ECEA 2019

5th International Electronic Conference on Entropy and Its Applications

18–30 November 2019

Chaired by Prof. Geert Verdoolaege

Sponsored by: entropy MDPI

Fast tuning of topic models: an application of Rényi entropy and renormalization theory

Sergei Koltcov [1,*], Vera Ignatenko [1] and Sergei Pashakhin[1]

[1] Internet Studies Lab, National Research University Higher School of Economics, 55/2 Sedova St., St. Petersburg, Russia, 192148.

* Corresponding author: skoltsov@hse.ru

NATIONAL RESEARCH UNIVERSITY

INTERNET STUDIES LAB

Abstract

- In practice, the critical step in build machine learning models of big data (BD) involves costly in terms of time and computing resources procedure of parameter tuning with grid search.
- We have shown that topic modeling (a clustering method for large document collections) demonstrates self-similar behavior under the condition of a varying number of clusters. Such behavior allows using a renormalization technique.
- A combination of renormalization procedure with Rényi entropy approach allows for fast searching of the optimal number of clusters.
- In this work, the renormalization procedure is developed for the Latent Dirichlet Allocation (LDA) model with variational Expectation-Maximization algorithm.
- The numerical experiments were conducted on two document collections with a known number of clusters in two languages.
- This work presents results for three versions of the renormalization procedure: (1) a renormalization with the random merging of clusters, (2) a renormalization based on minimal values of Kullback-Leibler divergence and (3) a renormalization with merging clusters with minimal values of Rényi entropy.
- Our work shows that the renormalization procedure allows finding the optimal number of topics 26 times faster than grid search without significant loss of quality.

Topic modeling

Topic modeling is based on the assumption that a document collection has a finite set of word distributions. Each such word distribution could be called a 'topic' or a thematical cluster. The probability of encountering a word w in a given document d is expressed as follows:

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \phi_{wt}\theta_{td},$$

where t is a topic, ϕ_{wt} constitute matrix Φ (the distribution of words by topics), and θ_{td} form matrix Θ (the distribution of documents by topics).

We consider the Blei model [Blei, D.M.; Ng, A.Y.; Jordan, M.I.; 2003] of Latent Dirichlet Allocation with variational Expectation-Maximization algorithm, where the distribution of topics by documents is assumed to be Dirichlet distribution with T-dimensional parameter α (T is the number of topics).

An example of matrix Φ (the distribution of words by topics)

