy) [44], where S_0 is the entropy of the zero state (chaos), and S is the entropy of the nonequilibrium state. Similarly to the Lyapunov entropy function, we can construct a function based on the difference of free energies: $\Lambda_F = F(T) - F_0$, where F_0 is the free energy of the initial state (chaos), and F(T) the free energy at a given value of T [44]. **5.** Since the topic modeling algorithm is a procedure of restoring latent distributions from a collection, the number of distributions is a variable parameter. The optimal number of such distributions corresponds to the situation where an information maximum (and, consequently, an entropy minimum) is reached. **6**. In this paper, it is also assumed that the equilibrium state of an information system can be characterized by the fact that the overall set of words with high probabilities ceases to change with the change in the number of topics. This means that the difference between two topic solutions, calculated using Jaccard index [45], is a constant at a certain interval of parameter T.

3.1. Density-of-states function

The total number of microstates in the information system that words in the topics can take, is W·T, where W is the number of unique words in the collection, and T is the number of topics/clusters. Let us define the density-of-states function as follows: $\rho(E) = \frac{\sum_{wt}^{WT} N(\varepsilon)_{wt}}{W \cdot T}$, where $N(\varepsilon)_{wt}$ is the number of states with energy E for all highly probable words w across all topics t (w and t are the indices of summation); W·T is the total number of states of all words. Proceeding from this, the relative Shannon entropy can be expressed in terms of the density of states as follows [33]:

$$S(E) = \ln(\rho(\varepsilon)) \tag{2}$$

It should be noted that the sum of probabilities for all microstates is always equal to one:

$$1 = \frac{1}{WT} \sum_{wt}^{WT} \rho(\varepsilon)_{wt} \tag{3}$$

It should also be noted that relative Shannon entropy is a subtype of gap statistic [24].

3.2. Energy of microstate and statistical sum of the information system

The energy of a microstate can be expressed as: $\varepsilon_{wt} = -\ln(p_{wt})$, where w is a word's number in the list of unique words, t is the topic number, and p_{wt} is the probability of the word w in topic t. In general, the energy range can be divided into a given number of intervals k, therefore, the density-of-states function can be expressed as:

$$\rho(E) = \frac{\sum_{wt}^{WT} N(\varepsilon)_{nt}}{WT} = \frac{\sum_{k}^{K} N_k}{WT}, \tag{4}$$

where N_k is the number of microstates with energy ε_k falling within interval k. The partition function, then, can be written in the following form: $Z = \sum_k^K e^{-\varepsilon_k/T}$.

3.3. Nonequilibrium free energy of the topic model

As already noted, in the course of topic modeling the transition to a strongly nonequilibrium state occurs, which is characterized by the fact that one part of the states has high probabilities $P_{nt} > 1/W$, and the other demonstrates low probabilities $P_{nt} < 1/W$, close to zero. From here on, we will consider only the states in which the information system resides with a non-zero probability. The entropy of a nonequilibrium system is described by the quantity $\Lambda_s = S - S_0$. Accordingly, the entropy and energy of the system are functions of the number of topics. Proceeding from the above, we can express the nonequilibrium free energy of the topic model in the following form:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\frac{\sum_{t=1}^{T} \sum_{w=1}^{W} P_{wt}}{T}\right) - T \cdot \ln\left(\frac{N_{k1}}{W \cdot T}\right), \tag{5}$$

where N_{k1} is the number of states in which $P_{wt} > 1/W$, (W·T) is the total number of all states, T is the number of topics (a variable parameter), W is the size of the dictionary of unique words, and E_0 and S_0 are the energy and the entropy of the system for the initial distribution that correspond to the maximum entropy. Quantities N_{k1} and P_{wt} are calculated for each topic model with parameter T being varied, so the quantity Λ_F is a function of T.

3.4. Information measure

As mentioned above, since a measure of information can be represented as entropy taken with the reversed sign, that is, the maximum entropy corresponds to the minimum information [12], the search for the optimal number of topics in complex systems can be reduced to the search for the minimum entropy. The classical version of entropy is the Gibbs–Shannon entropy (Shannon entropy) [46]: $S = -\sum_i p_i \cdot \ln(p_i)$, which in the case of uniform distribution coincides with the Boltzmann entropy. Here, we also consider two main q-deformed entropies, Rényi and Tsallis, since they are suitable for analyzing the behavior of a nonequilibrium information system, and they are the ones that we propose to minimize in order to find the optimal number of topics. The free energy of nonequilibrium information system Λ_{F} is, on the one hand, $\Lambda_{\mathrm{F}} = \mathrm{F} = \mathrm{E} - \mathrm{TS}$, and, on the other hand, $\mathrm{F} = \ln(\mathrm{Z})/\mathrm{T}$. With this in mind, the partition function $Z = \sum_k^K e^{-\varepsilon_{k/T}}$ with $\rho_k = e^{-\varepsilon_{k/T}}$, then the Rényi entropy, within the thermodynamic formalism [30,31] $S^R_{q=1/T} = \frac{\ln(\sum_k p_k^q)}{q-1}$, can be expressed in terms of free energy through the use of escort distribution: [46]:

