Hyper-parameters Tuning in Topic Modeling based on Renyi Entropy and Deformed Perplexity

Anonymous Author(s)

ABSTRACT

We propose a novel approach for estimating the optimal values of hyper-parameters in topic modelling based on Renyi entropy and deformed perplexity. This approach is inspired by the concepts of statistical physics, where the collection of documents and the set of words can be considered an information statistical system residing in a nonequilibrium state. We introduce a notion of 'deformed perplexity' which is expressed in terms of Renyi entropy and can also be used for tuning the values of hyper-parameters. We apply this approach to three topic models: pLSA, BigARTM and LDA Gibbs sampling, by studying the functional dependence of the topic modeling results on the values of hyper-parameters in terms of Renyi entropy. We experimentally demonstrate the effectiveness of the proposed approach for three datasets of different sizes and in different languages (i.e. English, Russian and French).

CCS CONCEPTS

• **Computing methodologies** → **Topic modeling**; *Regularization*; • **Theory of computation** → Unsupervised learning and clustering;

KEYWORDS

Topic Modeling, Renyi Entropy, Regularization

ACM Reference Format:
Anonymous Author(s). 2019. Hyper-parameters Tuning in Topic Modeling based on Renyi Entropy and Deformed Perplexity. In *Proceedings of ACM The Web Conference (WWW'19).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Nowadays, topic modelling (TM) is represented be three major groups of models: 1. Models based on Gibbs sampling procedure, 2. Models based on Expectation-Maximisation (EM) algorithm, and 3. Hierarchical topic models. Unfortunately, all of them share a common problem: they all lack criteria to optimize their parameters, such as the number of topics, parameters of Dirichlet distribution or regularization coefficients [3, 37]. The task of TM is equivalent to stochastic matrix decomposition, where a larger matrix F containing distribution of words w by documents d is approximated by the multiplication of two matrices $\Theta = (\theta_{td})$ and $\Phi = (\phi_{wt})$ of lower dimensions. However, stochastic matrix decomposition is defined not uniquely but with accuracy up to a non-degenerate transformation [38]. If $F = \Phi\Theta$ is a solution then $F = (\Phi S)(S^{-1}\Theta)$ is also a solution for all non-degenerate S under which $\Phi' = \Phi S$ and $\Theta' = S^{-1}\Theta$ are stochastic matrices. In terms of TM, ambiguity in retrieving the multidimensional density of distribution mixture means that the algorithm starting from different initial approximations will conjugate to different points of the solution set. This is expressed in the fact that different runs of the algorithm on the same source data give different output matrices Θ and Φ. The problems that have non-unique or non-stable solutions are termed ill-posed [36]. A general approach to avoiding multiple solutions is given by Tikhonov regularization [36]. The essence of regularization is to redefine a priori information that allows for narrowing the set of solutions by introducing restrictions on matrices Θ and Φ [31] and by modifying the sampling procedure [2]. Furthermore, regularization can be achieved by introducing a combination of conjugate functions [11] and different types of regularization procedures [37, 38]. Thus, TM parameter optimization is a significant problem that still needs an extensive research. As a partial solution, we propose an approach based on the concepts of statistical physics. Here, a collection of documents is considered an information thermodynamic system. For such a system, Renyi entropy can be introduced within the thermodynamic formalism [33] analogously to [21]. The values of hyper-parameters or regularization parameters are independently set and must be determined by searching for the minimum nonextensive entropy of the system. So, the optimal values of parameters correspond to the situation when an information measure is at its maximum (correspondingly, an entropy is in its minimum [8]). It is important to note that the proposed approach does not apply to hierarchical models, which would demand its modification and, therefore, a special research. The rest of this paper is organised as follows. The second section briefly discusses three different topic models characterized by different types of regularization with parameters and hyper-parameters. In the third section, approaches to determining parameters of topic models are studied. The fourth section begins with introduction of Renyi entropy and deformed perplexity that are proposed as the criteria to optimize parameters and hyper-parameters in topic models. The fifth section presents the experiments carried out on several datasets. Finally, the overall analysis of the obtained results is presented in the sixth section.

2 BACKGROUND AND RELATED WORK

2.1 Basics of Topic Modelling.

TM is a family of mathematical algorithms based on the following assertions [17]:

1. Let D be a collection of textual documents, W be a set of all unique terms (vocabulary). Each document $d \in D$ is a bag of terms $w_1, ..., w_{n_d}$ from the vocabulary W.

2. It is assumed that there exists a finite number of topics T, and each entry of a word w in document d is associated with at least one topic $t \in T$. A topic is considered to be a combination of words

WWW'19, May 2019, San Francisco, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

which are often (in statistical sense) reproduced together in a large number of documents.
3. A collection of documents is considered a stochastic and independent sample of triples (w_i, d_i, t_i), $i = 1, ..., n$ from a discrete distribution $p(w, d, t)$ on a finite probability 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 words in the set of documents is not important for TM ('bag-of-words' model). Similarly, the order of documents in a collection is also insignificant.
In TM, the appearance probability $p(w|d)$ of a term w in a document d can be expressed as follows:

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

where $p(w|t) = \phi_{wt}$ is the appearance probability of a word w under a topic t, $p(t, d) = \theta_{td}$ is the probability of a topic t in a document d.

Thus, constructing a topic model from a set of documents means solving a problem, where it is necessary to find the set of latent topics T based on observable variables d and w, i.e. the goal is to obtain a set of one-dimensional conditional probabilities $p(w|t) \equiv \phi_{wt}$ for each topic t, which form a matrix $\Phi \equiv (\phi_{wt})_{w \in W, t \in T}$ expressing the distribution of words over topics and a set of one-dimensional conditional probabilities $p(t|d) \equiv \theta_{td}$ for each document d, which form a matrix $\Theta \equiv (\theta_{td})_{t \in T, d \in D}$ expressing the distribution of topics over documents. Different types of topic models are related to different regularization algorithms. In the following, we consider several algorithms, which are used in our experiments.

2.2 Probabilistic Latent Semantic Analysis (PLSA)

In the framework of this model, the determination of the matrices Φ and Θ is performed as described in [17]. The entire dataset is generated as:

$$p(D) = \prod_{d \in D} \prod_{w \in W} p(d, w)^{n(d,w)} = \prod_{d \in D} \prod_{w \in W} p(d)^{n(d,w)} p(w|d)^{n(d,w)}$$

$$= \prod_{d \in D} \prod_{w \in W} p(d)^{n(d,w)} \sum_{t \in T} p(w|t)^{n(d,w)} p(t|d)^{n(d,w)}$$

where $p(d, w)$ is the joint probability distribution, $n(d, w)$ counts the appearance frequency of the term w in the document d. Note that this model involves a conditional independence assumption, namely, d and w are independently conditioned on the state of the associated latent variable [17].

The estimation of the one-dimensional distributions is based on log-likelihood maximization with linear constraints:

$$L(\phi, \theta) = \sum_{d \in D} \sum_{w \in W} n(d, w) \ln \left[p(d) \sum_{t \in T} \phi_{wt}\theta_{td} \right] \to \max_{\phi, \theta} L(\phi, \theta),$$

where $\phi_{wt} \geq 0$, $\sum_{w \in W} \phi_{wt} = 1$, $\theta_{td} \geq 0$, $\sum_{t \in T} \theta_{td} = 1$.

The determination of the local maximum of $L(\phi, \theta)$ is carried out using Expectation-Maximization (E-M) algorithm. The initial approximation of ϕ_{wt} and θ_{td} is chosen randomly or uniformly before the first iteration.