Words with high probability

	1	2	3	4	5	6	7	8	9
1	turkish: 0,011209	israel: 0,014142	space: 0,020407	windows: 0,014858	please: 0,015528	god: 0,020857	file: 0,016943	which: 0,007806	car: 0,018218
2	armenian: 0,010977	jews: 0,009726	nasa: 0,011303	dos: 0,011373	mail: 0,015515	his: 0,010509	image: 0,011841	their: 0,007559	article: 0,0080
3	armenians: 0,008735	israeli: 0,008491	gov: 0,006948	drive: 0,009181	me: 0,014473	who: 0,009749	files: 0,010098	government: 0,006550	cars: 0,00659
4	were: 0,008166	who: 0,007504	earth: 0,006274	card: 0,008830	e: 0,011283	jesus: 0,009381	jpeg: 0,008802	may: 0,005726	any: 0,00499
5	their: 0,008119	article: 0,007339	henry: 0,005341	mac: 0,007855	ca: 0,010973	bible: 0,006158	windows: 0,008260	other: 0,005012	engine: 0,004
6	armenia: 0,006552	jewish: 0,006495	launch: 0,004740	apple: 0,007013	thanks: 0,010650	church: 0,005834	color: 0,007459	people: 0,004895	out: 0,004861
7	people: 0,006493	were: 0,005950	orbit: 0,004283	system: 0,006904	any: 0,008069	christ: 0,005742	gif: 0,007035	states: 0,004573	new: 0,00421
8	turkey: 0,005936	arab: 0,005898	shuttle: 0,004232	use: 0,005862	am: 0,007969	christian: 0,005668	bit: 0,006634	such: 0,004483	get: 0,003651
9	turks: 0,005750	their: 0,005445	moon: 0,004221	problem: 0,005720	anyone: 0,007684	which: 0,005123	program: 0,005892	its: 0,004243	also: 0,00362
10	cramer: 0,005588	war: 0,005075	mission: 0,003879	scsi: 0,005654	email: 0,007125	him: 0,004921	format: 0,005868	new: 0,004154	speed: 0,003
11	men: 0,005135	peace: 0,004540	solar: 0,003236	does: 0,005479	university: 0,006294	christians: 0,004498	some: 0,005762	state: 0,004147	oil: 0,003528
12	article: 0,004264	its: 0,003932	toronto: 0,003226	pc: 0,005387	know: 0,005499	us: 0,004302	images: 0,005762	these: 0,003989	when: 0,0034
13	genocide: 0,004252	people: 0,003819	which: 0,003215	any: 0,005379	send: 0,005413	our: 0,004241	use: 0,005609	also: 0,003797	had: 0,00341
14	p: 0,004217	muslims: 0,003747	pat: 0,003195	disk: 0,005120	address: 0,005226	their: 0,004032	any: 0,005480	national: 0,003735	just: 0,003327
15	had: 0,004147	which: 0,003634	its: 0,003081	video: 0,005003	interested: 0,005189	paul: 0,003959	version: 0,004419	been: 0,003563	up: 0,003282
16	been: 0,004147	arabs: 0,003613	also: 0,003050	memory: 0,004828	list: 0,005177	sin: 0,003940	than: 0,004172	our: 0,003351	ford: 0,00321
17	who: 0,004147	only: 0,003521	more: 0,002967	software: 0,004403	sale: 0,004991	were: 0,003757	don: 0,003724	public: 0,003330	than: 0,00317
18	soviet: 0,004055	any: 0,003377	satellite: 0,002925	using: 0,004361	fax: 0,004643	lord: 0,003738	get: 0,003724	were: 0,003296	good: 0,00311
19	history: 0,003962	them: 0,003202	system: 0,002842	get: 0,004236	internet: 0,004519	man: 0,003511	which: 0,003583	year: 0,003241	drive: 0,00310
20	university: 0,003904	sandvik: 0,003119	into: 0,002780	monitor: 0,004161	d: 0,004407	love: 0,003511	quality: 0,003536	right: 0,003186	like: 0,002991
21	gay: 0,003787	policy: 0,003119	sky: 0,002759	mouse: 0,003803	net: 0,004246	people: 0,003481	does: 0,003477	united: 0,003172	dealer: 0,002
22	new: 0,003648	world: 0,002965	spacecraft: 0,002728	work: 0,003769	new: 0,004221	only: 0,003462	graphics: 0,003394	american: 0,003172	price: 0,00271
23	greek: 0,003555	state: 0,002955	new: 0,002676	which: 0,003753	apr: 0,004196	also: 0,003285	also: 0,003335	well: 0,003110	very: 0,00271
24	serdar: 0,003485	when: 0,002800	first: 0,002666	driver: 0,003753	info: 0,003861	faith: 0,003217	display: 0,003229	more: 0,003021	much: 0,0026
25	argic: 0,003474	his: 0,002780	high: 0,002593	when: 0,003744	mark: 0,003861	believe: 0,003187	software: 0,003182	two: 0,003007	front: 0,00265

An example of matrix Θ (the distribution of documents by topics)