$$S^R_{q=1/T} = \frac{F}{T-1}, \mathrm{q} = 1/\mathrm{T} \tag{6}$$

In this case, temperature T is considered as a formal parameter (the number of topics/clusters), which can be changed during the computational experiment. In turn, the Tsallis entropy, written in the form of $S^{Ts}_{q=1/T} = \frac{1-\sum_k p_k^q}{q-1}$, can also be expressed in terms of the Rényi entropy:

$$S^{Ts}_q = \frac{e^{(q-1)\cdot S^R_q} - 1}{q-1}, \tag{7}$$

and, consequently, in terms of free energy [46]. Thus, the modification of parameter $\mathrm{q} = 1/\mathrm{T}$ also allows us to investigate the behavior of the Tsallis entropy in topic modeling. It should be noted that with this approach, the entropy divergence is achieved at q=1, that is, the information obtained in topic modeling for one topic is zero. On the other hand, at $\mathrm{T}\to\infty$, we get a uniform probability distribution of words over topics, which also corresponds to the maximum entropy or minimum information.

4. Numerical investigation of the application of q-deformed entropy to determine the number of topics

4.1. Datasets

In this study computational experiments were performed on the following datasets.

- The well-known English-language dataset '20newsgroups' [47]: 15,404 news texts; 50,948 unique words. According to the description of the dataset, its data is organized into 20 different newsgroups, each corresponding to a different topic. As some of the newsgroups are very closely related to some others, the actual number of topics, according to the authors, equals to approximately 15. When conducting topic modeling on this dataset, the number of topics was varied in the range: T=[2; 120] in increments of 2 topics.
- All posts of the top 2000 bloggers of the Russian-language section of social network LiveJournal for January–April 2014: 101,481 posts; 172,939 unique words. This dataset contains a mixture of short conversational messages and long posts in a journalistic style. For this dataset, the number of topics was varied in the range: T=[2; 330] in increments of 2 topics.

These datasets were selected due to the following reasons. First of all, they are datasets in different languages. Therefore, our computational experiments show the applicability of the entropy approach to collections in different languages, and also reveal those model features that are language-independent. Secondly, since larger collections (LJ) usually contain a greater quantity of topics, that is, they are more diverse, it is logical to assume that they have more local entropy minima requiring separate attention. Thirdly, the English-language collection had been earlier used to test various clustering models [48], which make it possible to compare the results of topic modeling with the results of cluster analysis.

4.2. Experimental design

We studied behavior of Rényi and Tsallis entropies as functions of the number of topics, using the following software implementations of the four topic modeling algorithms:

A. pLSA (E–M algorithm)–BigARTM (http://bigartm.org/).
B. VLDA (E–M algorithm)–Latent Dirichlet Allocation package, http://chasen.org/~daiti-m/dist/lda/.
C. LDA GS (Gibbs sampling)–GibbsLDA++ (http://gibbslda.sourceforge.net/).
D. GLDA (Gibbs sampling) (https://linis.hse.ru/en/soft-linis)

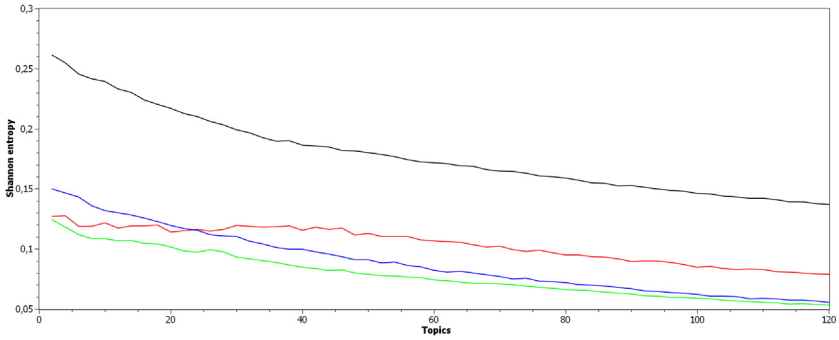

Fig. 1. Shannon entropy as a function of the number of topics on the 20 newsgroups dataset. Black: LDA GS, blue: pLSA; green: VLDA; red: GLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

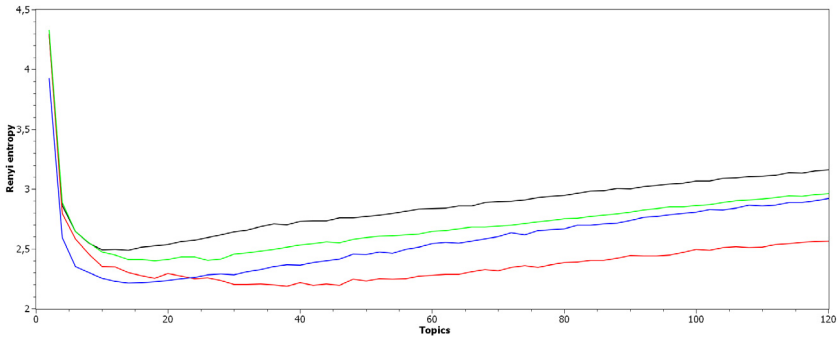

Fig. 2. Rényi entropy as a function of the number of topics on the 20 newsgroups dataset. Black: LDA GS, blue: pLSA; green: VLDA; red: GLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

All source codes were integrated into the single software tool 'TopicMiner' (https://linis.hse.ru/en/soft-linis/) as a set of dynamic link libraries (dll). Thus, computational experiments were carried out on collections that were processed in exactly the same way.