E - step: using Bayes' rule, conditional probabilities $p(t|d, w)$ are calculated for all $t \in T$ and each $w \in W$, $d \in D$ [18], namely:

$$p(t|d, w) = \frac{p(d, w|t)p(t)}{p(d, w)} = \frac{p(d|t)p(w|t)p(t)}{p(d) \sum_{s \in T} p(w|s)p(s|d)} =$$

$$= \frac{p(w|t)p(t|d)}{\sum_{s \in T} p(w|s)p(s|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}.$$

M-step: using conditional probabilities, new approximations of ϕ_{wt}, θ_{td} are estimated, namely:

$$\phi_{wt} = \frac{\sum_{d \in D} n(d, w)p(t|d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)p(t|d, w)},$$

$$\theta_{td} = \frac{\sum_{w \in W} n(d, w)p(t|d, w)}{\sum_{t \in T} \sum_{w \in W} n(d, w)p(t|d, w)}.$$

Thus, alternating E and M steps in a cycle, $p(t|d)$ and $p(w|t)$ can be estimated. Note that this model has no additional parameters except of 'the number of topics', which defines the size of matrices Φ, Θ.

2.3 The Additive Regularization of Topic Models (ARTM) (Models based on E-M algorithm)

The algorithm of additive regularization is an extended version of pLSA [38]. This model was developed to combine different types of topic models in order to create a new model with the desired properties for specific applications [20]. The main idea of the algorithm is based on regularized Maximum Likelihood Estimation, i.e., the following optimization problem is considered:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

where

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln(\sum_{t \in T} \phi_{wt}\theta_{td}).$$

$$R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta),$$

The regularization term R is given not only by functions R_i, but also by values of regularization coefficients τ_i. The local extremum of the above optimization task is searched using E-M algorithm. In the framework of this model a large number of different regularizers is introduced [38]. Despite the large number of possibilities of this approach, the method of additive regularization does not help with choosing regularization parameters τ_i. Basically, the selection of regularization coefficients is carried out manually taking into account the perplexity stabilization [11], [37]. Generally, the problem of selecting regularizers and their coefficient values is still in the research core for this type of models.

Let us briefly discuss two particular sparsing regularizers (called SparsePhi, SparseTheta). The functionality of such regularizers is mainly to control the sparseness of Φ and Θ in order to ensure that each document and each word are related to a small number of topics [38]. Sparsing regularizer for Φ and Θ can be written as follows [38]:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td},$$

where $(\beta_w)_{w\in W}$, $(\alpha_t)_{t\in T}$ are given discrete distributions, for instance, uniform distributions. Note that the coefficients α_0, β_0 are included in the maximization procedure. Moreover, as it was noted in [37], the regularizers can conflict with each other, worsen the convergence of the topic model for points away from the set of solutions or lead to degeneracy of the model.

2.4 LDA (Model based on Gibbs sampling procedure)

LDA Gibbs sampling model is a topic model, in which each topic is smoothed by the same regularizer in the form of Dirichlet function [15]. In this model, each document is considered as one-dimensional grid and each word of the document is considered a node. A node can reside in one of T states (topics). The goal of examining such a Potts model is to estimate the distribution of nodes (words) over the set of states, i. e. by topics. The difference between Potts model and topic modeling is that in TM, the amount of documents can be huge, and the probability of a node to belong to a state (topic) is defined not only by the distribution of words over topics in one document but also by the distribution of topics over documents. According to Blei et al. [11], it is assumed to use Dirichlet distributions with one-dimensional parameters β and α, correspondingly, in order to simplify the derivation of analytical expressions for the matrices Φ and Θ. On this basis, the probability of the ith word in a given document d is defined as follows [15]:

$$\begin{aligned} p(w_{i,d}) = \sum_{j=1}^{T} p(w_{i,d}|z_{i,d}=j)p(z_{i,d}=j) = \sum_{j=1}^{T} \phi_{wj}\theta_{dj} = \\ = \sum_{j=1}^{T} \frac{c_{d,j}+\alpha}{\sum_{j=1}^{T} c_{d,j}+\alpha T} \cdot \frac{c_{w,j}+\beta}{\sum_{w=1}^{W} c_{w,j}+\beta W}, \end{aligned} \tag{2}$$

where $z_{i,d}$ is a latent variable (topic), $p(w_{i,d}|z_i = j)$ is the probability of the word w_i in document d under the jth topic, $p(z_{i,d} = j)$ is the probability of choosing a word from topic j in the current document d, $w_{i,d}$ is the ith word in document d, counter $c_{d,j}$ is the number of words in document d assigned to topic j, counter $c_{w,j}$ is the number of word w is assigned to topic j; $\sum_{j=1}^{T} c_{d,j}$ is the total number of words in document d (i.e. length of document d), $\sum_{w=1}^{W} c_{w,j}$ is the total number of words assigned to topic j. Correspondingly, θ and ϕ can be obtained as follows:

$$\theta_{dj} = \frac{c_{d,j}+\alpha}{\sum_{j=1}^{T} c_{d,j}+\alpha T}, \tag{3}$$

$$\phi_{wj} = \frac{c_{w,j}+\beta}{\sum_{w=1}^{W} c_{w,j}+\beta W}. \tag{4}$$

The algorithm of calculation consists of three phases. The first one is the initialization of matrices, counters and parameters α and β, in addition to specifying the number of iterations. Counters, which define the initial values of matrices Φ and Θ, are set as constants. So, matrices are filled with constants, for example, Φ can be filled with uniform distribution, where all elements of the matrix are equal to $1/W$, where W is the number of unique words in a collection of documents.

The second phase (sampling procedure) is an exhaustive search through all the documents and all words in each document in a cycle. Each word w_i in a given document d is matched with the topic number, which is generated as follows:

$$p(z_i = j|z_{-i}) \approx \frac{c_{d,j}^{-i}+\alpha}{\sum_{j=1}^{T} c_{d,j}^{-i}+\alpha T} \cdot \frac{c_{w_i,j}^{-i}+\beta}{\sum_{w=1}^{W} c_{w,j}^{-i}+\beta W},$$

where $c_{d,j}^{-i}$ is the number of words from document d assigned to topic j not including the current word w_i, $c_{w,j}^{-i}$ is the number of instances of word w assigned to topic j not including the current instance i, $c_{d,j}^{-i}$ and $c_{w,j}^{-i}$ are called counters. Here, the probabilities of belonging of the current word to different topics are calculated, then the most probable topic is assigned to the current word. The initial probability of word-topic matching is defined only by $1/T$ and $1/W$ when considering a uniform distribution as the initial approximation of matrix Φ. However, after each word matching to a topic, the values of counters change and, hence, after an important number of iterations, counters contain the full statistics of document collection under study.

At the third phase, Φ and Θ are calculated according to the equations 3 and 4. Finally, the matrices are ready for manual analyses, where for sociological analysis, only the most probable words and documents for each topics are considered. Note that the coefficients α and β defining Dirichlet distribution are parameters of this model, which one has to select.

3 PROBLEMS OF HYPER-PARAMETERS ESTIMATIONS

To determine the values of parameters in topic modelling, two functions are most often employed for this purpose: 1) perplexity, 2) Kullback-Leibler divergence.