Documents with high probability

	1	2	3	4	5	6	7	8	9	10	11
1	4445: 0,947937	3678: 0,909014	1735: 0,911512	5607: 0,826132	14286: 0,767241	5230: 0,945484	11157: 0,995329	4604: 0,884863	13706: 0,838942	8383: 0,911319	12352: 0,7151
2	3600: 0,935178	4474: 0,864130	1500: 0,890000	11325: 0,802752	2735: 0,745614	4834: 0,904255	11201: 0,992803	8282: 0,790441	15161: 0,801020	8578: 0,907226	5617: 0,70979
3	4051: 0,934798	3699: 0,830153	1845: 0,860224	1095: 0,787736	3273: 0,726190	9613: 0,866795	10425: 0,990019	8594: 0,754167	958: 0,790076	8984: 0,897832	5806: 0,694631
4	3877: 0,929806	3543: 0,829609	7112: 0,851889	5966: 0,771028	14709: 0,717262	5129: 0,864833	4376: 0,808943	14477: 0,753968	14161: 0,787938	8599: 0,888172	5569: 0,68750
5	4624: 0,921500	4475: 0,822581	579: 0,829373	809: 0,769608	5298: 0,711538	9641: 0,859694	4205: 0,806905	15107: 0,746894	13329: 0,772109	9003: 0,850495	5539: 0,686441
6	4648: 0,919545	4560: 0,820675	1531: 0,810078	11714: 0,763566	4346: 0,676471	9989: 0,849826	5508: 0,781553	8948: 0,742623	825: 0,771845	8291: 0,840603	8427: 0,68243
7	4050: 0,913194	4533: 0,815789	1288: 0,809278	11368: 0,759146	12405: 0,671875	4850: 0,845865	7073: 0,764957	8278: 0,710253	10429: 0,768072	9017: 0,823612	12103: 0,66891
8	4449: 0,906816	4466: 0,814685	1318: 0,808094	12255: 0,757843	3218: 0,656863	9894: 0,845679	10875: 0,755435	14063: 0,702649	7040: 0,758278	8289: 0,780303	5235: 0,64084
9	4583: 0,898798	3397: 0,802174	1154: 0,787726	594: 0,755814	4319: 0,628788	10075: 0,836890	4387: 0,753947	14230: 0,696991	2325: 0,755144	14639: 0,664706	6077: 0,63043
10	4598: 0,895711	4465: 0,795238	1577: 0,786122	860: 0,745989	15004: 0,621396	10001: 0,836883	10903: 0,744275	14264: 0,696311	2647: 0,747059	8582: 0,644578	12673: 0,6234
11	3881: 0,893411	4522: 0,792576	1866: 0,778128	12494: 0,744505	15285: 0,615118	9895: 0,830909	12739: 0,739224	7544: 0,691915	7051: 0,746154	8992: 0,622917	8488: 0,62152
12	4545: 0,886643	3778: 0,791322	657: 0,766393	11974: 0,740000	3676: 0,608333	9658: 0,830590	5621: 0,729592	13922: 0,686141	13717: 0,744737	8460: 0,609743	12629: 0,6145
13	4502: 0,885827	4011: 0,790503	1567: 0,762443	11955: 0,734043	820: 0,603659	9971: 0,826923	10981: 0,728916	8776: 0,683140	13196: 0,740964	8716: 0,601266	5114: 0,60365
14	8621: 0,880488	4504: 0,788690	1155: 0,761696	4669: 0,733918	11840: 0,602564	9960: 0,820313	10739: 0,726804	11422: 0,675223	3414: 0,738550	8580: 0,595261	12193: 0,60091
15	8591: 0,879870	3900: 0,782927	1931: 0,755708	843: 0,727778	4315: 0,597561	9956: 0,819372	10726: 0,724604	14353: 0,660000	13839: 0,738095	8386: 0,592437	4933: 0,597561
16	3937: 0,879132	3706: 0,781250	1195: 0,752427	866: 0,726293	4317: 0,594828	4957: 0,818452	11205: 0,723545	9034: 0,653966	14884: 0,737805	8292: 0,577586	6571: 0,59558
17	4597: 0,878702	3953: 0,778409	1854: 0,750000	838: 0,723776	343: 0,594340	9465: 0,817460	10547: 0,721014	13923: 0,636842	14384: 0,737179	8310: 0,573089	1150: 0,592391

