



NATIONAL RESEARCH  
UNIVERSITY

Laboratory for Social and Cognitive  
Informatics,  
St. Petersburg School of Physics,  
Mathematics, and Computer  
Science, Department of Informatics



# **TOPIC MODELING – ADDITIONAL MODELS, TOPIC NUMBER**

Koltcov S.

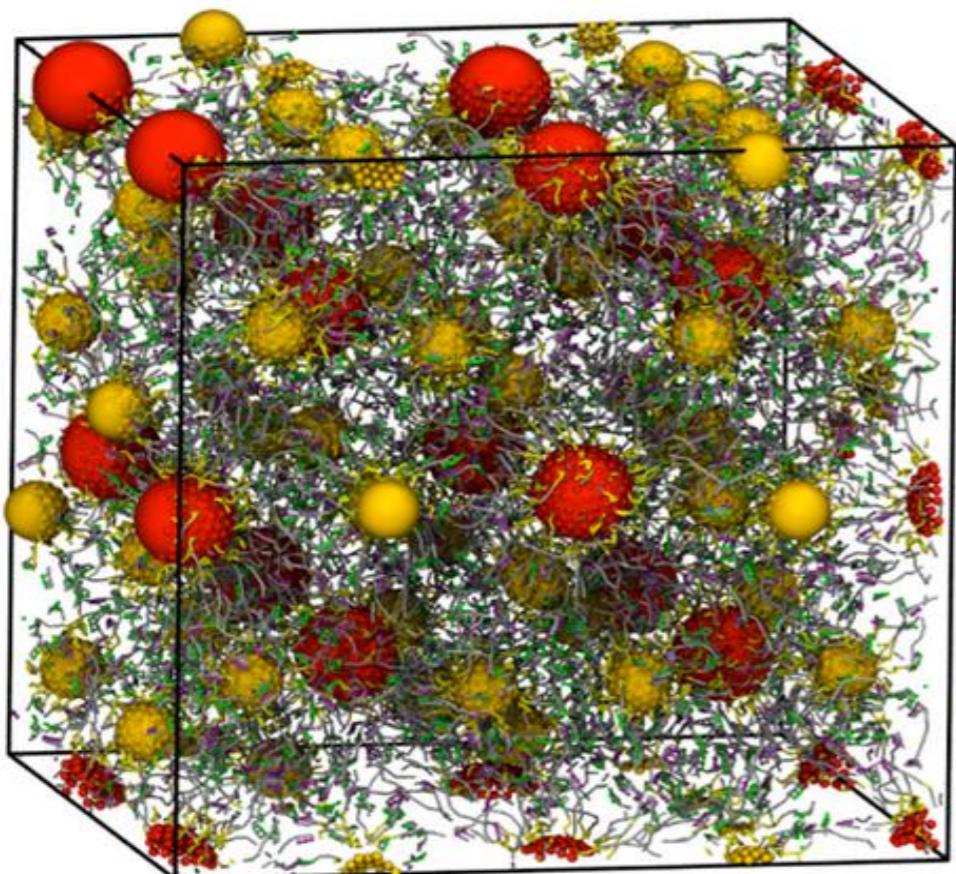
Saint Petersburg, 2024



## Semi-Supervised Latent Dirichlet Allocation (Gibbs sampling)

Одна из идей регуляризации (добавление априорных знаний) основано на том, что можно зафиксировать набор слов по отношению к заданным номерам тем. Соответственно, когда происходит процедура сэмплирования, то для таких слов номера тем не меняются. Для остальных слов производится обычное сэмплирование Гиббса.

$$p(z, w, \alpha, \beta) \propto \begin{cases} z = t & \text{Начальное распределение слов (anchor words)} \\ q(z, w, \alpha, \beta) & \text{Gibbs sampling} \end{cases}$$



Модель SLDA ведет себя как процесс кристаллизации, в котором anchor words являются центрами кристаллизации. Так как фиксированные слова принадлежат определенным темам, то слова, которые часто совместно встречаются с этими словам будут формировать темы. Соответственно такие темы будут более стабильными.

**Недостатком такого подхода является необходимость знания данного набора слов. Однако в ряде случаев это оправданно.**



## GRANULATED LDA (based on Gibbs sampling)

Гранулированный вариант LDA основан на идеи, что каждый документ можно рассматривать как гранулированную поверхность (по аналогии с моделью Поттса). Каждая гранула – набор слов, которые локализированы внутри одной гранулы, соответственно все слова внутри гранулы могут принадлежать одной теме.

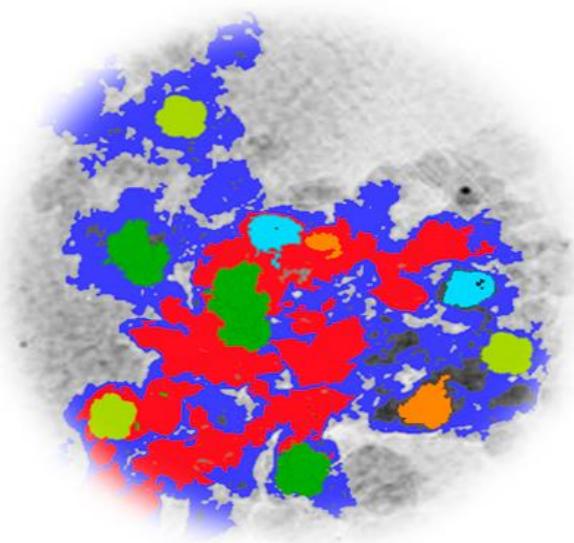
### DOCUMENT

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

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

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

### KEYWORDS

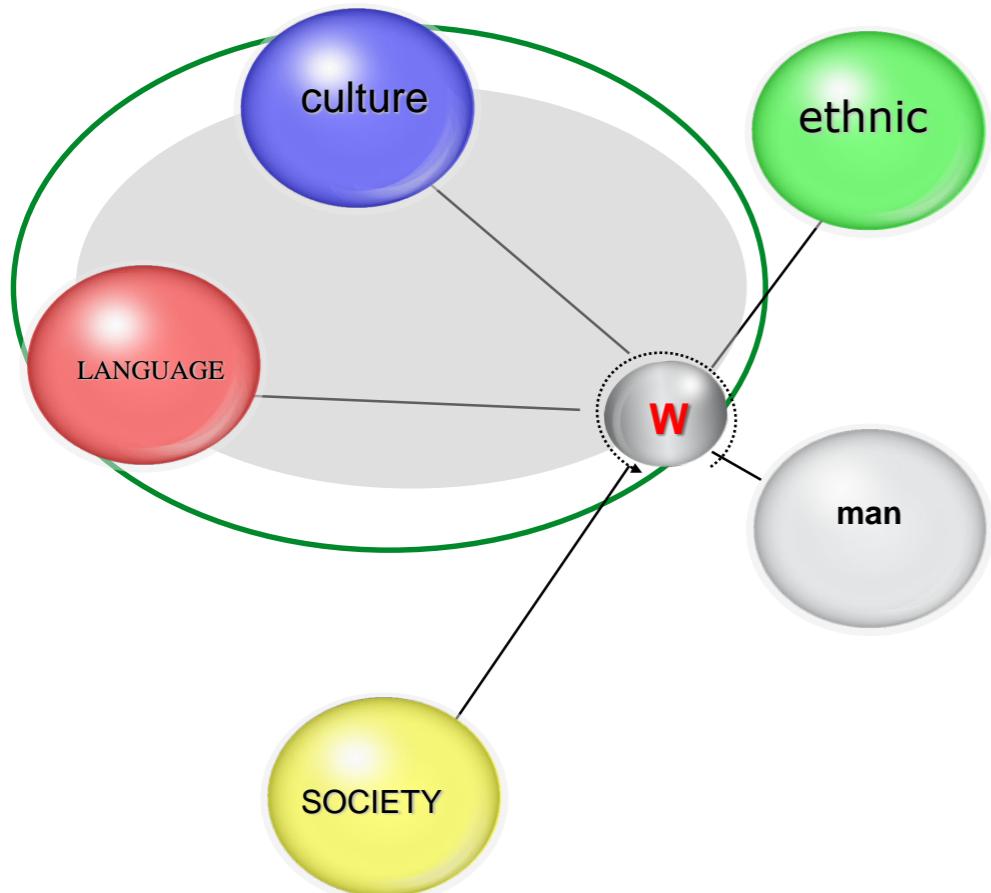




## GRANULATED LDA (based on Gibbs sampling)

**Entrance:** коллекция документов D, число тем |T|, число итераций, размер гранулы

**Initialization:** инициализация матриц  $\phi(w,t)$ ,  $\theta(t,d)$   $d \in D$ ,  $w \in W$ ,  $t \in T$ ;



**Запуск внешнего цикла(i) по документам**  
**Запуск внутреннего цикла (j) по словам**

1. Генерируем случайное число k. Max value of k is number of words in documents i.
2. Выбираем слово под номером k из документа i.
3. Рассчитываем номер темы t для слова k.
4. Присваиваем всем словам внутри гранулы один и тот номер темы.

**Конец внутреннего цикла.**

**Конец внешнего цикла.**

$$\theta_{dj} = \frac{C_{d,j}^{DT} + \alpha}{C_{d,j}^{DT} + T\alpha}$$

$$\phi_{m,j} = \frac{C_{m,j}^{WT} + \beta}{\sum_m C_{m,j}^{WT} + V\beta}$$



## Стабильность различных моделей

Topic model	Topic quality metrics		Topic stability metrics	
	coherence	tf-idf coherence	stable topics	Jaccard
pLSA	-237.38	-126.08	54	0.47
pLSA + $\Phi$ sparsity reg., $\alpha = 0.5$	-230.90	-126.38	9	0.44
PLSA + $\Theta$ sparsity reg., $\beta = 0.2$	-240.80	-124.09	87	0.47
LDA, Gibbs sampling	-207.27	-116.14	77	0.56
LDA, variational Bayes	-254.40	-106.53	111	0.53
SLDA	-208.45	-120.08	84	0.62
GLDA, $l = 1$	-183.96	-125.94	195	0.64
GLDA, $l = 2$	-169.36	-122.21	195	0.71
GLDA, $l = 3$	-163.05	-121.37	197	0.73
GLDA, $l = 4$	-161.78	-119.64	200	0.73

RESULT: регуляризация может существенно влиять на результаты тематического моделирования. Регуляризация может как улучшать так и убивать стабильность тематического моделирования.

(модель LDA является регуляризованной версией pLSA, где регуляризация заключается в факте добавления информации о Dirichlet функциях).



## Энтропийный подход к проблеме настройки параметров ТМ

1. В силу того, что ТМ ориентировано на работу с большими данными (миллионы документов и уникальных слов), то такой набор данных можно рассматривать как макроскопическую систему из большого числа частиц, которая характеризуется термодинамическими величинами, такими как **энергия, энтропия и свободная энергия**. При этом **под температурой в таких системах можно понимать число кластеров или число тем**. То есть, можно рассматривать то, как меняется свободная энергия (**F=E-TS**) при изменении количества кластеров. Так как свободная энергия выражается через энтропию Ренни, то в дальнейшем будем анализировать поведение энтропии Ренни.
2. Выбор энтропии Ренни обусловлен тем, что она является функцией состояния, и определяется температурой для систем с фиксированным числом частиц.
3. Тематическое моделирование является процедурой, в которой система выводится из равновесного состояния (вероятность всех состояний одинакова) в состояние, когда небольшая часть состояний имеет высокую вероятность, а остальная часть состояний имеет очень маленькую вероятность. Например, LDA (Gibbs sampling) – исходное распределение матрицы **Φ** (terms - topics) =  $1/N$ , матрица **Θ** (topics - doc) =  $1/T$ .