3.1 Perplexity and likelihood

The perplexity is a standard criterion for topic models that evaluates the efficiency of the model to predict the new data. Specifically, the perplexity of a set of M testing documents (d_i, $i = 1, ..., M$) is defined as [11, 27]: $\text{perplexity}(D_{\text{test}}) = exp\left(-\frac{\sum_{i=1}^{M} \log p(d_i)}{\sum_{i=1}^{M} N_i}\right)$, where N_i is the number of words in document d_i. The lower the perplexity score is the better the parameters' values are. Generally, perplexity can be expressed in terms of entropy in the following form: $\text{perplexity} = 2^{\text{entropy}}$ or $\text{perplexity} = e^{\text{entropy}}$ [12], [14], where entropy is the Gibbs-Shannon entropy. The use of perplexity for the selection of parameters for topic models is discussed in a number of works [11, 15, 30]. In [30], the convergence of topic models such as LDA Gibbs sampling and HD-LDA is studied, where it has been observed that the perplexity behaves as a monotone decreasing function of iteration number. Thus, the perplexity is a convenient clue for determining the optimal number of iterations in topic models [11, 15, 34]. Moreover, the perplexity is used in work [30] also for determining the optimal number of topics. The authors demonstrated that the perplexity decreases monotonously, by increasing the number of topics. Such a behaviour is typical for Gibbs sampling algorithm and for hierarchical models.

The use of perplexity has some limitations, which are reviewed in [13]. The authors demonstrated that the value of perplexity depends on the vocabulary size of the given collection, used for topic modeling. The dependence of perplexity value on type of topic model and size of vocabulary is shown in [34] as well. Thus, the comparison of topic models, conducted on different datasets and different languages using the perplexity is complicated [5, 40] due to the aforementioned reasons and therefore perplexity-based methods are not stable.

Some works show another behaviour of perplexity, for example, authors of [4] show that the perplexity as a function of hyper-parameters has a notable unique minimum for collapsed Gibbs sampling (CGS) model, variational Bayesian inference (VB) model and collapsed variational Bayesian inference (CVB). Also, authours of [40] show that the perplexity as a function of topic number has a notable minimum for hierarchical topic model, and maximal values of perplexity correspond to minimum and maximum of numbers of topics (i.e. for $T \to 1$ and $T \to \infty$). In [5], it has been shown that the perplexity, used for a model with feature regularization, has clear minimum for some values of varying parameters. Also, the maximum value of perplexity corresponds to the maximum value of varying parameter. Thus, it can be noticed that different types of perplexity behaviour can be found in literature on TM without an explanation of such behaviour.

Another measure, which is often used when analyzing the results of topic modeling, is logarithm of likelihood [15] $p(\hat{w}|T)$, where $\hat{w}$ is a corpus . Usually, the calculation of this value is carried out when the perplexity stops changing and no further iterations are needed. Correspondingly, the hyper-parameters and number of topics are selected when finding maximum of logarithm of likelihood [15]. Notice that logarithm of likelihood is a version of perplexity and different types of probability logarithm behaviour are shown in literature as well as for perplexity.

3.2 Kullback-Leibler divergence

Another measure, which is widely used for analysing topic models is Kullback-Leibler divergence (KL) (or relative entropy) [10, 25]. In the field of TM, symmetric Kullback-Leibler divergence is most comonly used, which was proposed by Steyvers and Griffiths [35] for determining the number of stable topics. The dissimilarity between two topics, j_1 and j_2, is measured as follows:

$$KL(j_1, j_2) = \frac{1}{2}\sum_{k=1}^{W} \phi_k'^{(j_1)} \log_2 \phi_k'^{(j_1)} / \phi_k''^{(j_2)} + \\ + \frac{1}{2}\sum_{k=1}^{W} \phi_k''^{(j_2)} \log_2 \phi_k''^{(j_2)} / \phi_k'^{(j_1)},$$

where ϕ' and ϕ'' correspond to the estimated topic-word distributions from two different runs. The topics of the second run are re-ordered to correspond as best as possible (using a greedy algorithm) to the topics of the first run [35].

Further, based on this measure, the algorithm for finding stable topics for different topic models is proposed in [24]. In this approach, pairwise comparison is carried out for all topics of a solution with all topics of another solution. Thus, if a topic is stable then it is regularly reproduced in each run of the algorithm. Work [24] shows that different topic models give different number of stable topics applied to the same dataset.

Let us note that in the field of statistical physics, KL divergence is closely related to free energy. In the framework of Boltzmann-Gibbs statistics, KL divergence can be expressed as: $KL(p|q) = \beta[F(p) - F(q)]$ [1] , where p and q are the probability distributions of a system residing in non-equilibrium and equilibrium states, respectively. $F(p)$ and $F(q)$ denote the free energies of the system [33] in non-equilibrium and equilibrium states, respectively, and T is the temperature of the system. Let us remark that the free energy principle tries to explain how (biological, physical, economics) systems maintain their order (non-equilibrium steady state) by restricting themselves to a limited number of states. Free energy is expressed in terms of Shannon-Gibbs entropy and internal energy by following formula: $F = E - TS$, where E is internal energy, S is Shannon-Gibbs entropy, T is temperature. Consequently, KL divergence is the difference of the off-equilibrium and equilibrium free energies. The difference between free energies is a key feature of thermodynamic approach [23], which is discussed subsequently.

3.3 Selection of hyper-parameters in topic models

In general, the hyper-parameters α and β of LDA model have a smoothing effect on the results of TM, that influences the sparsity of matrices Φ and Θ [16]. The sparsity of matrices influences, in turn, the number of topics, which can appear in a document collection. Consequently, the number of topics may implicitly depend on the values of hyperparameters. Work [15] suggests a rule to select hyperparameters: $\alpha = 50/T$ and $\beta = 0.01$, where T is the number of topics. Such values of parameters were widely used in different studies [2, 26, 29]. On the other side, the effect of changing the values of hyperparameters was studied in the following form: $\alpha' = m \cdot \alpha$, $\beta' = n \cdot \beta$ varying n and m [39], where the influence was analyzed using logarithm of probability. It has to be noted that the behaviour of probability logarithm is different in [15, 39] due to the different used datasets. Therefore, the use of probability logarithm is not always justified when working with different types of datasets, such as in cross-national studies where the results of topic models in different languages has to be compared. Asuncion et al. [4] describe how to select the values of hyperparameters as well as the optimal number of topics for four different topic models (collapsed Gibbs sampling, variational Bayesian inference, 2 versions of collapsed variational Bayes) using perplexity and Minka' algorithm [28].

4 RENYI ENTROPY, DEFORMED PERPLEXITY AND KULLBACK-LEIBLER DIVERGENCE IN TOPIC MODELLING

The author of work [21] proposes an entropy-based approach for analysing the results of topic modelling. Here, 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' [6]. 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 Renyi 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. 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 presumably should be constant. It is these stable structures that should be revealed through topic modeling.

So, textual collections can be considered statistical system, for which one can calculate such quantities as energy, entropy and free energy. Analogous approach is widely used in pattern recognition [10].

The calculation of free energy value in TM is based on the following assertions [23]: 1) The result of TM is a matrix Φ denoting a distribution of unique words by topics, whose size is $W \cdot T$; 2) This matrix defines the total number of micro-states of a textual statistical system. Each element of the matrix corresponds to a micro-state, which is characterized by the belonging probability of each word to each topic.

Correspondingly, the energy of a micro state can be expressed as $\epsilon_{wt} = -\ln(p_{wt})$, where p_{wt} is the probability of the word w under the topic t. Let the density-of-states function be defined as $\rho(E) = \frac{N(E)}{WT}$, where $N(E)$ is the number of states with tat least energy E. The relative entropy of the system can be expressed as $S(E) = \ln(\rho(E))$, where $\rho(E)$ characterizes the relation between the initial distribution (when all elements of the ensemble have the same belonging probabilities to different topics, namely $p = \frac{1}{W}$) and the distribution obtained as the result of topic modelling.