Renyi entropy approach

The entropy approach to topic modeling (TM) tuning is based on computing Rényi entropy for each topic solution while varying the number of topics and hyperparameters [Koltcov, S.; 2018], [Koltsov S.; Ignatenko V.; Koltsova O;2019]. For TM, the Rényi entropy is expressed as follows:

$$S_q^R = \frac{\ln(Z_q)}{q-1} = \frac{q\ln(q\,\tilde{P}) + q^{-1}\ln(\tilde{\rho})}{q-1}$$

where *q=1/T*, *T* is the number of clusters or topics, $\tilde{\rho} = \frac{N}{WT}$ is the density-of-states function, *W* is the number of unique words in the dataset, N is the number of words with high probability (i.e. with ϕ_{wt}>1/W),

$\tilde{P} = \frac{1}{T}\sum_{wt}\phi_{wt}\,\mathbb{1}_{\{\phi_{wt}-\frac{1}{W}\}}$is the sum of probabilities of all words with high probability,

$\mathbb{1}_{\{x-y\}} = 1$ if $x \geq y$ and $\mathbb{1}_{\{x-y\}} = 0$ if $x < y$.

It has been showed that the **minimum point** of Rényi entropy corresponds to the number of topics identified by human coders [Koltcov, S.; 2018]. Hence, the search for the S_q^R minimum could substitute at least partly manual labor of marking up document collections, substantially simplifying TM tuning on uncoded datasets.

Renormalization

The algorithm of renormalization consists of the following steps:

1. We choose a pair of topics for merging according to one of the three possible criteria (they will be discussed further). Let us denote the chosen topics by $t_1$and t_2.
2. We merge the chosen topics. The word distribution of a 'new' topic resulted from merging of $t_1$and t_2 is stored in column $\phi_{\cdot t_1}$of matrix Φ:
$$\phi_{wt_1} \coloneqq \phi_{wt_1} \exp\big(\psi(\alpha t_1)\big) + \phi_{wt_2} \exp\big(\psi(\alpha t_2)\big),$$
where ψ is a digamma function. Then, we normalize the obtained column $\phi_{\cdot t_1}$ so that $\sum_t \phi_{wt_1} = 1$ and recalculate $\alpha_{t_1} \coloneqq \alpha_{t_1} + \alpha_{t_2}$. Then, we delete column $\phi_{\cdot t_2}$ from matrix Φ and element $\alpha_{t_2}$from vector α. Further, vector α is normalized so that $\sum_t \alpha_t = 1$. We have *T−1* topics at the end of this step.
3. Since a new topic solution (matrix Φ) is formed in the previous step, we recalculate the global Rényi entropy for this solution.

Renormalization

Criteria of merging:

1. Merging of similar topics, where the similarity is estimated with symmetric Kullback-Leibler (Jensen-Shannon) divergence, and the topic pair is chosen based on the minimal value computed pairwisely.
2. Merging based on the minimum of Rényi entropy, where the topics with local minima values are summed together. Here local Rényi entropy is computed for a single topic.
3. Merging randomly selected columns.

Datasets

- Dataset in Russian (Lenta.ru). This dataset contains news articles in the Russian language where each news item was manually assigned to one of ten topic classes by the dataset provider [https://www.kaggle.com/yutkin/corpus-of-russian-news-articles-from-lenta]. However, as some of these topics could be considered folded or correlated, this dataset could be represented by 7--10 topics. We considered a class-balanced subset of this dataset, which consisted of 8,624 news texts (containing 23,297 unique words).
- Dataset in English (20 Newsgroups dataset [http://qwone.com/~jason/20Newsgroups/]). This well-known dataset contains articles assigned by users to one of 20 newsgroups. Since some of these topics can be unified, this document collection can be represented by 14-20 topics [Basu, S.; Davidson, I.; Wagstaff, K. , 2008]. The dataset is composed of 15,404 documents with 50,948 unique words.

Numerical experiments

Results for the dataset in Russian.