## Энтропийный подход к проблеме выбора числа тем и гипер - параметров

4. В рассматриваемой информационной термодинамической системе общее количество слов и документов является константой, то есть, изменение объема отсутствует.
5. Под темой понимается состояние (аналог направления спина), которое может принимать каждое слово и документ в коллекции. При этом как слово, так и документ может принадлежать к разным темам с различной вероятностью.
6. Информационная термодинамическая система является открытой и обменивается только энергией с внешней средой за счет изменения числа кластеров и гипер - параметров. Выбор параметров можно организовать при помощи поиска минимума деформированной энтропии системы (максимум информации). Следует отметить, что данный принцип также можно использовать для отбора типов регуляризаций.
7. В качестве меры неравновесности такой информационной системы можно использовать разность свободных энергий:  $\Delta F = F(T) - F_0$ , где  $F_0$  – свободная энергия начального состояния,  $F(T)$  – свободная энергия при заданной величине  $T$ , а также разницу энтропий Ренни или энтропий Шарма – Миттаала.



## Renyi entropy - Entropic approach

**The entropy of the system** (Shannon's Entropy) in TM can be expressed through the logarithm of the number of states with energy E:  $S = \ln(\check{p}(T))$ ,

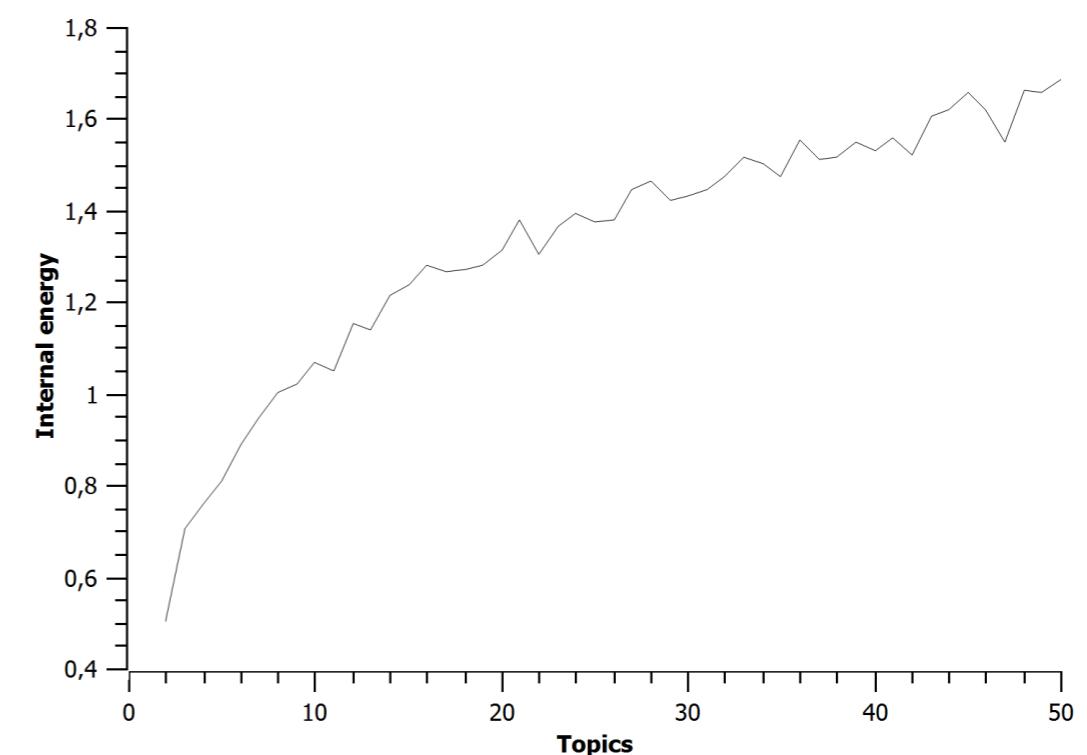
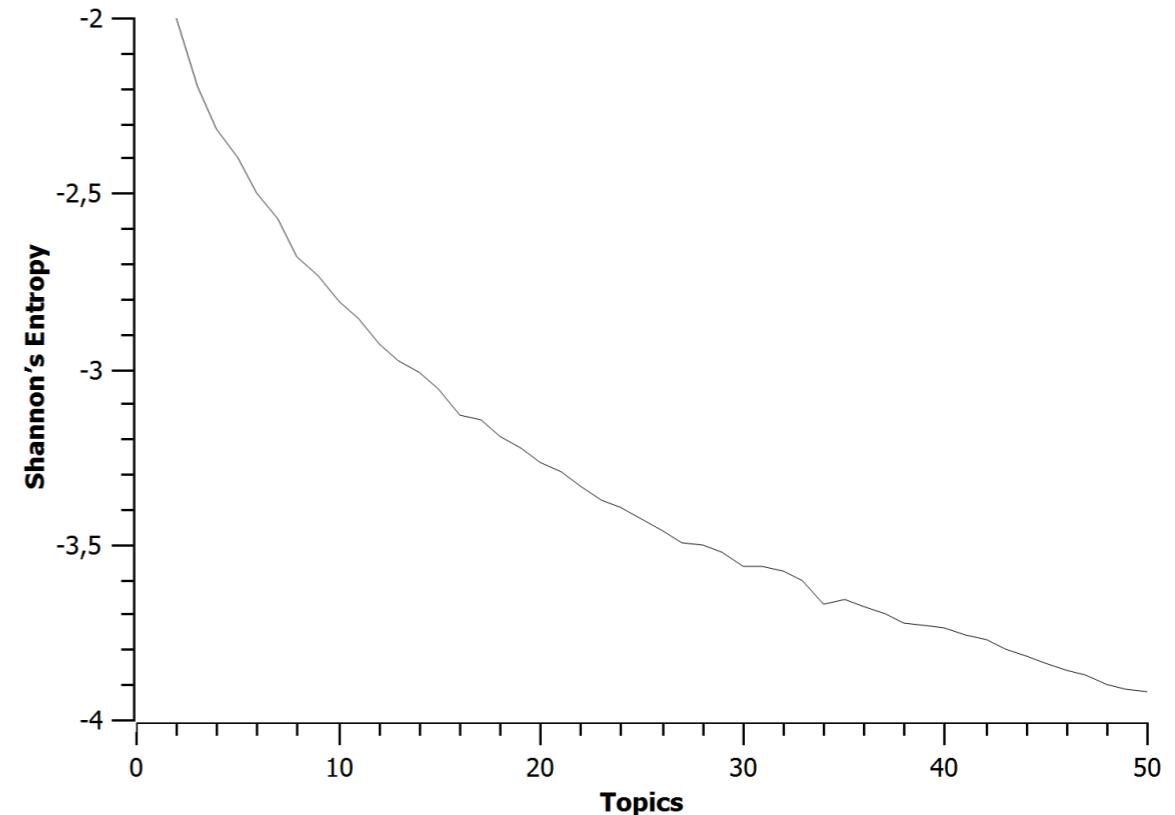
$\check{p} = \frac{\sum_{t=1}^T \sum_{n=1}^W N_{tn}}{WT}$ , where  $N(E)_{tn}$  – number of states with energy E, n,t - summation over all words and topics.

**The internal energy** of the whole system can be determined through the sums of the probabilities of the elements in the system:  $E = -T \cdot \ln \check{P}$ ,  
n is the number of the word in the dictionary, t is the topic number.

$$\check{P} = \sum_{i=1}^n p_i / T.$$

The Helmholtz free energy is expressed as follows:  
 $F = E - T \cdot S$

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T$$
$$= -\ln(\check{P}) - T \cdot \ln(\check{p})$$





## Эксперименты: Датасеты

В данной работе были использованы два датасета.

### 1. Англоязычный датасет ‘20 newsgroup dataset’ (News20).

(<http://qwone.com/~jason/20Newsgroups/>)

Размер англоязычного датасета составляет 15404 поста и N=50948 уникальных слов. Число тем варьировалось в диапазоне: T=[2; 120] с шагом в 2 темы. The data is organized into 20 different newsgroups, each corresponding to a different topic. В силу того, что часть тем сильно скоррелированы друг с другом, то реальное число кластеров в датасете порядка 15 штук.

### 2. Русскоязычный датасет ‘JJ’. Датасет состоит из 101481 русскоязычных постов из социальной сети ‘Живой журнал’ и N=172939 уникальных слов. Число тем варьировалось в диапазоне: T=[2; 330] с шагом в 2 темы. **Не размеченный датасет.**

В каждом эксперименте измерялось количество микросостояний, чьи вероятности больше величины  $1/W$ , где W – число слов в датасете. Таким образом, на основании формулы  $\rho(E) = \frac{\sum_{nt}^{NT} N(E)_{nt}}{NT}$  рассчитывалась функция плотности состояний от числа тем, которая усреднялась по трем запускам. Также рассчитывалась энергия каждого тематического решения в соответствии с формулой  $E(T) - E_0$ . Величина энергии также усреднялась по трем запускам. Величина свободной неравновесной энергии и энтропия Ренни рассчитывалась на основе усредненных значений плотности состояний и внутренней энергии.

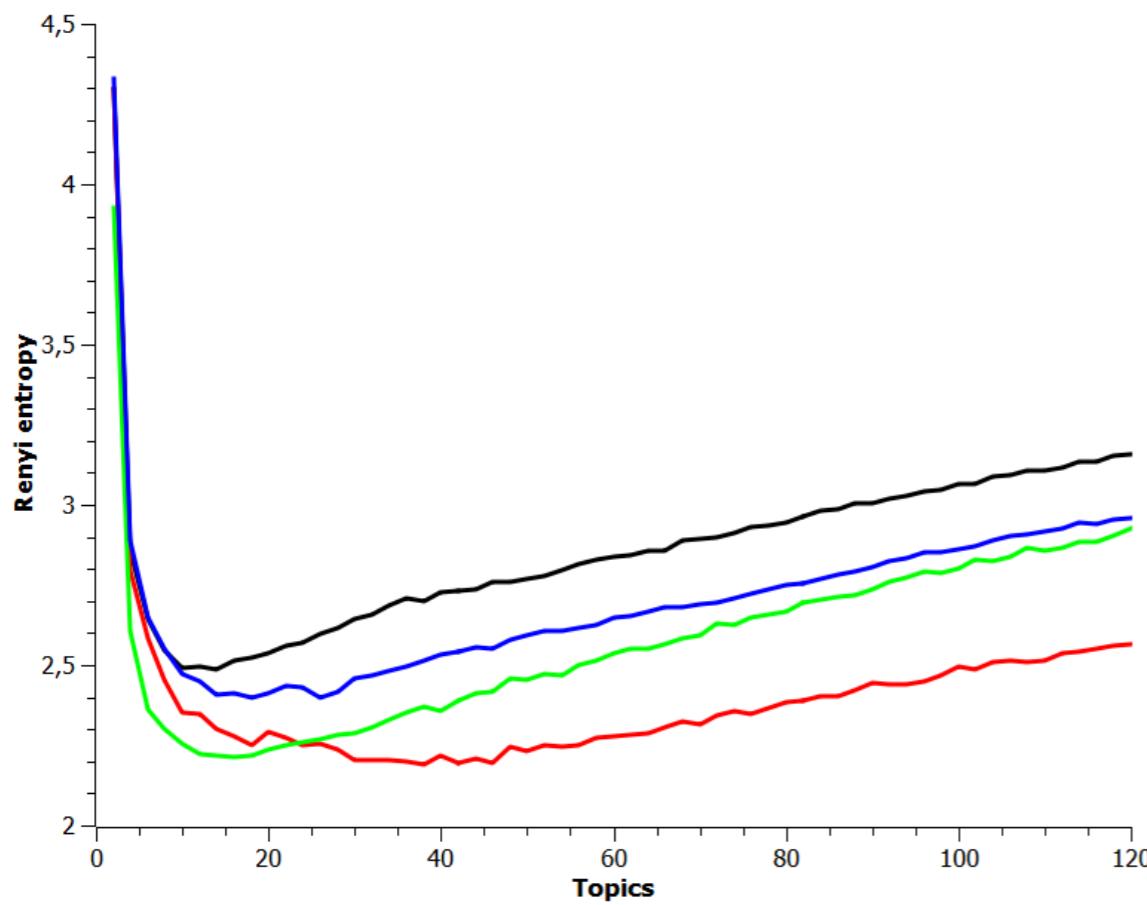


## Renyi entropy

The **Renyi entropy** is expressed in terms of free energy, and allows us to estimate the level of clustering (chaos minimum) depending on the number of topics and other parameters of thematic models.