Full internal relative energy of an ensemble of words in textual collection can be written as $E = \frac{\sum_t \sum_w p_{wt}}{T}$, and the relative free energy as a function of the number of topics can be written as:

$$\Lambda_F = F(T) - F_0 = E(T) - E_0 - (S(T) - S_0)T =$$
$$= -\ln(\frac{\sum_{t=1}^{T} \sum_{w=1}^{W} p_{wt}}{T}) - T\ln(\frac{N_1}{WT}),$$

where N_1 is the number of states satisfying $p_{wt} > 1/W$, F_0 and E_0 are the free and internal energies of the initial state, respectively, and S_0 is the Gibbs entropy of the initial state. $F(T)$, $E(T)$ and $S(T)$ are the free energy, internal energy and Gibbs entropy of the final state, respectively. In the framework of [23], the behavior of $\tilde{\Lambda}_F = \frac{F(T)-F_0}{T}$ is experimentally investigated under variation of the number of topics for Gibbs sampling model. Free energy of an ensemble of words can be expressed by Renyi entropy $S^R_{q=1/T}$ using escort distribution [9, 19]: $S^R_{q=1/T} = \frac{F}{T-1}$, $q = \frac{1}{T}$, where q is the deformation parameter. Let us note that Gibbs-Shannon entropy is a special case of Renyi entropy, namely, $S^R_q \to S^{GS}$ as $q \to 1$, where S^{GS} is Gibbs-Shannon entropy. The search of the optimal number of topics corresponds to the search of the minimum non-extensive entropy that in turn corresponds to the maximum of information [8].

The algorithm of finding the optimal number of topics consists of the four steps [21]: 1) Running a series of topic modeling procedures with different number of topics. The output is a series of matrices of words' distributions by topics. 2) Computing the value of the density-of-states function for each matrix. 3) Computing the value of Renyi entropy according to: $S^R_{q=1/T} = \frac{F}{T-1}$. 4) Finding the global minimum of Renyi entropy. This approach allows us to estimate the level of topic model entropy compared to the initial state (which corresponds to the maximum value of entropy). Since the information measure I satisfies: $I = -S$, the maximum entropy corresponds to the minimum information [8]. Thus, the searching of optimal values of parameters can be realized by searching the minimum of Renyi entropy and modified version of perplexity, which is defined below. Note that the approach described in [21] is dedicated to determining the optimal number of topics, whereas we propose in this paper to use analogous method to tune the values of hyper-parameters for LDA model and values of regularization parameters for BigARTM model.

Since perplexity is an exponential function of Gibbs-Shannon entropy, a deformed version of perplexity can be defined using the above Renyi entropy approach as: $\text{perplexity}_q = e^{S^R_q} = e^{\frac{F}{T-1}}$. Note that the deformed perplexity has the behaviour similar to behaviour of Renyi entropy (but more amplitude) and, hence, it has clear local minima analogous to local minima of Renyi entropy.

Both Renyi entropy and deformed perplexity have several advantages. Firstly, these values are expressed by the difference of free energies that means a close relation between deformed entropy (or deformed perplexity) and KL divergence. Secondly, the values of Renyi entropy and deformed perplexity do not depend on the size of dataset rather only on the parameters of topic models. Thirdly, Renyi entropy and deformed perplexity behave in a similar way for different datasets and adequately reflect the features of topic models under boundary conditions $T = 1, T \to \infty$. The difference in topic models behaviour is characterized by the location of a global and some local minima. Fourthly, both values can be used for selecting the optimal number of topics and also for choosing the values of hyper-parameters or regularization parameters by finding the minimum of Renyi entropy or deformed perplexity. In the next section of this work we demonstrate the applicability of our approach for selecting different parameters in different topic models.

5 DESCRIPTION OF DATA AND COMPUTER EXPERIMENTS.

The purpose of this part of the work is application of deformed entropy (Renyi entropy) and the deformed perplexity for the finding of optimal hyper-parameters in three topics models (pLSA, BigARTM and LDA Gibbs sampling). The experiments were conducted on three different datasets (Russian, English, and French-language datasets). The choice of regularization parameters is based on the principle of searching for a minimum of entropy / perplexity with variations of the number of topics and values of hyper-parameters. The choice of the minimum of Renyi entropy is due to the fact that the number of topics corresponding to the minimum of the entropy coincides with the human markups [21]. For this reason, two datasets (Russian and English) were chosen with a known number of topics. It means that we know in advance the number of topics and topics themselves in the collections. This allows us to evaluate the effect of the regularization procedure, and how the value of the coefficients affects the results of topic modeling.

- **Russian Dataset (lenta_ru)**: it consists of 8630 documents (containing 23297 unique words) in Russian language, each of which is manually annotated with a class among 10 topic classes. However, some of these topics are strongly correlated to each other, Human annotators had a small disagreement about topics, which really exist in dataset, therefore the number of topics can slightly vary in this dataset. As consequence, different topic models can give slightly different number of topics in the collection [21]. Therefore, the documents in this dataset can be represented by 7 to 10 topics.
- **English Dataset**: the 15404 English documents (containing 50948 unique words) composing this dataset are manually annotated with a topic class among 20 topic classes. This is a famous dataset which is called '20 Newsgroups' [32]. Related work [7] argued that 14 to 20 topics are suitable to represent the documents of this dataset.
- **French Dataset**: it contains 25000 documents in French language, where 18749 words appear in the whole dataset. We use this dataset to show that the behavior of entropy and deformed perplexity is similar to the correspondent behaviors for Russian and English datasets. In this case, using the formalism of the entropy approach, it is possible to evaluate the optimal number of topics in a dataset with a previously unknown distribution of topics.

The effectiveness of Renyi entropy to indicate the evaluation of optimal hyper-parameters values and number of topics is tested with detailed experiments, which are conducted on the above mentioned datasets. For more general evaluation, three different models, namely pLSA, BigARTM and LDA Gibbs sampling, are applied on all datasets. In addition to the wide employment of these models in TM, another reason to consider them in this work is their different characteristics. In other words, pLSA is a basic model, which is parametrized with only the number of topics, whereas BigARTM and LDA Gibbs sampling models are regularized versions of pLSA model with different regularization options. Accordingly, this work demonstrates how to select the optimal parameters of regularization using Renyi entropy and deformed perplexity.

5.1 Discussion of pLSA model for three datasets.

In this evaluation, pLSA model is applied on the three datasets to study the behaviour of Renyi entropy and deformed perplexity in dependence on number of topics. The details of the simulation on each dataset is described in the following:

1) The Russian dataset (lenta_ru): Figure 1 shows the behaviour of Renyi entropy as a function of the number of topics on the Russian dataset. The minimum of Renyi entropy is marked by symbol 'x' on the figure. Further we also mark the minimum of Renyi entropy by this symbol on corresponding figures. As demonstrated in Figure 1, the minimum of Renyi entropy indicates the optimal number of topics which is close to human coding.

2) English Dataset: According to [7], the documents in this dataset can be well represented by a number of topics between 14 and 20. Figure 2 presents Renyi entropy curve in terms of number of topics. Despite the fact that in the region of minimum entropy and deformed perplexity there are small fluctuations associated with the instability of topic models, the minimum of both functions is easily determined, and this minimum corresponds to 15 topics.