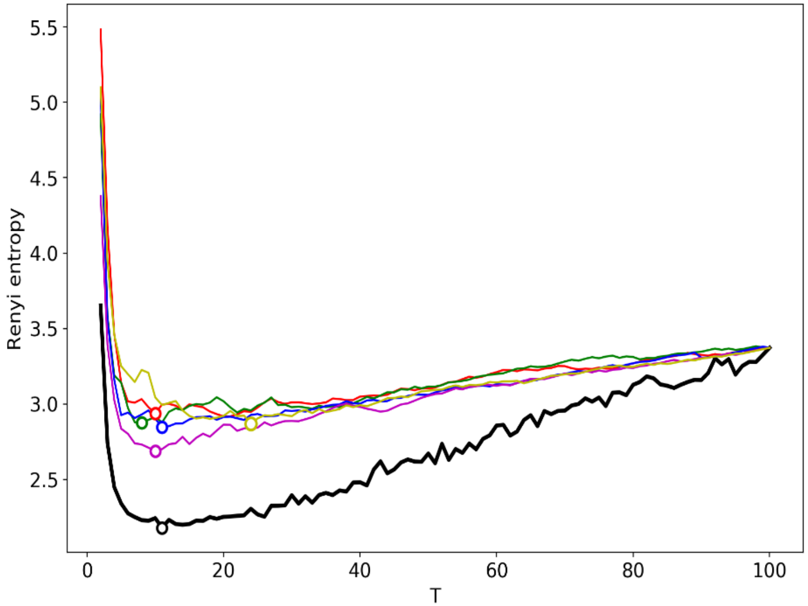

Fig. 1. Rényi entropy curves. Black: successive topic modeling. Other colors: renormalization with the random merging of topics.

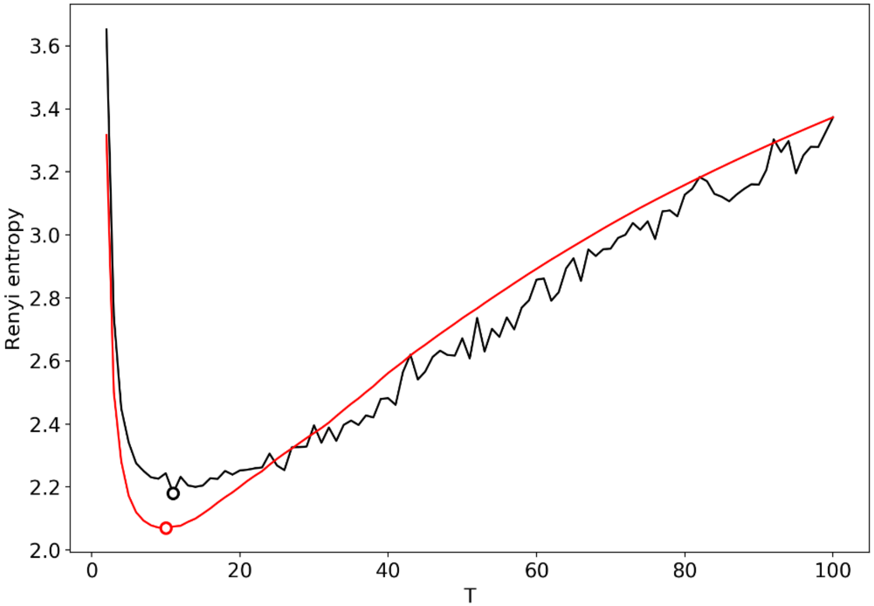

Fig. 2. Rényi entropy curves. Black: successive topic modeling; red is renormalization with minimum local entropy merging.

Numerical experiments

Results for the dataset in Russian.