$$S_{q=1/T}^R = \frac{A_F}{T-1}$$

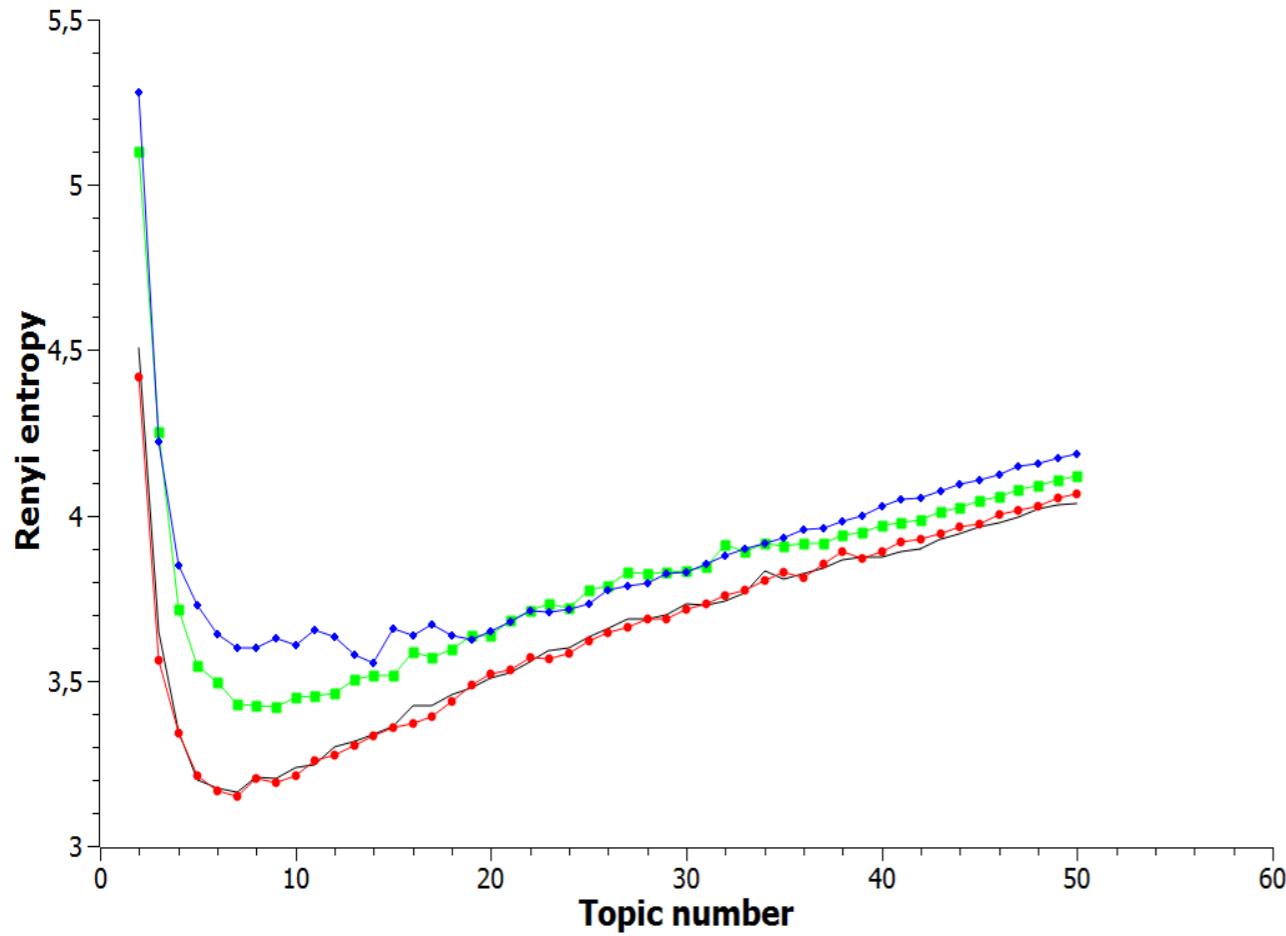
### Experiments: Results of calculating TM on the English dataset (20 topics news)



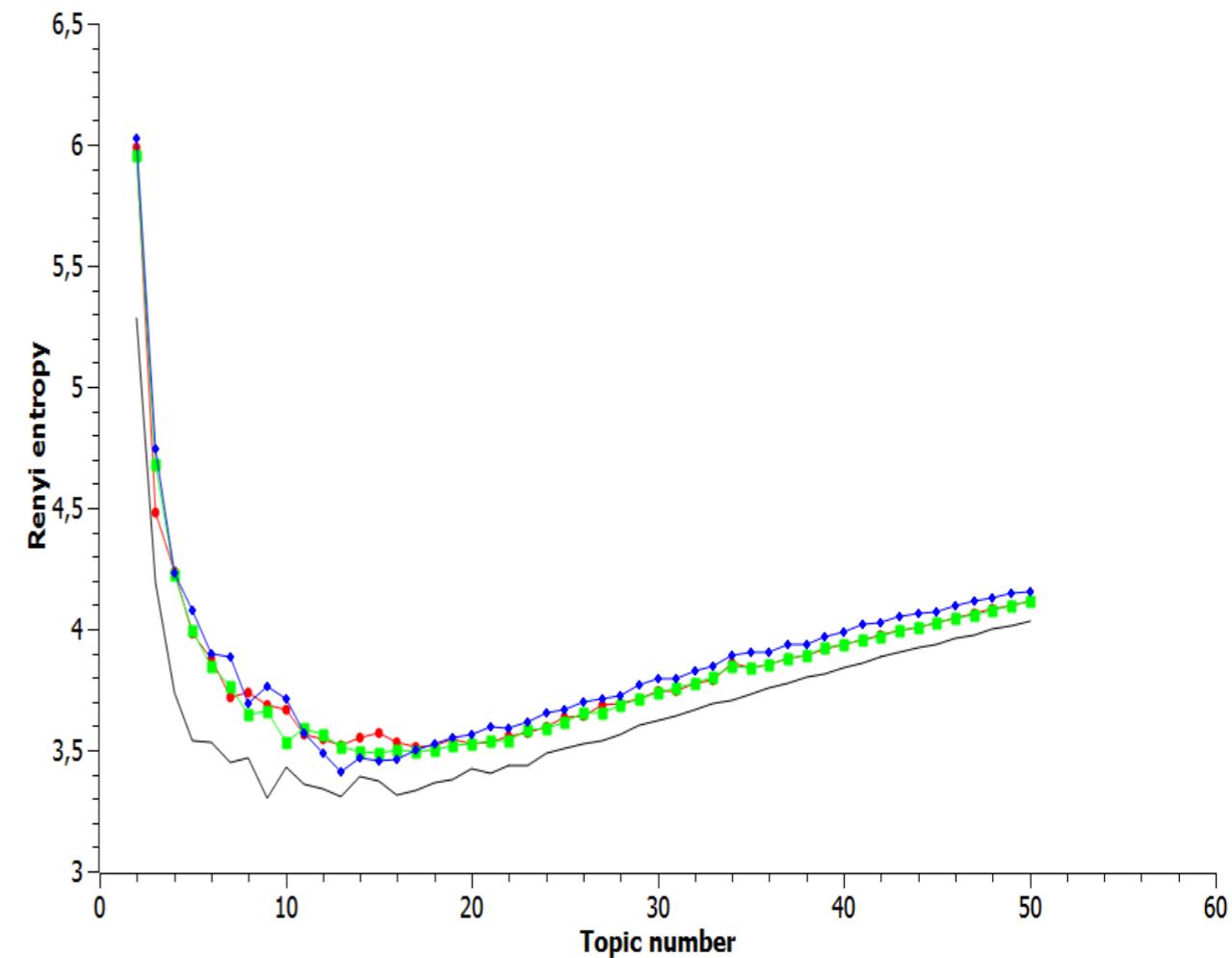
**LDA** (Gibbs sampling) – black color, **GLDA** (Gibbs sampling) – red color, **PLSA** (E-M algorithm) – blue color, **VLDA** (E-M algorithm) – green color.



## Simulation results - Renyi entropy



Renyi entropy distribution over topics (dataset '[lenta\\_ru](#)'). pLSA,— black line, LDA Gibbs sampling ( $\alpha=0.1, \beta=0.1$ ) – red line, LDA Gibbs sampling ( $\alpha=0.5, \beta=0.1$ ) – green line, Gibbs sampling ( $\alpha=1, \beta=1$ ) – blue line.



Renyi entropy distribution over topics ('[20 topics news](#)' datasets). pLSA,— black line, LDA Gibbs sampling ( $\alpha=0.1, \beta=0.1$ ) – red line, LDA Gibbs sampling ( $\alpha=0.5, \beta=0.1$ ) – green line, Gibbs sampling ( $\alpha=1, \beta=1$ ) – blue line.



## Коэффициент Жаккара и семантическая стабильность

Пусть у нас есть два множества X и Y. Тогда можно рассчитать следующие параметры:

- а – множество видов X, отсутствующих в Y;
- б – множество видов Y, отсутствующих в X;
- с – множество видов, общих для X и Y;

Тогда коэффициентов Жаккара называется следующая комбинация:

$$Jk = |c/(a+b-c)|.$$

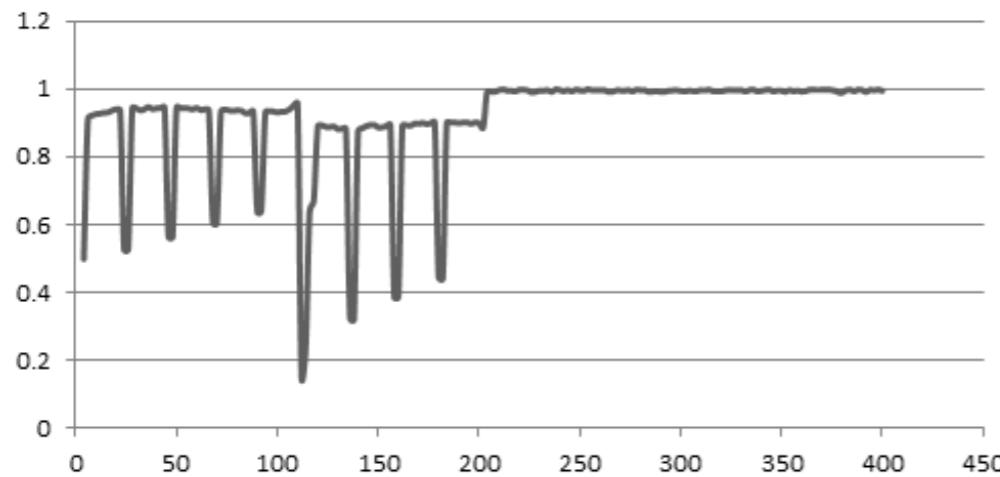
Коэффициент равен 1 в случае полного совпадения двух множеств и равны 0, если множества совершенно различны.