In each computational experiment, for each model, the number of microstates whose probabilities were greater than the preassigned value $P_{nt}>1/W$ was measured. Further, on the basis of formula (4), we calculated the function of dependence of density-of-states on the number of topics. Also, internal energy, entropy and free energy were calculated for each topic solution in accordance with formula (5). On the basis of the free energy, Rényi and Tsallis entropies were calculated using formulas (6) and (7) for each topic solution.

4.3. Discussion of the results of the experiments on the 20 newsgroups dataset

Figs. 1–3 plot Shannon, Rényi and Tsallis entropies dependence on the number of topics for all four topic models on the 20 newsgroups dataset. Each model was run three times, then the results of the calculation were averaged. The entropy values were calculated on the basis of the averaged values.

The LDA GS, pLSA and VLDA models produce similar curves without a pronounced minimum or maximum, while the GLDA model yields a small maximum in the range of 30–40 topics. The curves in Fig. 1 show that the greater the number of topics/clusters, the lower the entropy value and, correspondingly, the greater the value of information. However, this contradicts the actual experimental results, since an unlimited increase in the number of topics leads to the probability distribution of words over topics tending to a uniform distribution, which should correspond to an increase in entropy. This means that Shannon entropy is not suitable for analyzing such complex information systems, which in turn makes us to conclude that perplexity, the most commonly used quality measure of topic modeling, is not applicable either.

The Rényi entropy, in contrast to the Shannon entropy, has a global minimum, and shows the correct results on the boundary values of the number of topics. For $T=1$, the entropy should give a maximum, because topic modeling, just like any other cluster algorithm, does not give the distribution of clusters, so information about the cluster distribution is zero. At the same time, as noted above, an excessive increase in the number of clusters/topics (i.e., $T\rightarrow\infty$) leads to a uniform distribution of each word over topics, which also corresponds to an increase in entropy or a decrease in information. However, different models yield slightly different minimum Rényi entropy values, and different depths for those minima. In order to determine which of these models produces a more accurate result, we compare the results of topic modeling with alternative methods of determining the number of topics in the same collection. The authors of [48] tested a number of clustering algorithms

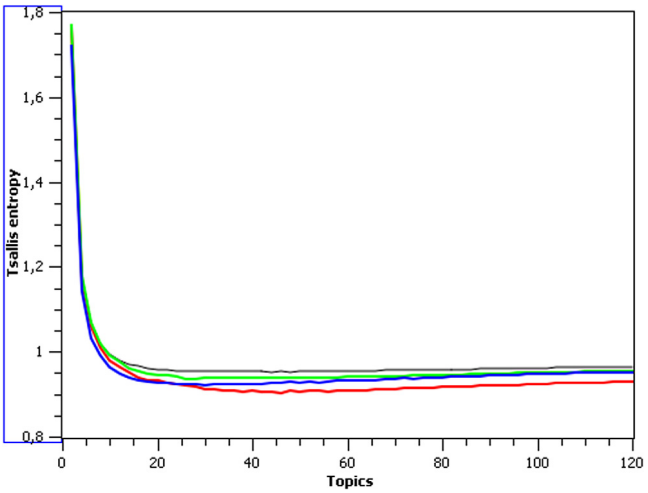

Fig. 3. Tsallis entropy as a function of the number of topics in the 20 newsgroups dataset. Black: LDA GS, blue: pLSA; green: VLDA; red: GLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

on the '20 newsgroup dataset' and showed that the optimal number of clusters varies from 15 to 20 for different cluster algorithms. This corresponds to the description of the dataset by its creators and coincides with the results yielded by pLSA, VLDA and LDA GS in our experiment.

Specifically, LDA GS and VLDA models also show the optimal number of topics around 15, the pLSA model shows 20, while the VLDA also shows the deepest minimum. However, the GLDA is significantly different, almost doubling the number of topics compared to other models, as well as to alternative methods of determining the number of topics. Thus, it can be concluded that the regularization procedure used in the GLDA model, although providing almost 100% stability [21], leads to a shift in the minimum Rényi entropy away from the correct value.

As can be seen from Fig. 3, the Tsallis entropy curves have significantly less pronounced minima, which makes it difficult to determine the optimal number of topics using them. In this case, the Tsallis entropy, like the Rényi entropy, gives the maximum values at the boundaries of the interval $[1; \infty]$. All models, with the exception of the GLDA, give a minimum of around 20 topics, which corresponds to the values obtained by the alternative method.