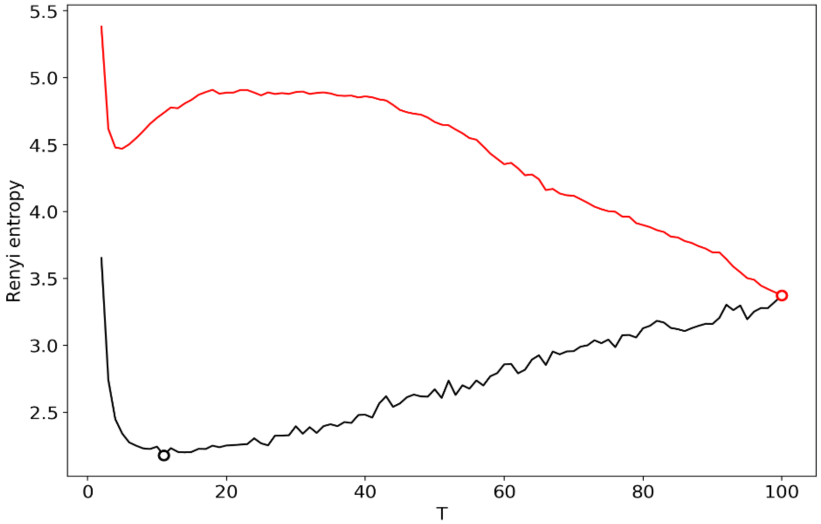

Fig. 3 Rényi entropy curves. Black: successive topic modeling. Red: renormalization with minimum KL divergence principle of merging.

Numerical experiments

Results for the dataset in English.

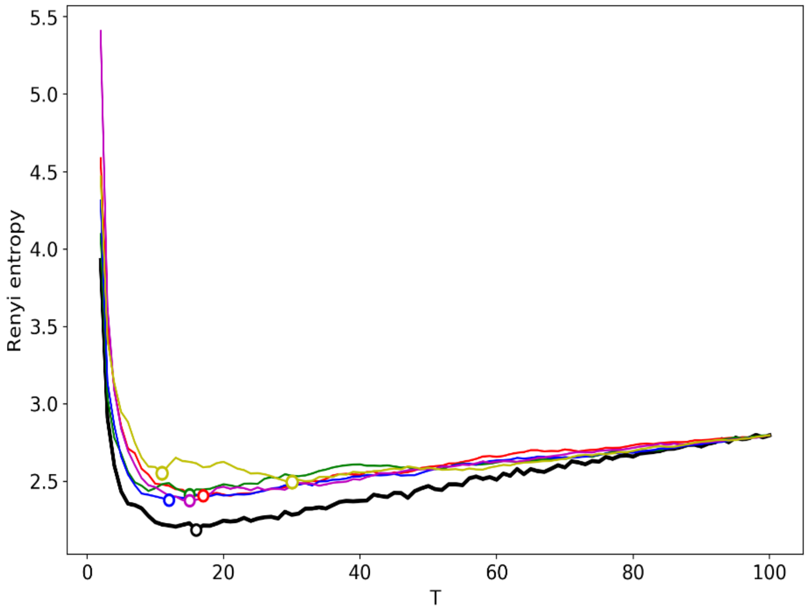

Fig. 4. Rényi entropy curves. Black: successive topic modeling. Other colors: renormalization with the random merging of topics.

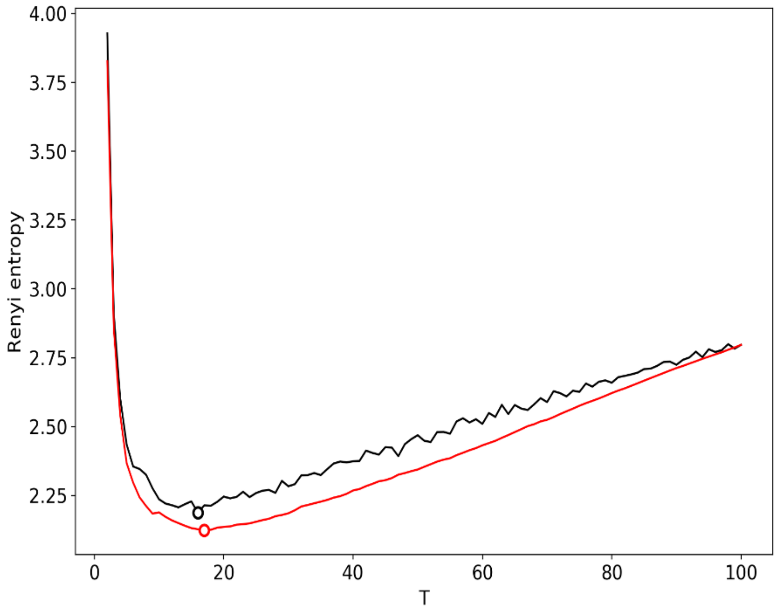

Fig. 5. Rényi entropy curves. Black: successive topic modeling; red is renormalization with minimum local entropy merging.