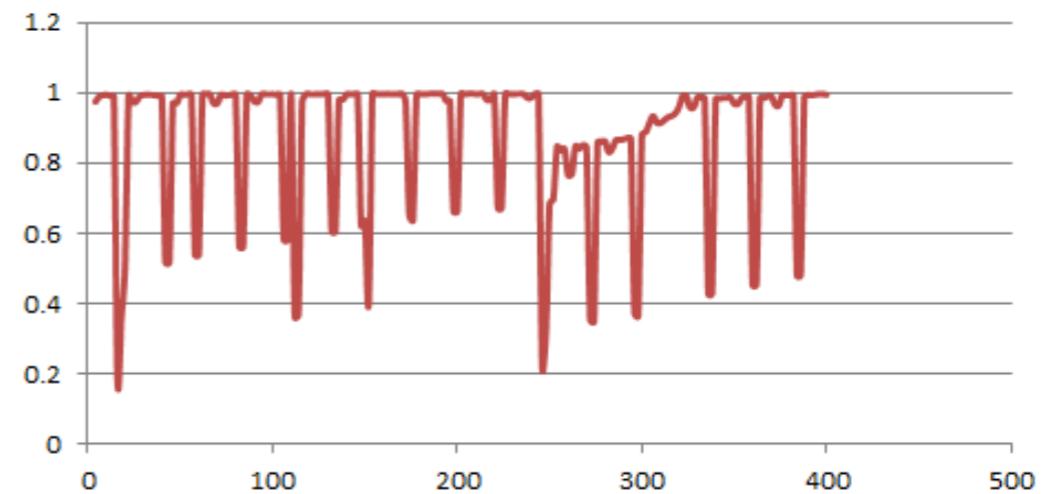
## Семантическая стабильность

Эксперименты: Результаты расчета ТМ на русскоязычном датасете.

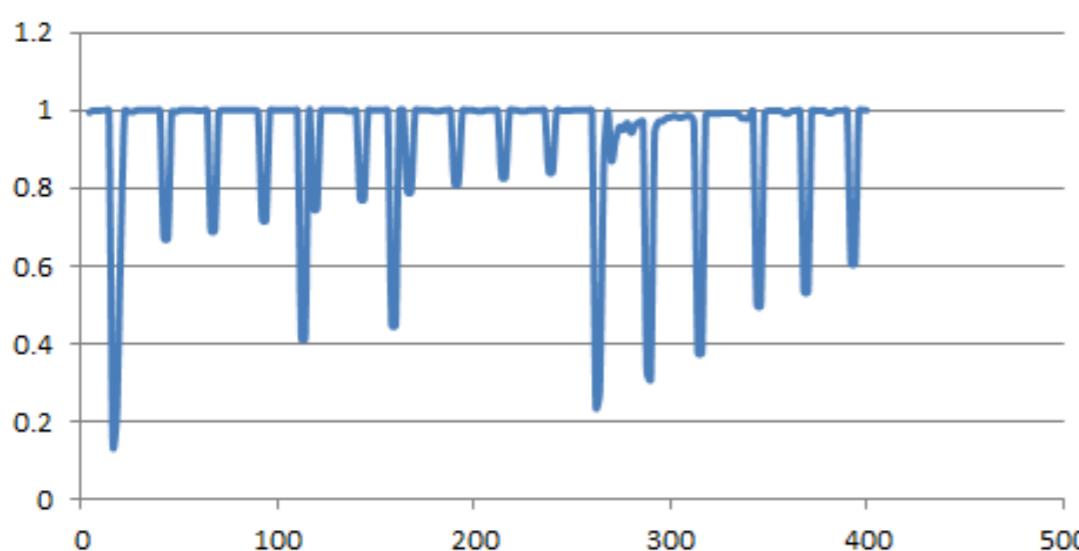
Jaccard (LDA Gibbs)



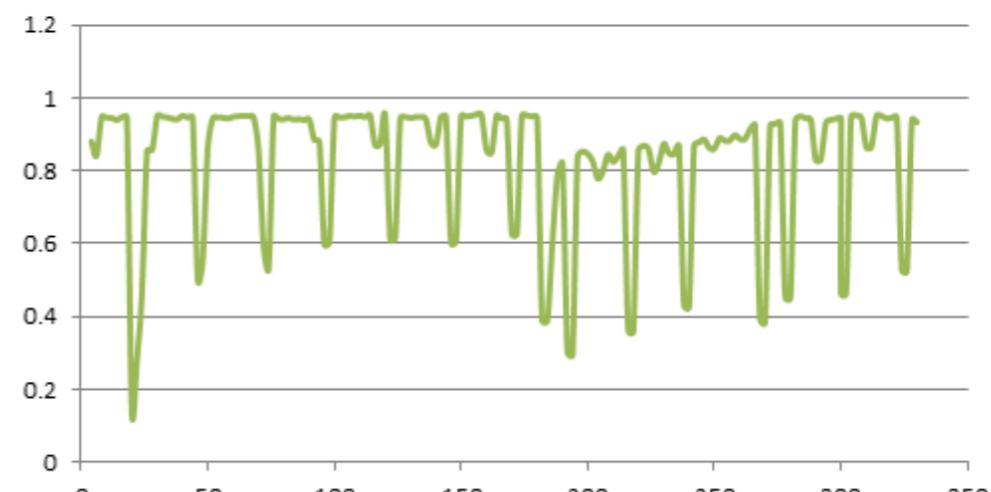
Jaccard (GLDA Gibbs)



Jaccard (PLSA)



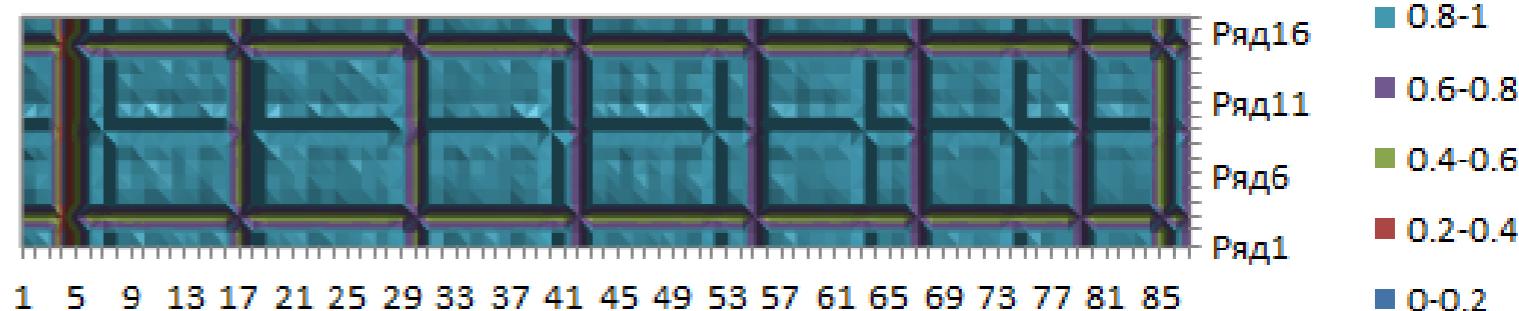
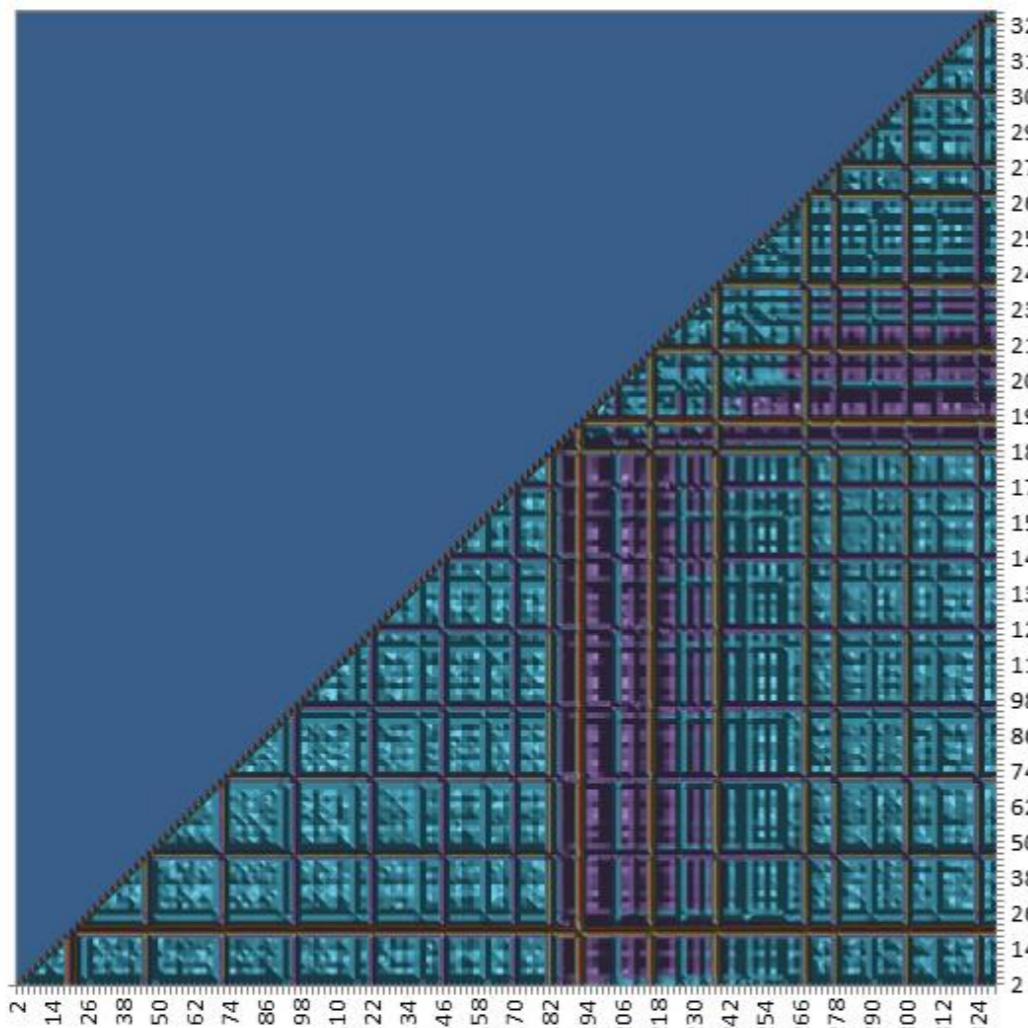
Jaccard (VLDA)





## Семантическая стабильность

### Эксперименты: Поверхность семантической стабильности



### Модель VLDA (Блей)

Цветом выделены  
уровни коэффициента  
Жаккара



## Sharma-Mittal Entropy in topic modeling

$$S_{S,M} = \frac{1}{1-r} \left[ \left( \sum_i p_i^q \right)^{\frac{1-r}{q-1}} - 1 \right]$$

Параметры  $q$  и  $r$  являются параметрами деформации.

Соответственно,  $q = 1/T$ , а в качестве параметра  $r$  можно использовать коэффициент Жаккара. В этом случае,  $1-r$  не что иное как entropy distance or Jaccard distance, где  $W'$  и  $W''$  являются наборами слов с высокой вероятностью разных тематических моделей (разное число тем).

$$\lim_{r \rightarrow 1} S_{S,M} = \lim_{r \rightarrow 1} \frac{1}{1-r} \left( 1 + \frac{1-r}{q-1} \ln \left( \sum_i p_i^q \right) + \sigma((1-r)^2) - 1 \right) = \lim_{r \rightarrow 1} \frac{1}{1-r} \left( \frac{1-r}{q-1} \ln \left( \sum_i p_i^q \right) + +\sigma((1-r)^2) \right)$$



## Sharma-Mittal Entropy in topic modeling

$$\lim_{r \rightarrow 0} S_{S,M} = \lim_{r \rightarrow 0} \frac{1}{1-r} \left( \left( \sum_i p_i^q \right)^{\frac{1-r}{1-q}} - 1 \right) = \left( \sum_i p_i^q \right)^{\frac{1}{1-q}} - 1 = e^{\frac{1}{1-q} \ln(\sum_i p_i^q)} - 1 = e^{S_q^R} - 1$$

Получаем, что предел энтропии Шарма-Митталь при  $r \rightarrow 0$  равен экспоненте от энтропии Ренъи без единицы, что можно рассмотреть как деформированную перплексию. Таким образом, когда  $r=0$ , то энтропия Шарма-Митталь выглядит следующим образом:  $S_{S,M} = e^{S_q^R} - 1$ , то есть энтропия становится чрезвычайно большой. Соответственно, изменение коэффициента Жакара приводит к резким скачкам энтропии, что позволяет использовать энтропию Шарма – Миттала для определения изменения семантической связности при вариации числа тем и гипер – параметров.