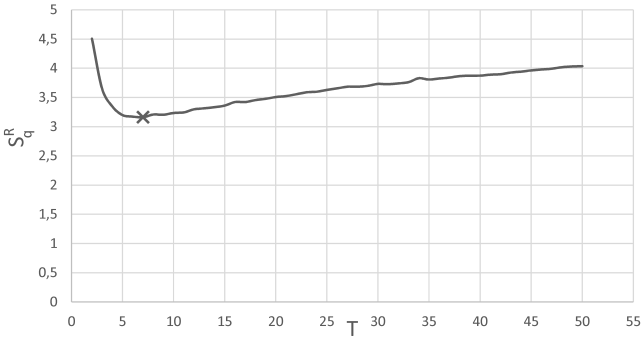

Figure 1: Dependence of Renyi entropy on the number of topics (pLSA on Russian dataset).

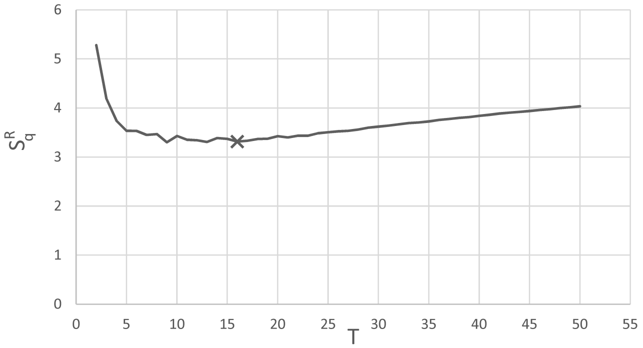

Figure 2: Dependence of Renyi entropy on the number of topics (pLSA on English dataset).

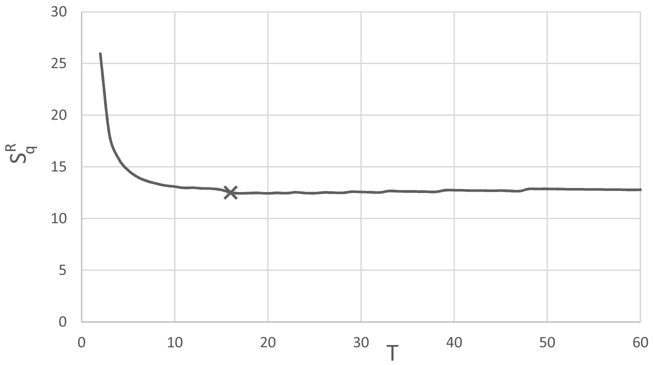

Figure 3: Dependence of Renyi entropy on the number of topics (pLSA on French dataset).

3) French dataset: Similarly to the Russian and English datasets, pLSA is applied on the French dataset. As shown in Fig. 3 Renyi entropy is relatively stable with different number of topics compared to the two other datasets. This can be explained with the high number of words per document which allows different clustering result. However, the obtained result indicates that the optimal number of topic corresponds to 17, which is approximately similar to the number of topics in news papers.

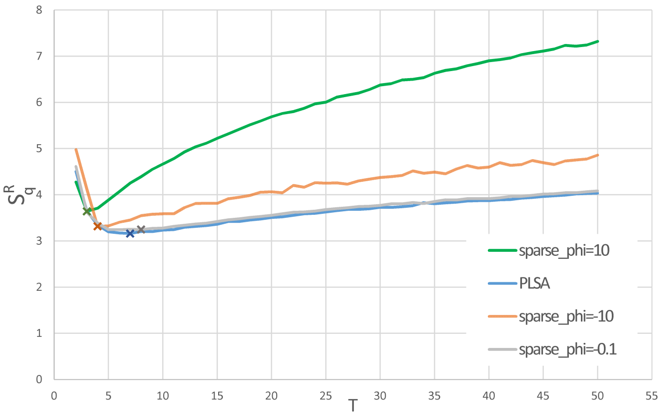

Figure 4: Dependence of Renyi entropy on the number of topics for different values of "sparse phi" regularizer (Russian dataset)

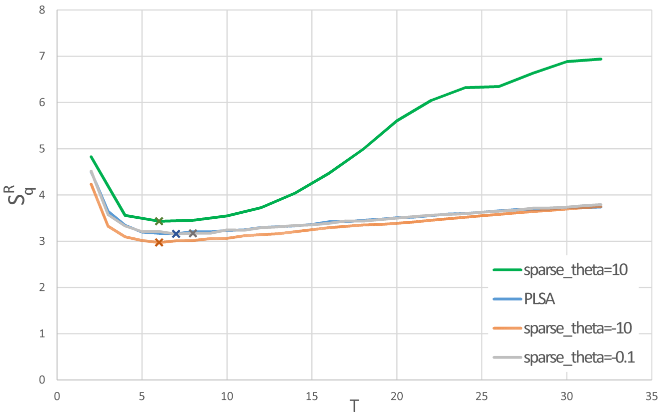

Figure 5: Dependence of Renyi entropy on the number of topics for different values of "sparse theta" regularizer (Russian dataset)

5.2 BigARTM model

In this model, we study the dependence of the optimal number of topics on different values of regularization coefficients. We consider two types of regularization, which are called "Sparse phi regularizer" and "Sparse theta regularizer". Each of the regularizers is characterized by a regularization coefficient, which was varied in the range [-10, 10] in our numerical simulations. So, we study the effect of the influence of two types of regularization on the results of topic modelling.

1) Russian-language dataset (lenta_ru).

1.1) "Sparse phi regularizer".

Since we know the evaluation of the number of topics in a given dataset by human markup, a shift of the minimum to the area of 1 or 2 topics means a strong over-regularization of the model. Thus, this regularization model is not suitable for using at large values of the regularization coefficient, but at small value it coincides with the plsa model.

1.2) "Sparse theta regularizer".

The distribution of Reny entropy minima under variation of regularization coefficient is given in Fig. 5. Here one can see that the values of the regularizer do not influence significantly to the optimal number of topics. So, the results are be more stable with "sparse theta regularizer" comparing to "sparse phi" regularization.

2) English-language dataset (20 Newsgroups).

2.1) "Sparse phi regularizer".

The change in the value of this regularizer leads to a significant distortion of the location of Renyi entropy minimum for English language dataset (Fig. 6). Shift of the minimum to the area of 4 or 5 topics means a strong over-regularization of the model. So, value of "sparse_phi" should rather be -0.1 than 10 or -10.

2.2) "Sparse theta regularizer".

Numerical results are represented in Fig. 7. Here one can also see the shift of the optimal number of topics under variation regularization coefficient.

3) French dataset

As shown in Fig. 8, changes in the value of 'Sparse phi" regularizer lead to a shift of the minimum Renyi entropy and the optimal number of topics. Variation of "Sparse theta" regularizer does not influence significantly on the optimal number of topics (Fig. 9).