4.4. Discussion of the results of the experiments on the LJ dataset

A large number of documents in a collection can lead to the appearance of additional local minima, which should also be investigated. Moreover, these additional local minima can be of the greatest interest for text researchers, because clustering into 15–20 topics often yields topics that are too general (such as sports, politics or art whose presence in the news flow is evident without research). It can be assumed that solutions that divide global topics into more specific but not excessively fractional topics (for example, "politics in the Middle East" and "European politics") correspond to local minima of nonextensive entropy. A large text collection, namely LJ dataset was used to verify this assumption. The Rényi and Tsallis entropy curves for the LJ dataset are shown in Figs. 4 and 5.

First of all, it should be noted that for the LJ dataset the models based on the EM-algorithm show a marked difference from the models based on Gibbs sampling. The LDA GS model demonstrates the presence of strong jumps of Rényi entropy, which are associated with significant fluctuations in the density distribution function. However, the VLDA and pLSA models do not reveal such jumps. Fluctuations in the density distribution in the Gibbs sampling models cannot be explained by the features of the sampling procedure, since research [42] conducted on the same dataset had accounted for this. Specifically, in [42] the LDA GS model was run three times for each value of T, while T was varied in increments of 1 in the range of [105–120], and in increments of 10 in the range of [120–600]. The jump in the region [110–120] was observed in all runs of the model. This means that the models based on the Gibbs sampling appear more sensitive than the other models.

The Tsallis entropy calculated on the LDA GS model also shows jumps in the ranges [110–120] and [190–200]; however, the amplitude of these jumps is much lower than that of the jumps in the Rényi entropy function. The higher smoothness of Tsallis entropy functions on both datasets derives from the fact that Tsallis entropy is more Lesche stable [10,49]. Thus, the lack of Lesche stability in the Rényi entropy turns out to be useful for revealing local minima necessary for our task.

5. Numerical experiments on semantic stability in topic models

Since, as indicated in Section 2.2, text systems do not have the property of indistinguishability of particles, in investigating their behavior, it is necessary to check whether the word distribution is reproducible from the semantic point of view when

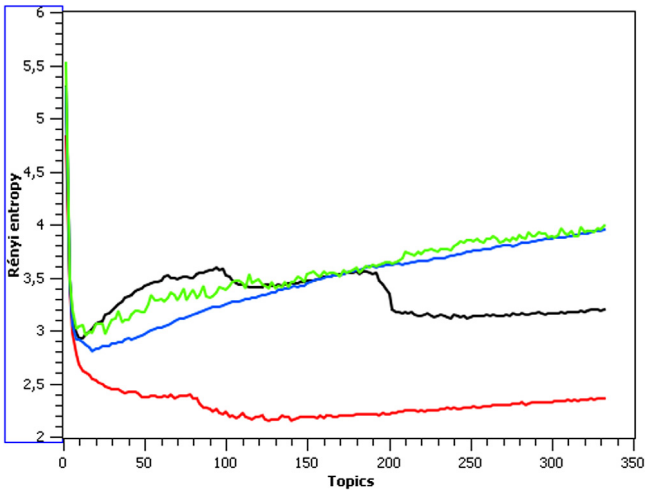

Fig. 4. Rényi entropy as a function of the number of topics in the LJ dataset. Black: LDA GS, blue: pLSA; green: VLDA; red: GLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

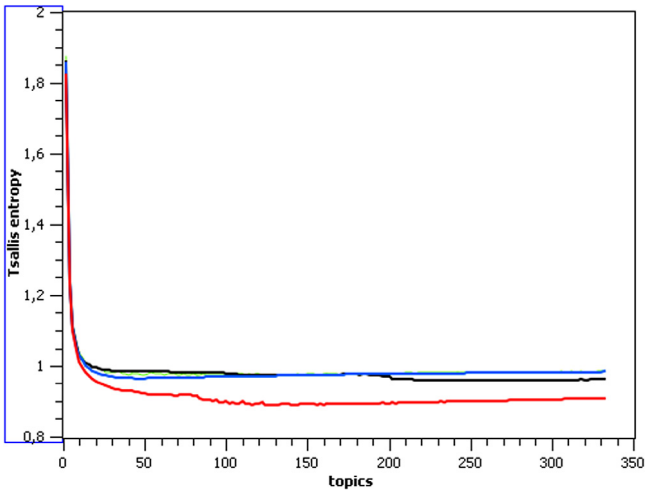

Fig. 5. Tsallis entropy as a function of the number of topics in the LJ dataset. Black: LDA GS, blue: pLSA; green: VLDA; red: GLDA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the parameter T (the number of topics) changes. In other words, it is essential to know how much the composition of words describing topics with high probability is T-invariant.

In this paper, T-invariance in topic models was measured using the Jaccard index [45] by the formula: $J_k = a/(a+b-c)$, where a is the set of the most probable words in topic solution T_1 that are absent from the list of most probable words in solution T_2; b is the set of words in solution T_2 absent from solution T_1, and c is the set of words common for solutions T_1 and T_2. The coefficient is equal to 1 if the two sets are identical and is equal to 0 if the sets have no common words. Highly probable words were defined as those whose probability was $P_w > 1/W$, where W is the number of unique words in the collection of documents.