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

Энтропия Шарма – Миттала выражается через экспериментально определяемые параметры: 1.  $\check{P}$  – сумма вероятностей слов (слова с высокими величинами вероятности),  $\check{\rho}$  – функция плотности слов в модели.



## Эксперименты: Датасеты - Модели

1. pLSA (на базе dll из BigARTM)
2. LDA (Gibbs sampling).

Каждая из моделей была использована при тематическом моделировании, где менялось число тем в диапазоне [2-50] с шагом в одну тему, и менялись гипер – параметры  $\alpha$ ,  $\beta$  в диапазоне [0-1] с шагом 0.1 для каждой темы и для каждого датасета. В каждой из моделей и для каждого датасета при вариации выше указанных параметров, были рассчитаны следующие метрики: 1. Log-likelihood. 2. Jaccard distance 3. Sharma - Mittal Entropy.

В данной работе были использованы два датасета.

1. Англоязычный датасет ‘20 newsgroup dataset’
2. Russian Dataset (from Lenta.ru news agency):

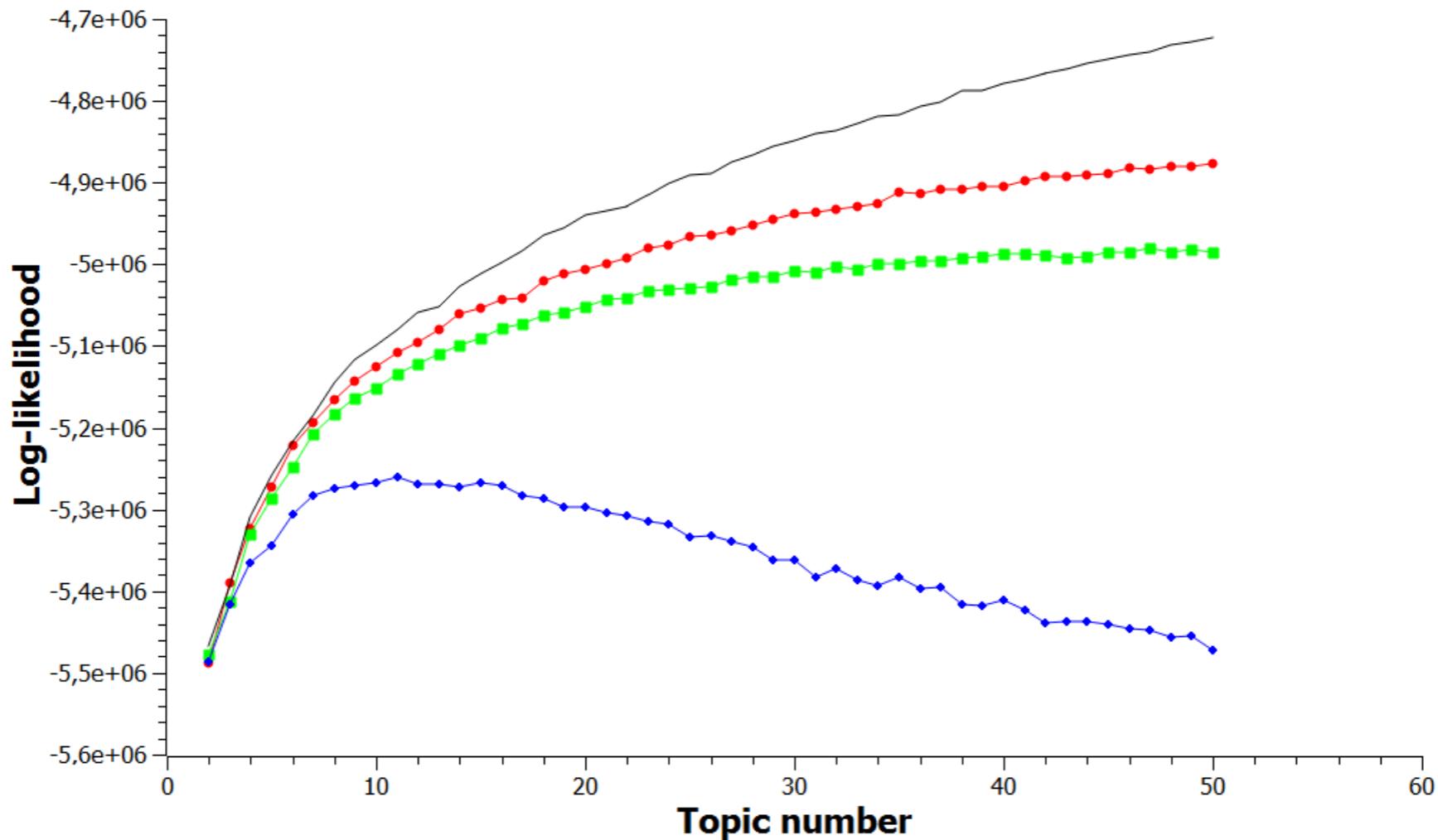
8,624 документов (содержит 23,297 уникальных слов).

Часть тем пересекаются между собой по смыслу, поэтому число тем лежит в диапазоне [7 - 10] штук.

**Table 1.** Statistics on the Russian dataset

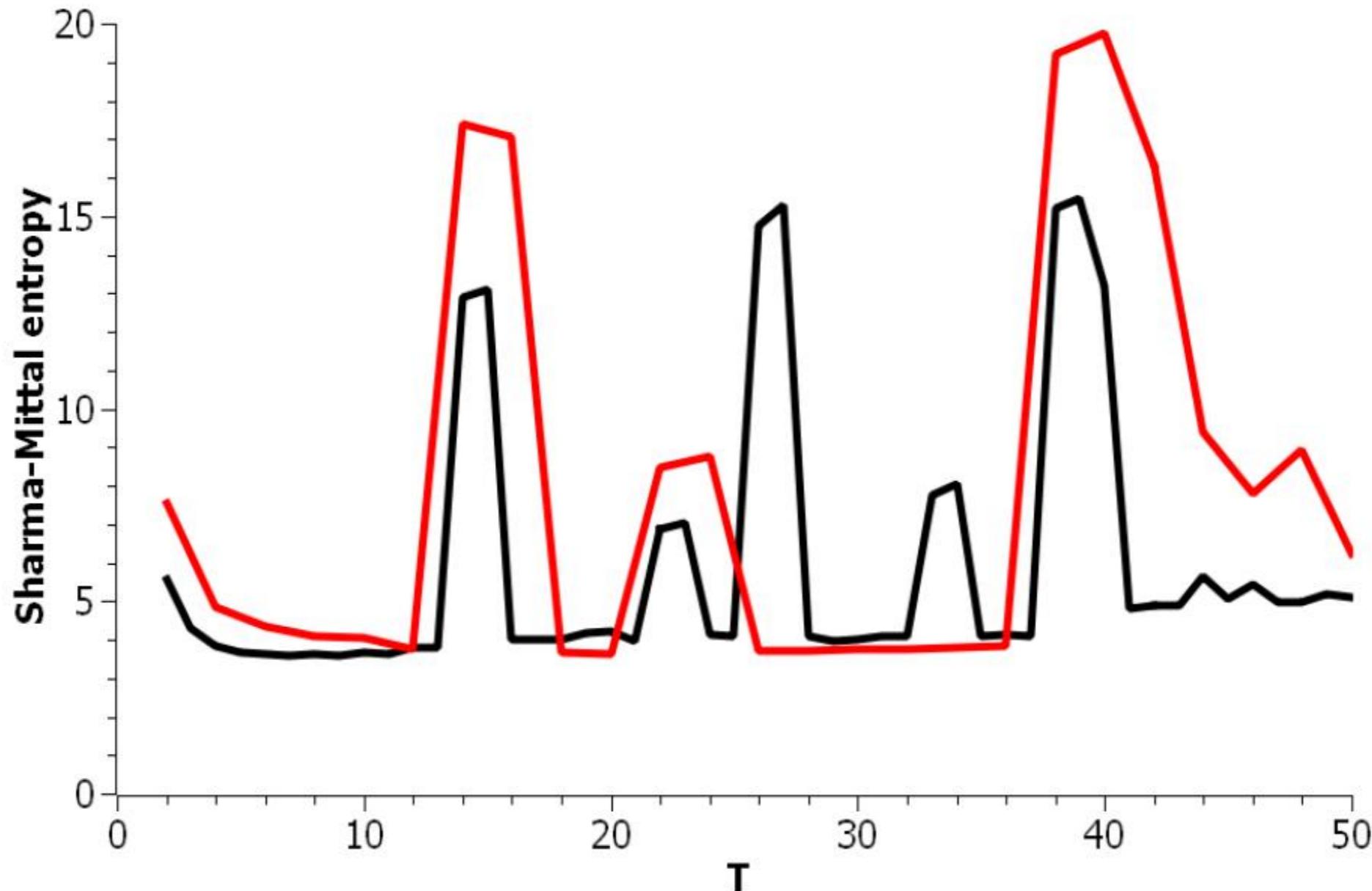
Category	Number of documents
business	466
culture	499
economy and finance	667
incidents	712
media	628
policy	1231
security services	863
science and tech	580
society and travel	1957
sport	1022

## Log-likelihood для русскоязычного датасета



Log-likelihood distribution over topics (dataset ‘lenta\_ru’).  
pLSA – Black line, LDA, LDA Gibbs sampling ( $\alpha=0.1, \beta=0.1$ ) – red line, LDA Gibbs sampling ( $\alpha=0.5, \beta=0.1$ ) – green line, Gibbs sampling ( $\alpha=1, \beta=1$ ) – blue line.

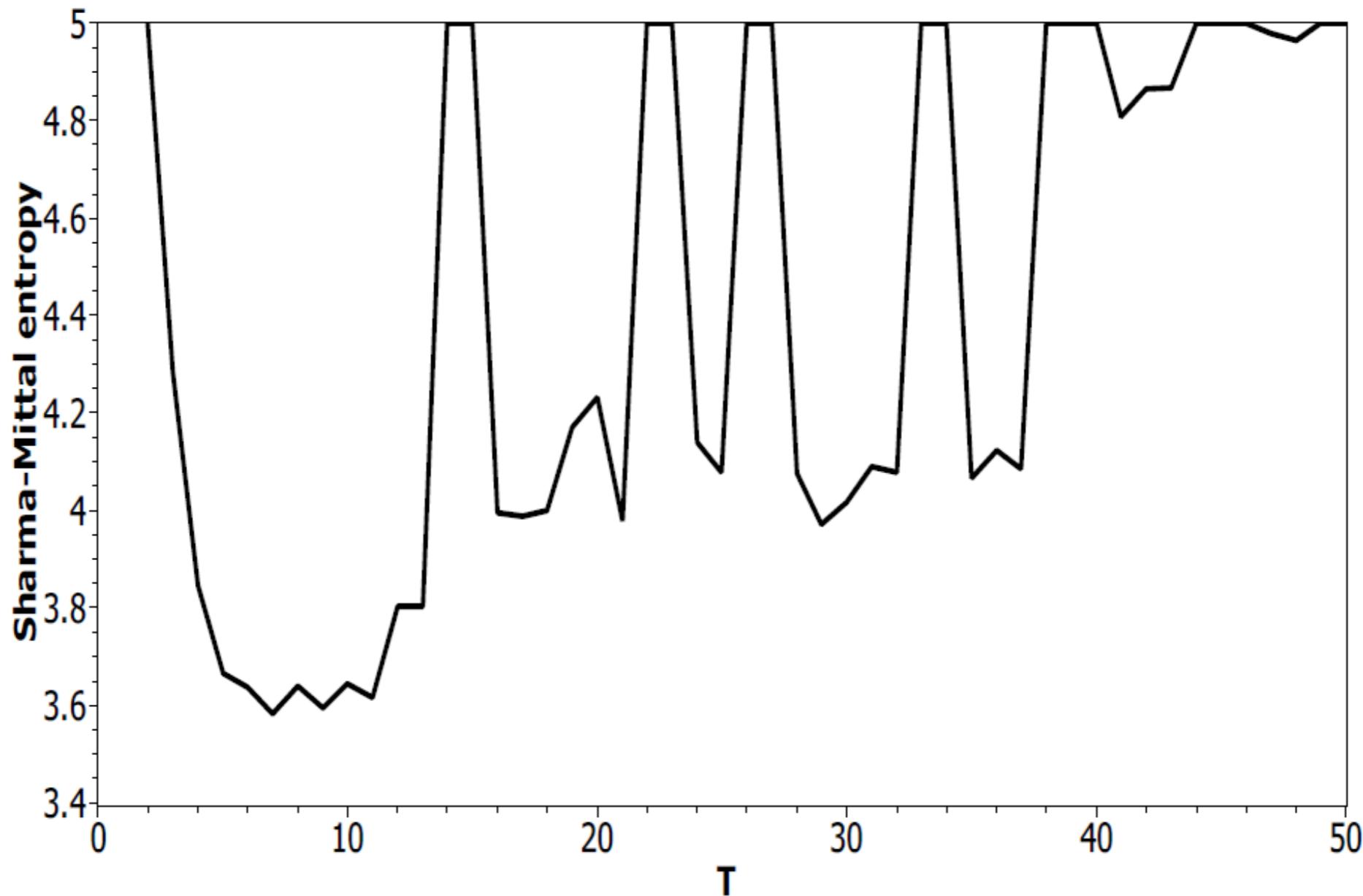
## Sharma-Mittal entropy



Sharma-Mittal entropy distribution over the number of topics  $T$   
(pLSA). Russian dataset - black, English dataset - red.