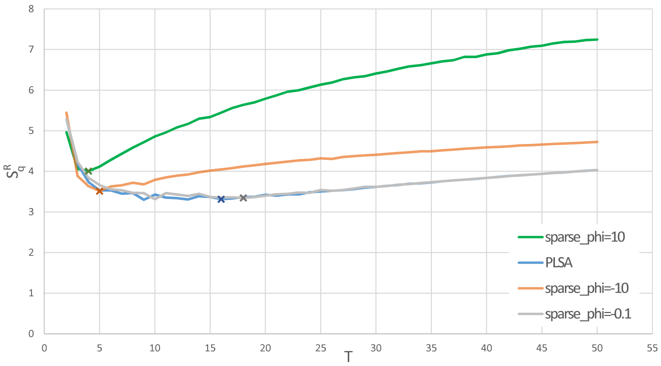

Figure 6: Dependence of Renyi entropy on the number of topics for different values of "sparse phi" regularizer (English dataset)

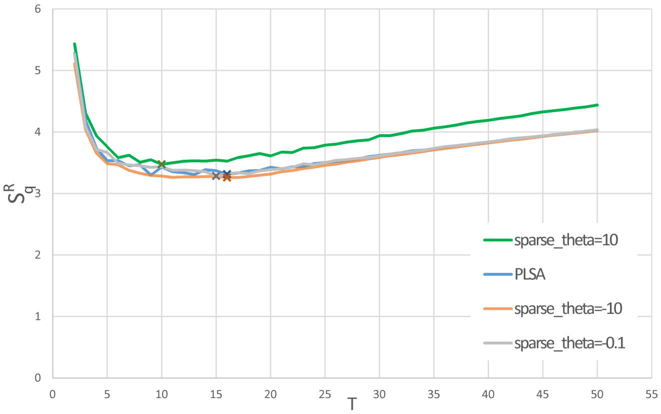

Figure 7: Dependence of Renyi entropy on the number of topics for different values of "sparse theta" regularizer (English dataset)

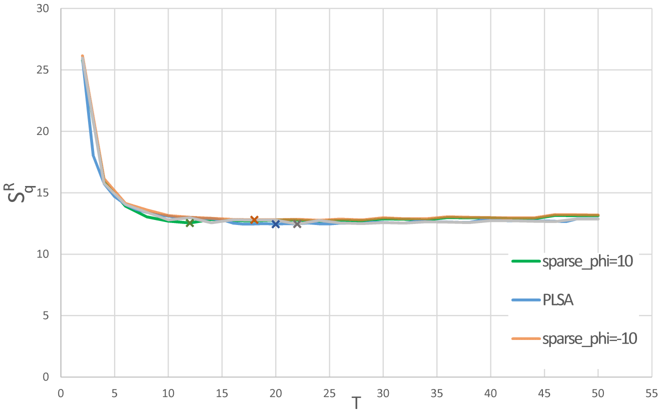

Figure 8: Dependence of Renyi entropy on the number of topics for different values of "sparse phi" regularizer (French dataset)

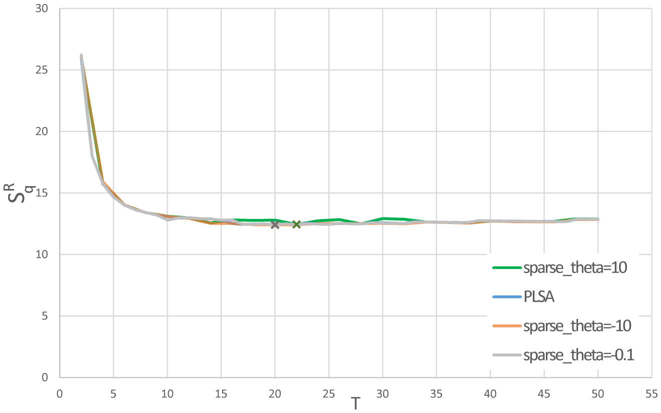

Figure 9: Dependence of Renyi entropy on the number of topics for different values of "sparse theta" regularizer (French dataset)

Application of Renyi entropy allows to see the tendency of regularization procedure for a topic model. Topic models possess fluctuations that can lead to unreliable jumps [22]. That is why we recommend to run topic modeling with the same values of parameters at least three times and then to average the result [22].

5.3 LDA Gibbs sampling model

Topic model based on Gibbs sampling procedure has three parameters: number of topics and two other parameters of Dirichlet distribution (α, β) that need to be selected. Within the framework of this paper, the number of topics was varied in the range [2, 50] and hyper-parameters were varied in the range [0.1, 1]. Renyi entropy was calculated for each model.

1) Russian dataset (lenta_ru).

Varying the parameters α, β and the number of topics leads to the appearance of a set of local entropy minima, that corresponds to the fact that the values of Renyi entropy minima and maxima depend on the number of topics and on the values of hyper-parameters as well. Figure 10 shows the dependence of Renyi entropy on the number of topics under variation of hyper-parameters. Small jumps on this figure are due to the fact that this figure is plotted for 100 different values of hyper-parameters simultaneously and due to fluctuations of TM. Despite these jumps one can see the common tendency of Renyi entropy curve for different values of hyper-parameters. Figure 11 shows the value of fluctuation in percent as a function of the number of topics for the given dataset. This plot shows that the variation of hyper-parameters leads to small value of fluctuation (about 4-8 % with respect to the average value) and does not significantly influence on the evaluation of the optimal number of topics.

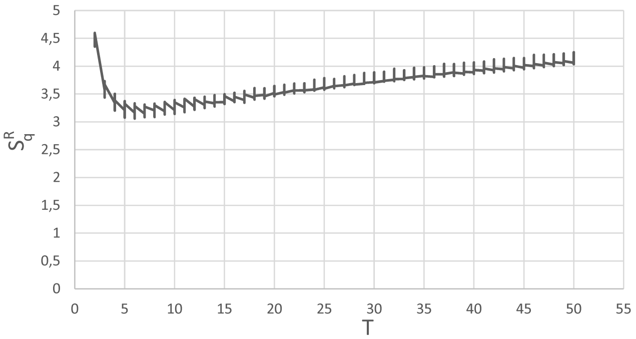

Figure 10: Dependence of Renyi entropy on the number of topics under variation of hyper-parameters for Russian dataset.

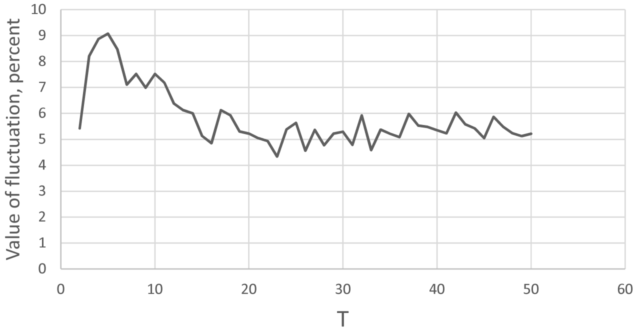

Figure 11: Fluctuation of the difference between the maximum and minimum of Renyi entropy values as a function of topic number for the Russian-language dataset.

Moreover, Fig. 12 illustrates the average value of Renyi entropy, calculated across Renyi entropy values for all values of hyper-parameters α and β for fixed number of topics. The plot shows that, despite the fluctuation, minimum of entropy corresponds to 6-7 topics that coincides with the results of pLSA model for this dataset. Thus, it can be considered that the hyper-parameters α and β do not play a big role for this dataset. Hence, one can take any value of hyper-parameters in the range [0.1, 1] and it will not influence significantly to the results of TM.

2) English dataset (20 Newsgroups). The variations in the parameters α, β and the number of topics for the English dataset also show the existence of a set of local minima of Renyi entropy (Fig. 13), however, the level of fluctuations is about 4-11% of the average Renyi entropy (Fig. 14) that is not essential. The minimum of the average Renyi entropy can be seen in Fig. 15, the common tendency

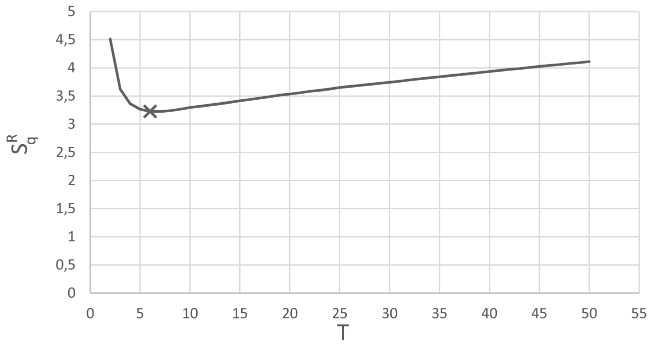

Figure 12: The average value of Renyi entropy as a function of the number of topics for the Russian-language dataset.