To determine the effect of the number of topics T on the total composition of top words across multiple solutions, T was varied in increments of 2 in the range from 2 to 120 on the 20 newsgroups dataset, and from 2 to 330 on the LJ dataset. Then a pairwise comparison of each topic solution was made with all the other solutions. As a result of this calculation, a Jaccard index matrix was generated. Each cell of it contains the Jaccard index $J_{t1,t2}$, calculated between the lists of top words of each pair of solutions for which parameter T takes the values of t_1 and $t_2 = t_1+2$. Since such a matrix is symmetric with respect to the diagonal elements, the coefficients were calculated for only half the matrix.

Figs. 6 and 7 show the Jaccard diagonal coefficients curves according to the LDA GS and VLDA models for the LJ dataset. We do not show the Jaccard coefficients for the 20newsgroups dataset since all models yielded approximately the same values of about 0.999 for all solution pairs.

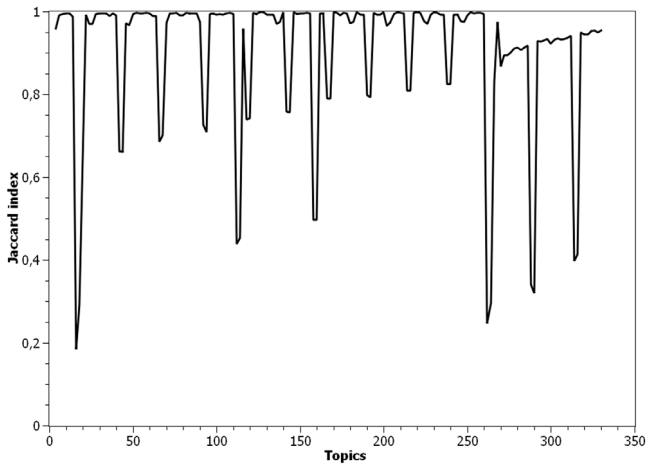

Fig. 6. The diagonal curve of the Jaccard index value for the LDA GS models.

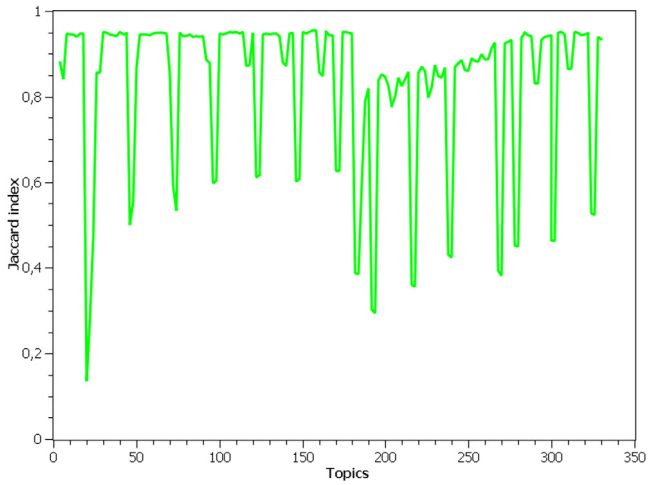

Fig. 7. The diagonal curve of the Jaccard index value for the VLDA models.

Figs. 6 and 7 show that, first, different models have a similar quasiperiodic semantic structure. This means that under the conditions of changing T, the lists of the most probable words are similar across most solutions, but periodically solutions occur that are very different from their neighbors. Second, T-invariance outside the atypical zone is not perfect, with Jaccard index abruptly dropping at some value of T in both models. And third, this drop occurs at different values of T for different models (around 260 for LDA GS and around 190 for VLDA).

Figs. 8 and 9 show the 'heat maps' of the Jaccard index for the two models.

Figs. 8 and 9 confirm the existence of T-invariance zones (with $J_{t1,t2} \cong 0.9$) and atypical zones (with $J_{t1,t2} \cong 0.5$) for both types of models. Moreover, the distributions of Jaccard indices shown in Figs. 8 and 9 demonstrate the presence of two quasiperiodic structures that overlap one another. This may indicate that the degree of T-invariance is described by several parameters. Analysis of these parameters is an extremely important task for future studies. At present, for the tasks of selecting the optimal number of topics in the topic models, it can be recommended to choose not only the global and local minima of Rényi or Tsallis entropies, but also to avoid zones where the T-invariance principle is violated.