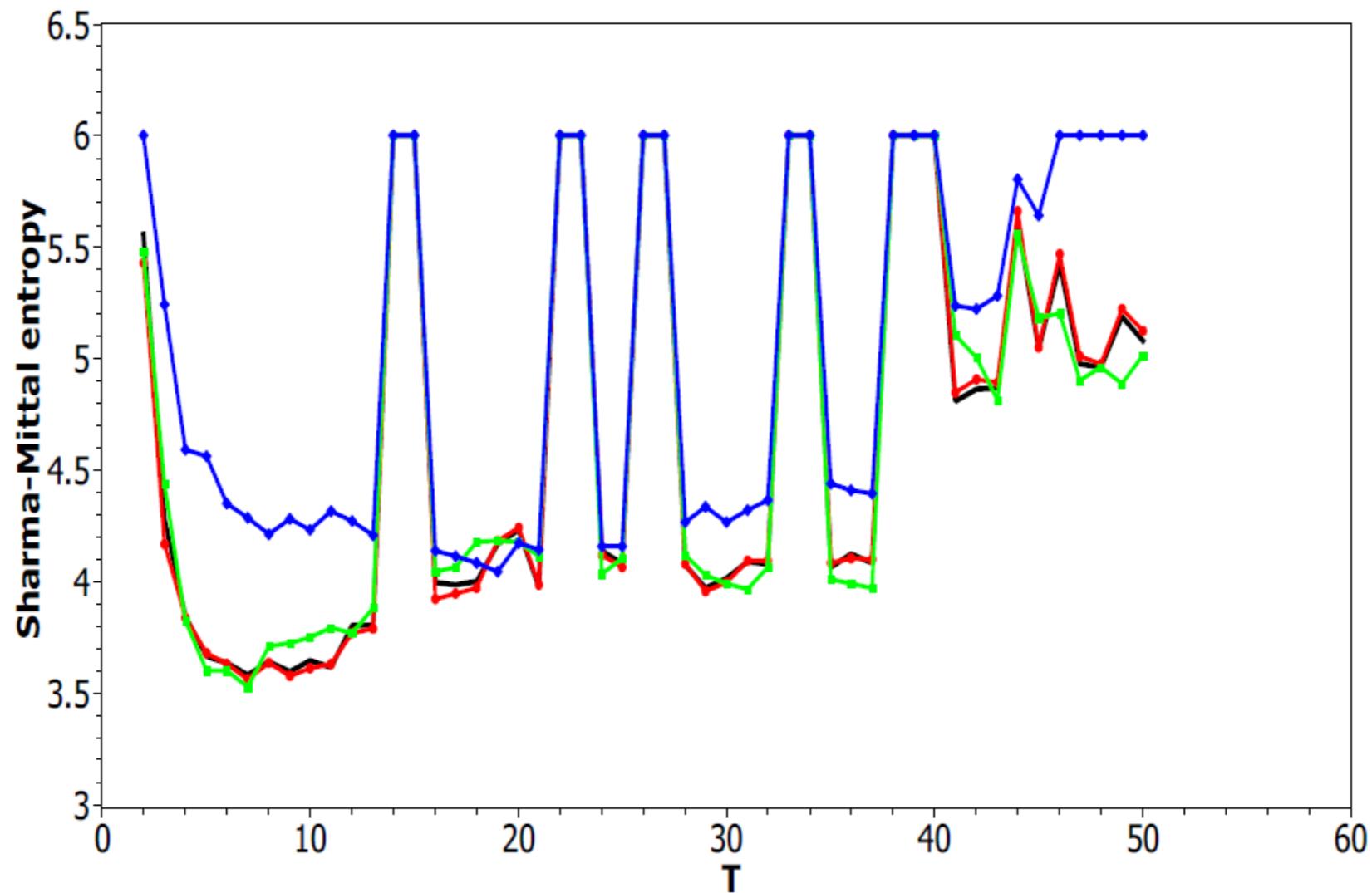


## Sharma-Mittal entropy при вариации числа тем



Sharma-Mittal entropy distribution over  $T$ , Russian dataset, pLSA

## Sharma-Mittal entropy при вариации Т и гипер - параметров



Sharma-Mittal entropy distribution over topics (Russian dataset). pLSA – black, LDA ( $a = 0.1, b = 0.1$ ) – red, LDA ( $a = 0.5, b = 0.1$ ) – green, LDA ( $a = 1, b = 1$ ) – blue.



# **Ренормализация тематических моделей**

Проблема подбора гипер - параметров в тематических моделях заключается в том, что требуется огромное время на расчеты модели при вариации параметров. Однако эту проблему можно частично решить за счет применения технологии ренормализации. Даный подход базируется на самоподобном поведении статистической системы.



## Мультифрактальное поведение тематических моделей

Тематическое решение при фиксированном числе тем представляет собой матрицу  $\phi_{wt}$ , в которой общее число ячеек  $T^*W$ , где  $T$  – число тем (столбцов в матрице),  $W$  – число уникальных слов (строк в матрице). Размер каждой ячейки  $\varepsilon=1/(WT)$ . Каждая ячейка матрицы содержит вероятность  $P_{ij}$  принадлежности слова  $w_i$  к теме  $T_j$ . При фиксированном размере словаря  $W=\text{const}$ , размер ячейки определяется только количеством тем в модели и при  $T \rightarrow \infty$ , размер ячеек стремится к нулю. Функция плотности распределения слов имеет

вид:  $\rho(E) = \frac{\sum_{ij}^{NT} N(E)_{nt}}{NT}$ ,  $n_i$ =число ячеек в тематическом решении, содержимое которых выше

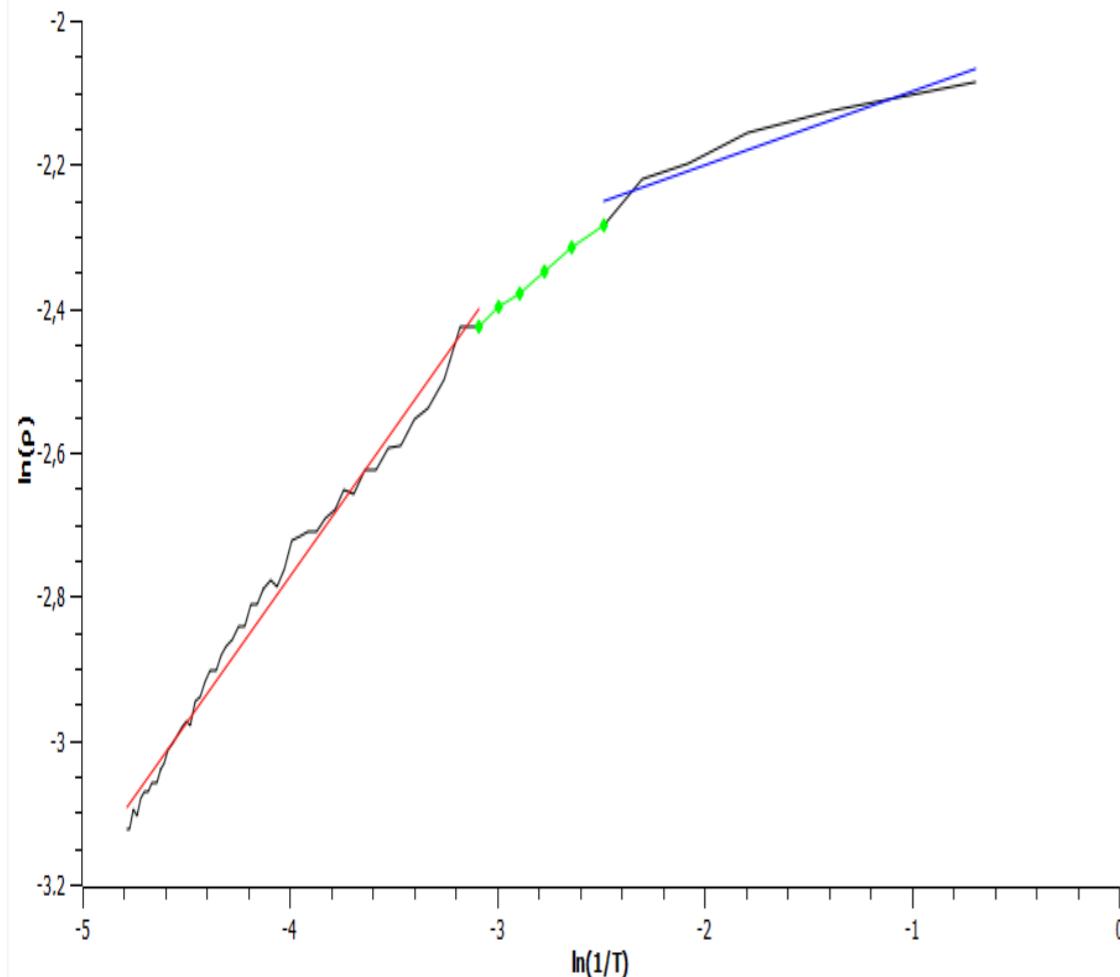
величины  $P_{ij}>1/N$ . Данная величина является функцией числа тем, и в ходе тематического моделирования меняется от 1 до некоторого значения  $\rho_i(\varepsilon)<1$ , которое зависит от количества тем. Таким образом, плотность  $\rho_i(\varepsilon)$  зависит от размера ячеек и степени  $D(\varepsilon)$ :  $\rho(\varepsilon) \cong \varepsilon^{-D(\varepsilon)}$ .