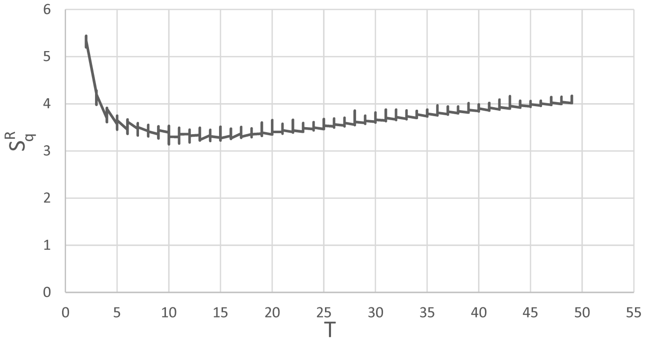

Figure 13: Dependence of Renyi entropy on the number of topics under variation of hyper-parameters for English dataset.

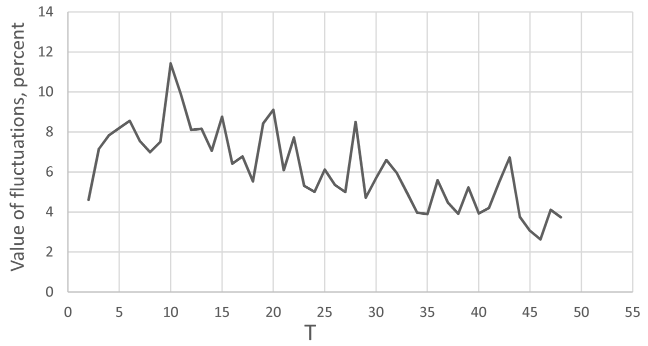

Figure 14: Fluctuation of the difference between the maximum and minimum of Renyi entropy values as a function of topic number for the English dataset.

can be seen in Fig. 13 . These figures show that the optimal number of topics for this dataset is about 14-15 topics, that coincides with the result of pLSA model as well.

3) French dataset. The behavior of Renyi entropy in dependence on the number of topics and values of hyper-parameters is shown in Fig. 16. The variations in the parameters α, β and the number of topics also show the existence of a set of local minima of Renyi entropy, the level of fluctuations is about 3-9% of the average Renyi entropy (Fig. 18). Fig. 17 shows that the optimal number of topics for this dataset is about 16 topics.

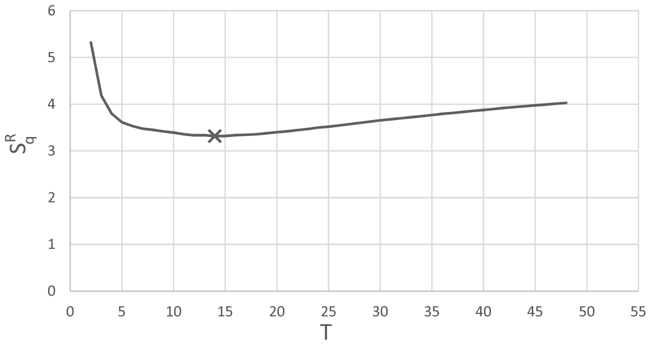

Figure 15: The average value of Renyi entropy as a function of the number of topics for the English dataset.

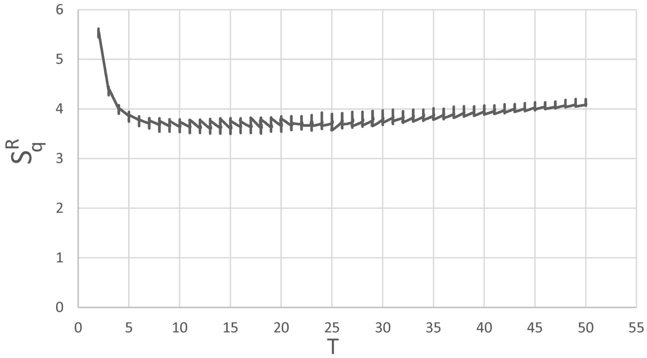

Figure 16: Dependence of Renyi entropy on the number of topics under variation of hyper-parameters for French dataset

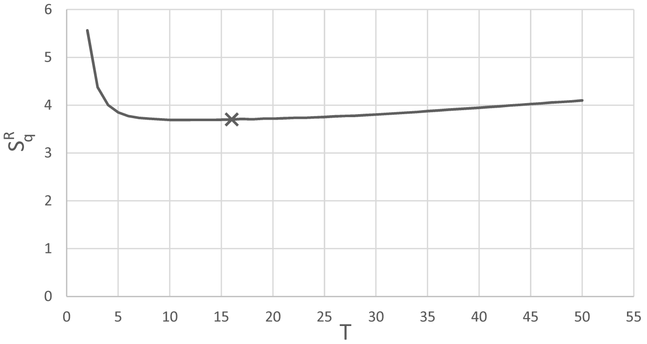

Figure 17: The average value of Renyi entropy as a function of the number of topics for the French dataset.

So, for LDA Gibbs sampling model we found out that the values of hyper-parameters do not lead to significant changes in the results of TM with respect to the optimal number of topics according to Renyi entropy approach.

6 CONCLUSION.

The three topic models, considered in this work, employ different variants of regularization, where regularization parameters play

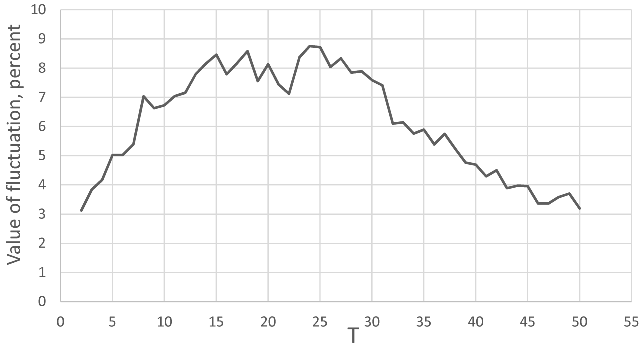

Figure 18: Fluctuation of the difference between the maximum and minimum of Renyi entropy values as a function of topic number for the French dataset.

different roles. However, the behavior of Renyi entropy and the deformed perplexity is similar for different models and datasets in different languages. At small values of the number of topics (1,2), the partitioning of datasets is of little avail since entropy is extremely large (correspondingly, the information value is almost zero). Further, increase of the number of topics leads to decrease of entropy. At some point entropy reaches the global minimum, but then, as the number of topics increases further, the entropy tends to the maximum, since in this case the overall distribution of words tends to flat distribution, which is known to have a large value of entropy (and a small value of information). Let us note that each of the datasets has a different number of topics corresponding to the minimum of entropy (maximum of information), while these minima coincide with human markups. The behavior of deformed perplexity is similar to the behavior of Renyi entropy.

In the additive regularization model, higher values of regularization parameters lead to larger deviation of the entropy minimum from the human markup. Therefore, large values of regularization parameters cannot be used in BigARTM model. When parameters in LDA Gibbs sampling model are varied, a set of local minima of Renyi entropy (and deformed perplexity) emerges, however, the average fluctuation is about 4-8%, while the minimum of average Renyi entropy coincides with the human markup. It can be concluded that the contribution of the change in the number of topics is more significant than the contribution of the change in hyper-parameters α and β to the results of topic modeling. Therefore, when conducting topic modeling it is sufficient to fix hyper-parameters, for example, by applyingStyers' and Griffith's empirical rule [15], where the parameters are inversely proportional to the number of topics, or simply use hyperparameters as constants, for example, $\alpha = \beta = 0.1$.