6. Conclusion

In this paper, we have proposed an entropy-based approach to the analysis of behavior of complex text systems, which allows finding the optimal number of topics in topic models. A major finding of our research is that we discover a theoretically grounded concept which, when represented as a function of the number of topics, has an extremum. The latter feature fundamentally distinguishes it from previously used monotonically decreasing functions (such as Shannon entropy or perplexity) that made it impossible to determine the threshold after which the increase in the number of topics becomes useless or even harmful. Specifically, in this paper the search for the optimal number of topics is based on minimizing Rényi

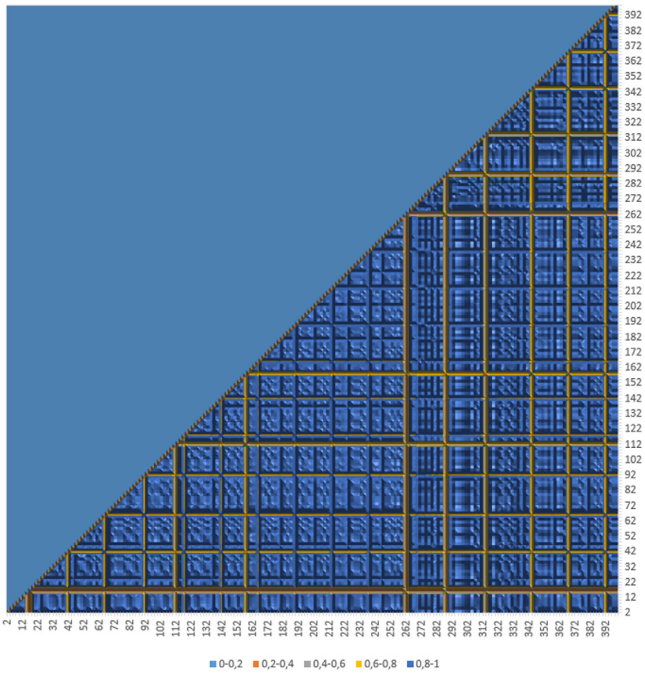

Fig. 8. Heat map of Jaccard index distribution for LDA GS models.

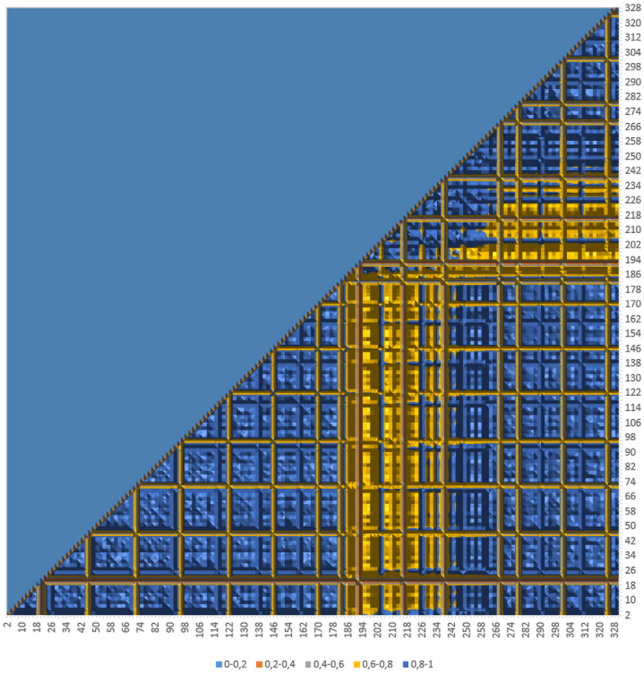

Fig. 9. Heat map of Jaccard index distribution for VLDA models.

and Tsallis entropies, with Rényi entropy giving the most pronounced minima. Both entropies successfully indicate a well-pronounced deterioration in the model with an increase in the number of topics in the interval from the optimum to infinity, which corresponds to empirical knowledge but which is not captured by the traditionally used monotone functions.

In addition, it is shown that topic models both on the basis of Gibbs sampling and on the basis of EM-algorithms give similar results in the area of the global minimum of nonextensive entropy. However, models based on Gibbs sampling show additional local minima that may be of interest for a comprehensive analysis of large text data in the social sciences.

Finally, the work introduces the concept of T-invariance and, while studying it, reveals the existence of two quasiperiodic semantic structures describing the dependence of the change of the total composition of top words of all topics on the number of topics. These structures are not yet included in the theoretical entropy model, but it is already clear that they are essential for determining the optimal number of topics. Further development of the proposed theoretical approach and the respective model can be obtained by extending it with the two-parameter Sharma–Mittal entropy. The latter generalizes the entropies of Rényi, Tsallis and Kaniadakis, and may help explain the observed quasiperiodic effect.

Acknowledgments

The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE), Russia in 2017. I am also thankful for Svetlana Alexeeva for help in data collection and to Olessia Koltsova for her aid in the paper translation.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2018.08.050.