Соответственно, степень  $D(\varepsilon)$  можно определить при помощи алгоритма ‘box counting’. Этапы расчета фрактальной степени: **1.** Многомерное пространство слов покрывается сеткой фиксированного размера, которая является матрицей  $\phi_{wt}$ . **2.** Подсчитывается количество ячеек, в которых вероятность слов больше величины  $P_{ij}>1/N$ . **3.** Рассчитывается величина  $\rho_i$  для заданной величины  $T_i$ . **4.** Меняем размер ячейки  $\varepsilon=1/T$ . **5.** Строится график зависимости  $\rho(\varepsilon)$  в билогарифмических координатах. **6.** Методом наименьших квадратов оценивается наклон на графике, и он представляет собой фрактальную размерность, взятую с обратным

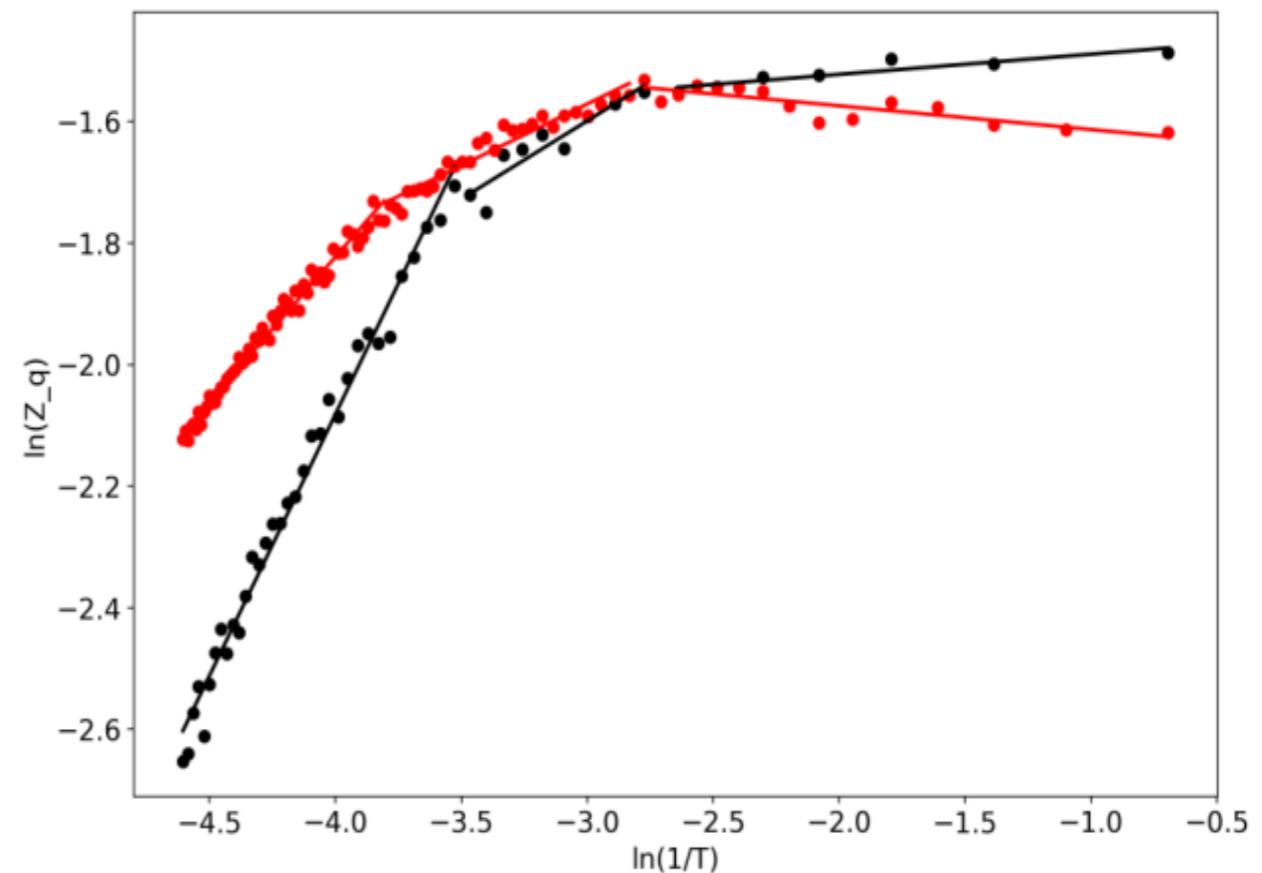
знаком:  $D(\varepsilon) = -\frac{\ln(\rho(\varepsilon))}{\ln(\varepsilon)}$ .



## Мультифрактальное поведение тематических моделей



Распределение фрактальной размерности тематических моделей для англоязычного датасета (LDA Gibbs sampling).



Поведение статсуммы модели VLDA в билогарифмических координатах.  
Black: Lenta dataset;  
Red: 20 Newsgroups dataset.

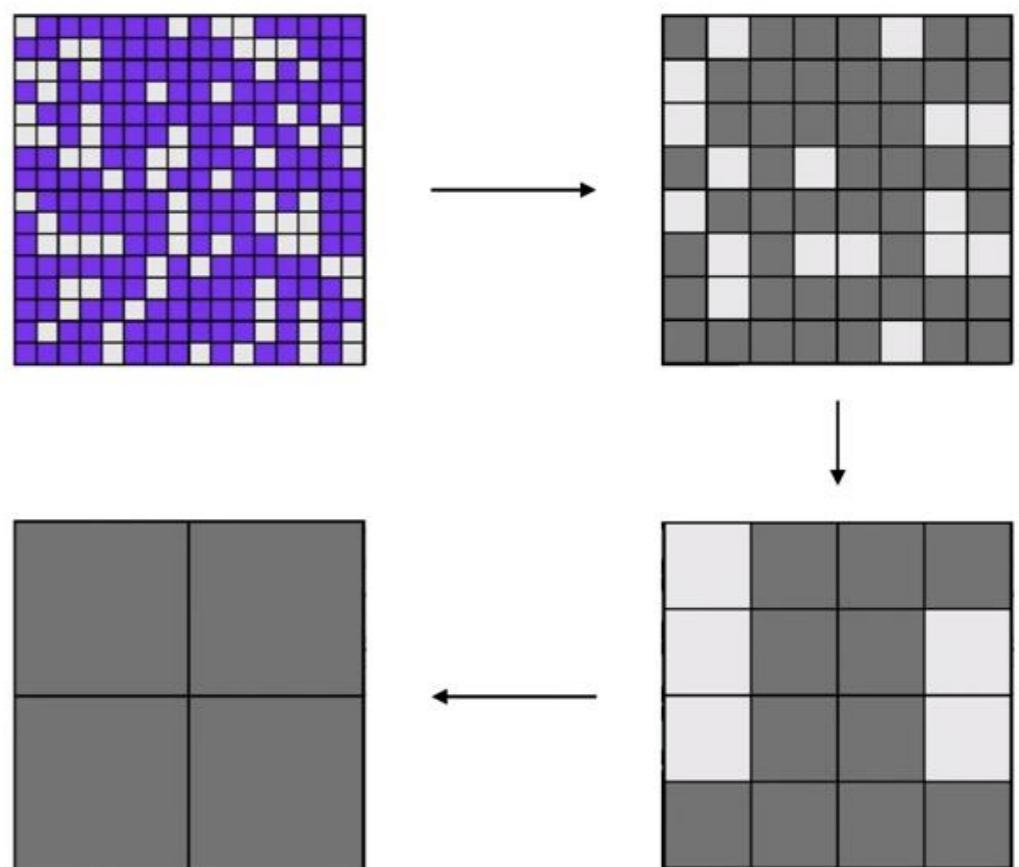
Линейная область соответствует ситуации, когда функция плотности распределения является самоподобной, то есть  $p(\lambda \cdot 1/T) = \lambda^d \cdot p(1/T)$ .



## Введение в теорию ренормализации

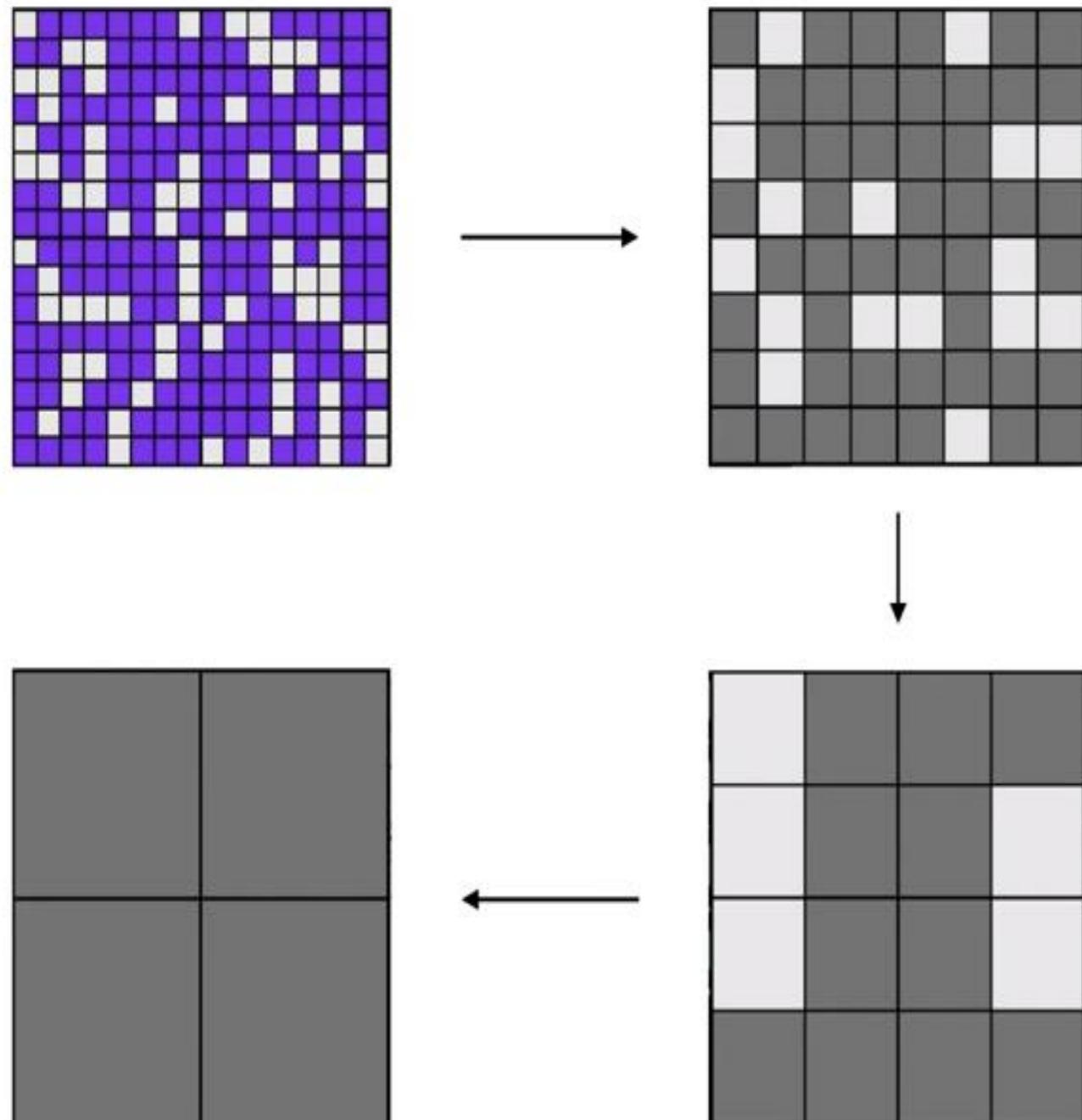
**Ренормализация** – математический формализм, который заключается в построении процедуры изменения масштаба системы, при котором поведение системы остается тем же самым. Широкое развитие процедуры перенормировки получили в теории фракталов, так как фрактальное поведение обладает свойством самоподобия.

**Суть процедуры ренормализации заключается в следующем.** Рассмотрим поверхность состоящих из совокупности узлов. Каждый из узлов, характеризуется направлением спина. В свою очередь, спин может занимать определение направление, количество которых зависит от задачи. Узлы с одинаковыми спинами составляют кластеры. Процедура скейлинга или ренормализации происходит по принципу блочного объединения, в котором несколько ближайших узлов заменяются на один узел. В качестве направления спина нового узла берется направление спинов, составляющие большинство в выбранном блоке. Процедуру блочного объединения проводится по все поверхности. Соответственно в результате появляется новая конфигурация спинов.