Numerical experiments

Results for the dataset in English.

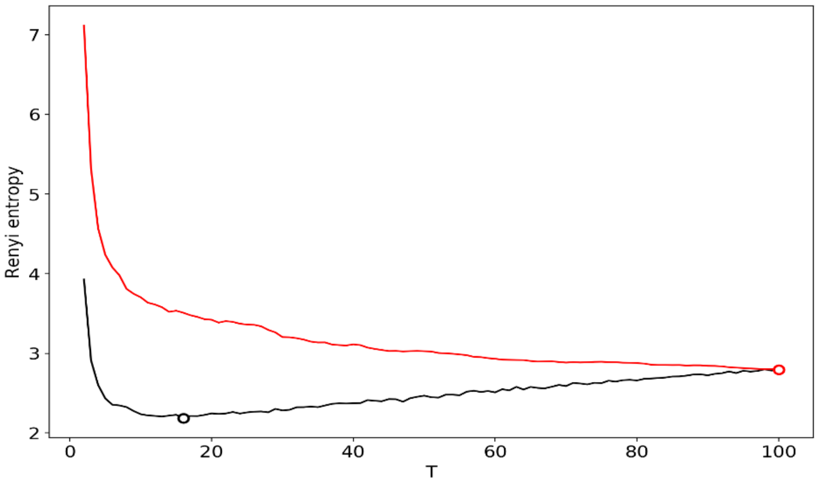

Fig. 3 Rényi entropy curves. Black: successive topic modeling. Red: renormalization with minimum KL divergence principle of merging.

Computational speed

Dataset	Successive TM Simulation	Renormalization (random)	Renormalization (minimum Rényi entropy)	Renormalization (minimum Kullback-Leibler divergence)
Russian dataset	780 min	1 min	1 min	4 min
English dataset	1320 min	3 min	3 min	10 min

The first column corresponds to successive runs of topic modeling for T= [2, 100] in the increments of one topic. Calculation of a single topic solution on 100 topics takes 26 min for the dataset in Russian and 40 min for the dataset in English. One can see that renormalization provides significant gain in time that is essential when dealing with big data.

Results and Discussion

- We demonstrated that renormalization based on merging of topics with minimum local Rényi entropy provides the best result in terms of accuracy and computational speed simultaneously. It was shown that for this type of renormalization, the global minimum point of Rényi entropy is almost equal (with an accuracy of ±1 topic) to the minimum point of Rényi entropy calculated according to successive topic modeling.
- Renormalization with the merging of random topics also leads to satisfactory results; however, it requires multiple runs and subsequent averaging over all runs.
- Renormalization based on minimum Kullback-Leibler divergence does not allow us to determine the optimal number of topics since there is no definite minimum of Rényi entropy.

Results and Discussion

- Let us note that renormalization is applicable to datasets in different languages and with a different number of topics in the collections.
- Application of renormalization allows us to speed up the searching of the optimal number of topics, at least in 26 times.
- The proposed renormalization approach could be adapted for other topic models including models with a sampling procedure of inference.
- Furthermore, our renormalization approach can be adapted for simultaneous estimation of the number of topics and fast tuning of other hyper-parameters of topic models, including regularization parameters.

Supplementary Materials

Samples of topic solutions and the source code of three types of renormalization are available online: https://www.sendspace.com/file/7pzm3j

Funding

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

NATIONAL RESEARCH UNIVERSITY