References

[1] Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Navigating the local modes of big data: The case of topic models, in: R. Michael Alvarez (Ed.), Computational Social Science: Discovery and Prediction, Cambridge University Press, 2016, pp. 51–97, http://dx.doi.org/10.1017/CBO9781316257340.004.
[2] D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel, Nature 439 (7075) (2006) 462–465, http://dx.doi.org/10.1038/nature04292, Nature Publishing Group.
[3] C. Song, T. Koren, P. Wang, A.L. Barabási, Modelling the scaling properties of human mobility, Nat. Phys. 6 (2010) 818–823, http://dx.doi.org/10.1038/nphys1760.
[4] Pablo M. Gleiser, Francisco A. Tamarit, Sergio A. Cannas, Self-organized criticality in a model of biological evolution with long-range interactions, Physica A 275 (1) (2000) 272–280, http://dx.doi.org/10.1016/S0378-4371(99)00425-2, Elsevier Science Publishers B.V..
[5] K. Friston, M. Levin, B. Sengupta, G. Pezzulo, Knowing one's place: A free-energy approach to pattern regulation, J. R. Soc. Interface 12 (105) (2015) http://dx.doi.org/10.1098/rsif.2014.1383, 20141383–20141383.
[6] Lisa Borland, Option pricing formulas based on a non-Gaussian stock price model, Phys. Rev. Lett. 89 (9) (2002) http://dx.doi.org/10.1103/PhysRevLett.89.098701.
[7] Rosario N. Mantegna, H. Eugene Stanley, Neil A. Chriss, An introduction to econophysics: Correlations and complexity in finance, Phys. Today 53 (12) (2000) http://dx.doi.org/10.1063/1.1341926, 70–70.
[8] Dimitrije Marković, Claudius Gros, Power laws and self-organized criticality in theory and nature, Phys. Rep. (2014) http://dx.doi.org/10.1016/j.physrep.2013.11.002.
[9] Lada Adamic, Eytan Adar, How to search a social network, Social Networks 27 (3) (2005) 187–203, http://dx.doi.org/10.1016/j.socnet.2005.01.007.
[10] Constantino Tsallis, Introduction to nonextensive statistical mechanics: Approaching a complex world, in: Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World, Springer New York, 2009, http://dx.doi.org/10.1007/978-0-387-85359-8.
[11] A.G. Bashkirov, On maximum entropy principle, superstatistics, power-law distribution and Renyi parameter, Physica A 340 (2004) 153–162, http://dx.doi.org/10.1016/j.physa.2004.04.002.
[12] Christian Beck, Generalised information and entropy measures in physics, Contemp. Phys. 50 (4) (2009) 495–510, http://dx.doi.org/10.1080/00107510902823517.
[13] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022, http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993.
[14] I. Chernyavsky, T. Alexandrov, P. Maass, S. Nikolenko, A two-step soft segmentation procedure for MALDI imaging mass spectrometry data, in: GCB, 2012, pp. 39–48.
[15] Nguyen Anh Tu, Dong-Luong Dinh, Mostofa Kamal Rasel, Young-Koo Lee, Topic modeling and improvement of image representation for large-scale image retrieval, Inform. Sci. 366 (2016) 99–120, http://dx.doi.org/10.1016/j.ins.2016.05.029.
[16] Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad, Knowledge discovery through directed probabilistic topic models: A survey, Front. Comput. Sci. China (2010) http://dx.doi.org/10.1007/s11704-009-0062-y.
[17] Thomas Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 50–57. http://dx.doi.org/10.1021/ac801303x.
[18] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Supplement 1) (2004) 5228–5235, http://dx.doi.org/10.1073/pnas.0307752101.
[19] Sergei Koltcov, Olessia Koltsova, Sergey I. Nikolenko, Latent Dirichlet allocation: Stability and applications to studies of user-generated content, in: Proceedings of the 2014 ACM Conference on Web Science, 2014, pp. 161–165 http://dx.doi.org/10.1145/2615569.2615680.
[20] S. Koltcov, S.I. Nikolenko, O. Koltsova, V. Filippov, S. Bodrunova, Stable topic modeling with local density regularization, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9934, LNCS, 2016, http://dx.doi.org/10.1007/978-3-319-45982-016.
[21] S. Koltsov, S.I. Nikolenko, O. Koltsova, Gibbs sampler optimization for analysis of a granulated medium, Tech. Phys. Lett. 8 (42) (2016) 837–839, http://dx.doi.org/10.1134/S1063785016080241.
[22] Catherine Sugar, James Gareth, Finding the number of clusters in a data set : An information theoretic approach, J. Amer. Statist. Assoc. 98 (2003) 750–763, http://dx.doi.org/10.1198/016214503000000666.
[23] Boris Mirkin, Clustering for Data Mining - A Data Recovery Approach, New York, 2005, http://www.amazon.com/dp/1584885343.
[24] Robert Tibshirani, Guenther Walther, Trevor Hastie, Estimating the number of clusters in a data set via the gap statistic, J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (2) (2001) 411–423, http://dx.doi.org/10.1111/1467-9868.00293.
[25] André Fujita, Daniel Y. Takahashi, Alexandre G. Patriota, A non-parametric method to estimate the number of clusters, Comput. Statist. Data Anal. 73 (2014) 27–39, http://dx.doi.org/10.1016/j.csda.2013.11.012.
[26] Glenn W. Milligan, Martha C. Cooper, An examination of procedures for determining the number of clusters in a dataset, Psychometrika (1985) http://dx.doi.org/10.1007/BF02294245.