Since BigARTM and LDA Gibbs sampling models are different versions of regularization compared to pLSA model, the comparison of Renyi entropy behavior allows us to suggest that some types of regularization significantly distort the true distribution of topics present in the datasets. For instance, the idea of regularization based on the use of conjugate functions (Dirichlet and multinomial distribution) turns out to be successful and, as our calculations show, such models allow for the "true" numbers of topics to be found.

Our research also shows that application of Renyi entropy and deformed perplexity allows to construct an effective strategy to find and select an appropriate regularization model and to formulate criteria for the selection of the values of regularization coefficients. In the future, the calculation of Renyi entropy or deformed perplexity can be embedded in the algorithms of topic modeling, that will significantly simplify its use in various areas of machine learning.

REFERENCES

[1] E. Akturk, G. B. Bagci, and R. Sever. 2007. Is Sharma-Mittal entropy really a step beyond Tsallis and Renyi entropies? (2007). arXiv:arXiv:cond-mat/0703277

[2] David Andrzejewski and Xiaojin Zhu. 2009. Latent Dirichlet Allocation with Topic-in-set Knowledge. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing (SemiSupLearn '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 43–48.

[3] Murat Apishev, Sergei Koltcov, Olessia Koltsova, Sergey I. Nikolenko, and Konstantin Vorontsov. 2016. Mining Ethnic Content Online with Additively Regularized Topic Models. *Computación y Sistemas* 20, 3 (2016), 387–403.

[4] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On Smoothing and Inference for Topic Models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*. AUAI Press, Arlington, Virginia, United States, 27–34.

[5] Ramnath Balasubramanyan, Bhavana Dalvi, and William W. Cohen. 2013. From Topic Models to Semi-supervised Learning: Biasing Mixed-Membership Models to Exploit Topic-Indicative Features in Entity Clustering. In *Proceedings, Part II, of the European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 8189 (ECML PKDD 2013)*. Springer-Verlag New York, Inc., New York, NY, USA, 628–642.

[6] A.G Bashkirov. 2004. On maximum entropy principle, superstatistics, power-law distribution and Renyi parameter. *Physica A: Statistical Mechanics and its Applications* 340 (2004), 153–162.

[7] S. Basu, I. Davidson, and K. Wagstaff (Eds.). 2008. *Constrained clustering : advances in algorithms, theory, and applications* (1st. ed.). Taylor & Francis Group Boca Raton.

[8] C. Beck. 2009. Generalised information and entropy measures in physics. *Contemporary Physics* 50, 4 (2009), 495–510.

[9] Christian Beck and Friedrich SchÃűgl. 1995. *Thermodynamics of Chaotic Systems an Introduction*. Cambridge University Press.

[10] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022.

[12] BUGRA 2014. Entropy and Perplexity on Image and Text. http://bugra.github.io/work/notes/2014-05-16/entropy-perplexity-image-text/

[13] A De Waal and E. Barnard. 2008. Evaluating topic models with stability. In *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*. PRASA, 79–84.

[14] Joshua T. Goodman. 2001. A Bit of Progress in Language Modeling. *Comput. Speech Lang.* 15, 4 (Oct. 2001), 403–434.

[15] T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, Suppl. 1 (April 2004), 5228–5235.

[16] Gregor Heinrich. 2004. *Parameter estimation for text analysis*. Technical Report.

[17] Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*. ACM, New York, NY, USA, 50–57.

[18] Thomas Hofmann. 2001. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Mach. Learn.* 42, 1/2 (Jan. 2001), 177–196.

[19] Yu L Klimontovich. 1989. Problems in the statistical theory of open systems: Criteria for the relative degree of order in self-organization processes. *Soviet Physics Uspekhi* 32, 5 (1989), 416.

[20] Denis Kochedykov, Murat Apishev, Lev Golitsyn, and Konstantin Vorontsov. 2017. Fast and Modular Regularized Topic Modelling. In *Proceedings of the 21st Conference of Open Innovations Association FRUCT (FRUCT'21)*. FRUCT Oy, Helsinki, Finland, Finland, 182–193.

[21] Sergei Koltcov. 2018. Application of Renyi and Tsallis entropies to topic modeling optimization. *Physica A: Statistical Mechanics and its Applications* 512 (2018), 1192 – 1204.

[22] Sergei Koltcov, Olessia Koltsova, and Sergey Nikolenko. 2014. Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. ACM, New York, NY, USA, 161–165.

[23] S. N. Koltcov. 2017. A thermodynamic approach to selecting a number of clusters based on topic modeling. *Technical Physics Letters* 43, 6 (01 Jun 2017), 584–586.

[24] Sergei Koltsov, Sergey Nikolenko, Olessia Koltsova, Vladimir Filippov, and Svetlana Bodrunova. 2016. Stable Topic Modeling with Local Density Regularization. In *Internet Science: Third International Conference*, Vol. 9934. Springer International Publishing, 176–188.

[25] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *Ann. Math. Statist.* 22, 1 (03 1951), 79–86.

[26] Daniel Maier, Annie Waldherr, P Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, U Reber, Tom HÃďussler, Hannah Schmid-Petri, and Silke Adam. 2018. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. (02 2018), 1–26.

[27] Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge, MA, USA.

[28] Thomas Minka. 2000. Estimating a Dirichlet Distribution.

[29] Marwa Naili, Anja Habacha Chaibi, and Henda Ben Ghézala. 2017. Arabic topic identification based on empirical studies of topic models. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées* Volume 27 - 2017 - Special issue CARI 2016 (Aug. 2017).

[30] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed Algorithms for Topic Models. *J. Mach. Learn. Res.* 10 (Dec. 2009), 1801–1828.

[31] David Newman, Edwin V Bonilla, and Wray Lindsay Buntine. 2011. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011 (NIPS 2011): Volume 1 of 3*, John Shawe-Taylor, Richard Zemel, Peter Bartlett, and Fernando Pereira (Eds.). Neural Information Processing Systems (NIPS), 1 – 9.

[32] Newsgroups 2008. 20 Newsgroups. http://qwone.com/~jason/20Newsgroups/

[33] Kenneth Rose, Eitan Gurewitz, and Geoffrey C. Fox. 1990. Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett.* 65 (Aug 1990), 945–948. Issue 8.

[34] Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. 2010. Learning Author-topic Models from Text Corpora. *ACM Trans. Inf. Syst.* 28, 1, Article 4 (Jan. 2010), 38 pages.

[35] Mark Steyvers and Tom Griffiths. 2007. *Probabilistic Topic Models.* Lawrence Erlbaum Associates.

[36] Andrey N. Tikhonov and Vasiliy Y. Arsenin. 1977. *Solutions of ill-posed problems.* V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York.

[37] Konstantin Vorontsov and Anna Potapenko. 2014. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. In *Analysis of Images, Social Networks and Texts (Communications in Computer and Information Science).* Springer International Publishing.

[38] K. V. Vorontsov. 2014. Additive regularization for topic models of text collections. *Doklady Mathematics* 89, 3 (May 2014), 301–304.

[39] Hanna M. Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why Priors Matter. In *Proceedings of the 22Nd International Conference on Neural Information Processing Systems (NIPS'09).* Curran Associates Inc., USA, 1973–1981.

[40] Weizhong Zhao, James J Chen, Roger Perkins, Zhichao Liu, Weigong Ge, Yijun Ding, and Wen. Zou. 2015. A heuristic approach to determine an appropriate number of topics in topic modeling. In *Proceedings of the 12th Annual MCBIOS Conference.* BioMed Central Ltd.