[27] Sung-Hyuk Cha, Taxonomy of nominal type histogram distance measures, in: Proceedings of the American Conference on Applied Mathematics, vol. 2, 2008, pp. 325–330 http://www.csis.pace.edu/~scha%5Cnhttp://dl.acm.org/citation.cfm?id=1415583.141564.
[28] Chun-Hung Cheng, Ada Waichee Fu, Yi Zhang, Entropy-based subspace clustering for mining numerical data, in: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '99, ACM Press, New York, New York, USA, 1999, pp. 84–93, http://dx.doi.org/10.1145/312129.312199.
[29] Edwin Aldana-Bobadilla, Angel Kuri-Morales, A clustering method based on the maximum entropy principle, Entropy 17 (1) (2015) 151–180, http://dx.doi.org/10.3390/e17010151, MDPI AG.
[30] Kenneth Rose, Eitan Gurewitz, Geoffrey C. Fox, Statistical mechanics and phase transitions in clustering, Phys. Rev. Lett. 65 (8) (1990) 945–948, http://dx.doi.org/10.1103/PhysRevLett.65.945.
[31] Wassim M. Haddad, VijaySekhar Chellaboina, Sergey G. Nersesov, Thermodynamics: A Dynamical Systems Approach, Princeton University Press, 2005.
[32] Greg J. Stephens, Thierry Mora, Gašper Tkačik, William Bialek, Statistical thermodynamics of natural images, Phys. Rev. Lett. 110 (1) (2013) 18701.
[33] Gašper Tkačika, Thierry Morab, Olivier Marrec, Dario Amodeide, Stephanie E. Palmerdf, Michael J. Berry IIeg, William Bialek, Thermodynamics and signatures of criticality in a network of neurons, Proc. Natl. Acad. Sci. USA 112 (37) (2015) 11508–11513, http://dx.doi.org/10.1073/pnas.1514188112.
[34] R.C. Venkatesan, A. Plastino, Deformed statistics free energy model for source separation using unsupervised learning, (2) (2011) 5 Arxiv Preprint. http://arxiv.org/abs/1102.5396.
[35] André F.T. Martins, Noah A. Smith, Eric P. Xing, Pedro M.Q. Aguiar, Mário A.T. Figueiredo, Nonextensive information theoretic kernels on measures, J. Mach. Learn. Res. 10 (2009) 935–975, http://jmlr.csail.mit.edu/papers/v10/martins09a.html.
[36] R.C. Venkatesan, A. Plastino, Generalized statistics framework for rate distortion theory, Physica A 388 (12) (2009) 2337–2353, http://dx.doi.org/10.1016/j.physa.2009.02.003.
[37] Abdiel Ramírez-Reyes, Alejandro Raúl Hernández-Montoya, Gerardo Herrera-Corral, Ismael Domínguez-Jiménez, Determining the entropic index q of Tsallis entropy in images through redundancy, Entropy 18 (8) (2016) http://dx.doi.org/10.3390/e18080299, MDPI AG.
[38] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, Sheng Tang, A density-based method for adaptive LDA model selection, Neurocomputing 72 (7–9) (2009) 1775–1781, http://dx.doi.org/10.1016/j.neucom.2008.06.011.
[39] R. Arun, V. Suresh, C.E. Veni Madhavan, M. Narasimha Murty, On finding the natural number of topics with latent Dirichlet allocation: Some observations, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, pp. 391–402 http://dx.doi.org/10.1007/978-3-642-13657-3_43, 6118 LNAI.
[40] Yee Whye Teh, Michael I. Jordan, M. Beal, David Blei, Hierarchical Dirichlet processes, J. Am. Stat. … (2006) 1–41, http://dx.doi.org/10.1017/CBO9781107415324.004.
[41] D. Blei, T.L. Griffiths, M.I. Jordan, J.B. Tenenbaum, Hierarchical topic models and the nested Chinese restaurant process, Adv. Neural Inf. Process. Syst. 16 (2004) 106, http://dx.doi.org/10.1016/0169-023X(89)90004-9.
[42] S.N. Koltcov, A thermodynamic approach to selecting a number of clusters based on topic modeling, Tech. Phys. Lett. 43 (6) 584–586 http://dx.doi.org/10.1134/S1063785017060207.
[43] I. Prigogine, I. Stengers, La Nouvelle Alliance, Gallimard, Paris, 1979.
[44] Yu L. Klimontovich, Problems in the statistical theory of open systems: Criteria for the relative degree of order in self-organization processes, Sov. Phys. Usp. 32 (1989) 416–433.
[45] Paul Jaccard, The distribution of the flora in the alpine zone, New Phytol. 11 (1912) 37–50, http://dx.doi.org/10.1111/j.1469-8137.1912.tb05611.x.
[46] C. Beck, F. Schlögl, Thermodynamics of Chaotic Systems, Cambridge University Press, Cambridge, 1993.
[47] The 20 Newsgroups data set, http://qwone.com/~jason/20Newsgroups/.
[48] S. Basu, I. Davidson, K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications, Chapman & Hall, 2008.
[49] B. Lesche, Instabilities of Renyi entropies, J. Stat. Phys. 27 (1982) 419.