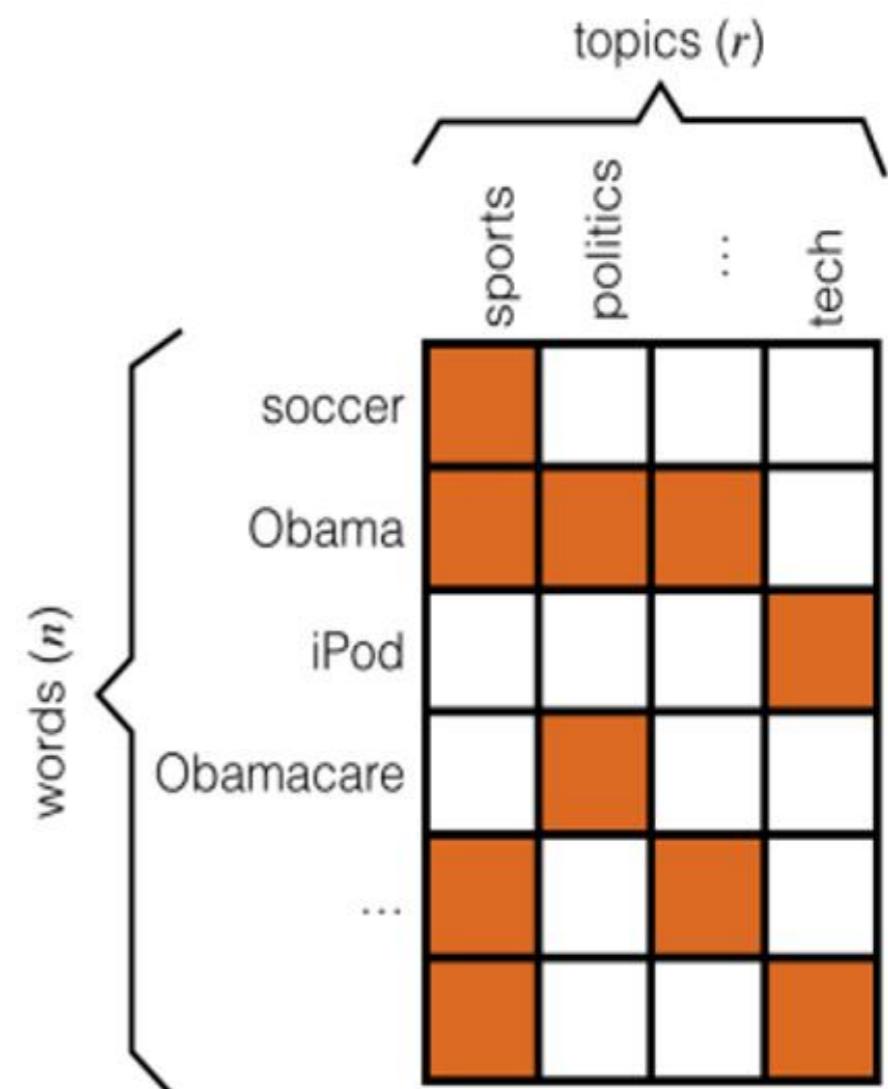




## Введение в теорию ренормализации тематической модели



Можно рассмотреть матрицу ‘**слова** – **темы**’ с точки зрения блочного объединения. Однако в качестве блоков использовать темы.





## Ренормализации в тематическом моделировании

Общая процедура ренормализации тематических моделей заключается в следующем. Результатом тематического моделирования является матрица  $\phi_{wt}$ , которая состоит из ряда одномерных распределений слов по темам. Размеры матрицы определяются числом слов  $w$ , и числом тем  $T$  (или обратной величиной  $q=1/T$ ). В данной работе рассматривается фиксированный словарь уникальных слов, поэтому масштаб ренормализации будет зависеть только от параметра  $q=1/T$ . **Процедура огрубления, то есть ренормализация, заключается в последовательном огрублении одного тематического решения при вариации числа тем и подсчете энтропии Ренъи на каждом шаге ренормализации.** **Процедура огрубления — это процедура слияния двух тем в одну тему.** После объединения двух тем, новая тема также нормируется, так как сумма вероятностей всех слов в одной теме всегда равна единице вне зависимости от числа тем. В силу того, что расчет матрицы  $\phi_{wt}$  зависит от типа алгоритма, математическая формулировка расчета ренормализации специфична для модели. Кроме того, результат слияния зависит от того какие темы попарно объединяются. В рамках данной работы рассматриваются три принципа объединения тем:

1. Объединение тем, которые похожи друг на друга. Критерием схожести тем является минимальная величина среди всех попарных значений меры Кулбака – Лейблера.
2. Объединение тем на основе минимума локальной энтропии Ренъи.
3. Принцип случайного объединения, где номера тем генерируются случайным образом.



## Принципы объединения двух тем

1. **Принцип попарного объединения тем на основе минимума Кулбака – Лейблера.** В данном принципе предполагается, что объединять нужно темы, которые имеют минимум KLB. Для этого производится попарный расчет по следующей формуле:

$$D_{KL}(p \mid q) = \frac{1}{2} \sum_{i=1} p(x_i) \cdot \ln \left( \frac{p(x_i)}{q(x_i)} \right) + \frac{1}{2} \sum_{i=1} q(x_i) \cdot \ln \left( \frac{q(x_i)}{p(x_i)} \right)$$

2. **Принцип объединения тем на основе минимума энтропии Ренъи.** В данном подходе рассчитывается энтропия Ренъи для каждой темы по отдельности. Далее производится сортировка всех тем по величине энтропии Ренъи. Две темы объединяются в одну, если у этих тем наименьшие значения.

3. **Объединение случайно выбранных колонок.** В данном подходе генерируются два случайных числа в диапазоне 1-Т, далее производится объединения колонок, чьи номера сгенерированы случайным образом. Данный принцип объединения характеризуется большой скоростью расчета.



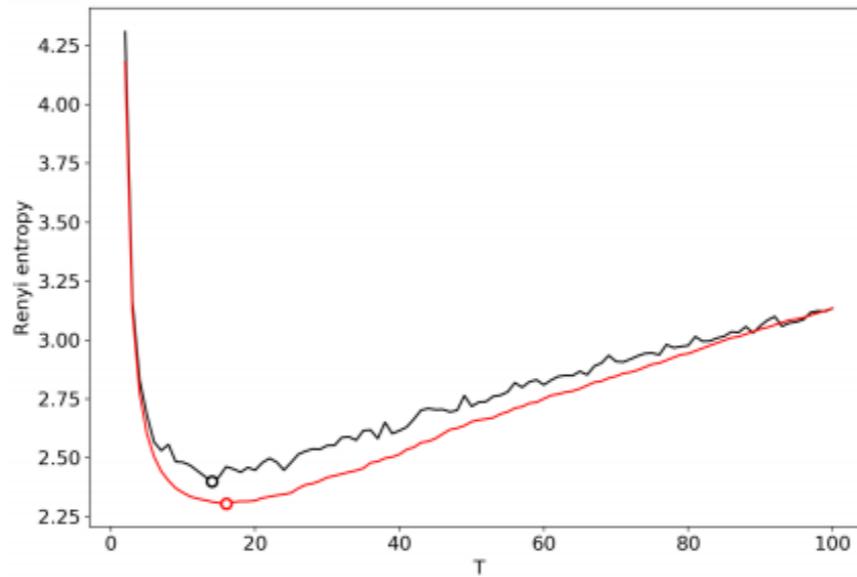
## Ренормализации в модели LDA Gibbs sampling

Алгоритм ренормализации для процедуры сэмплирования Гиббса состоит из следующих шагов:

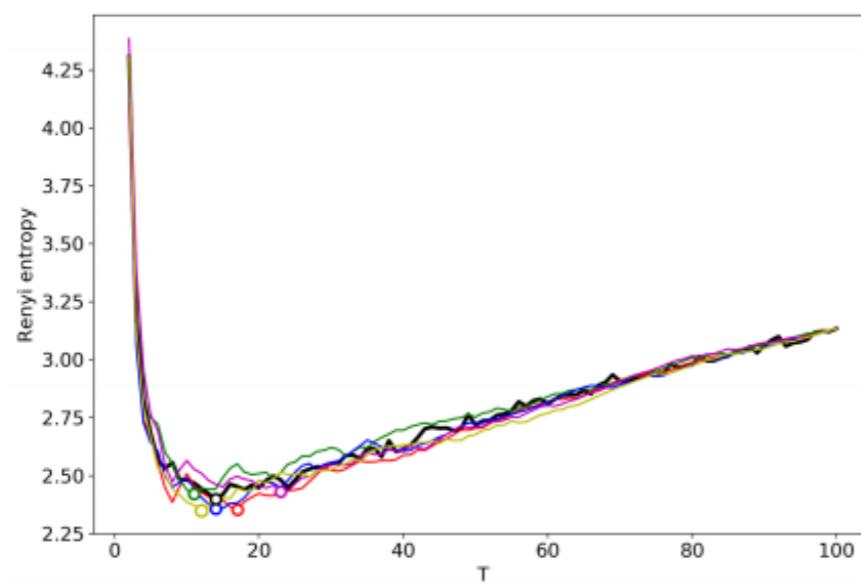
- 1. Выбор пары тем для склейки.** В рамках ренормализации мы сначала выбираем 2 темы для склеивания разными способами: Обозначим 2 выбранные темы одним из перечисленных способов за  $t_1$  и  $t_2$ .
- 2. Склейивание выбранных тем.** Склейивание выбранных заключается в суммировании частот слов  $c_{wt}$  двух выбранных тем. Затем, на основании новых значений счетчика производится расчет матрицы  $\phi_{wt}$ . Формула ренормализации модели выглядит следующим образом:  $\phi_{wt_1} := \frac{c_{wt_1} + c_{wt_2} + \beta}{(\sum_{w \in W} c_{wt_1} + c_{wt_2}) + \beta W}$ .
- 3. Перенормировка.** Новый столбец  $\phi_{\cdot t_1}$  уже удовлетворяет свойству:  $\sum_{w \in W} \phi_{wt_1} = 1$ . Далее мы удаляем столбец  $\phi_{\cdot t_2}$  из матрицы  $\Phi$ . Отметим, что на этом шаге происходит уменьшение числа тем на одну тему, то есть в конце этого шага имеем  $T - 1$  тем.

По итогам ренормализации строится кривая энтропии Ренъи, как функция от параметра ренормализации, то есть от числа тем.

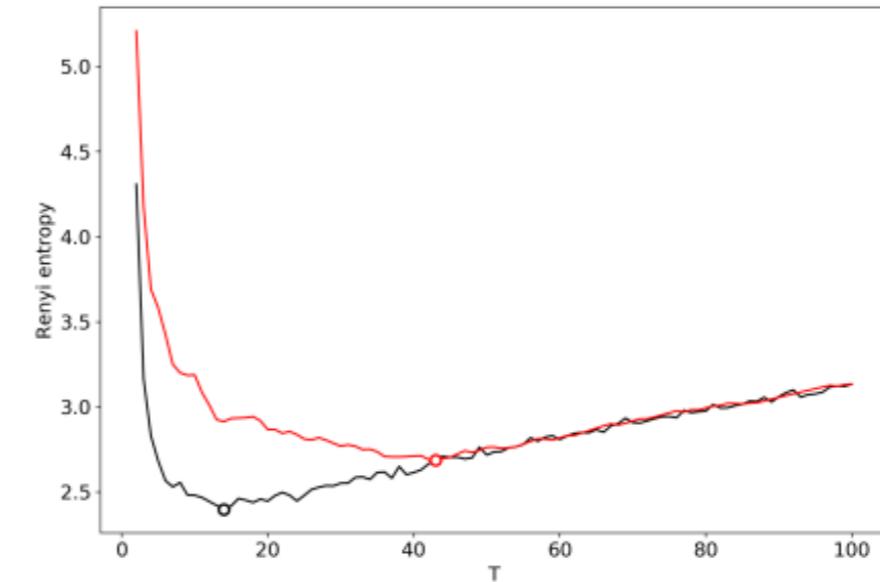
## Ренормализация модели LDA Gibbs sampling (20Newsgroups)



Объединение тем с наименьшими локальными энтропиями Ренъи.



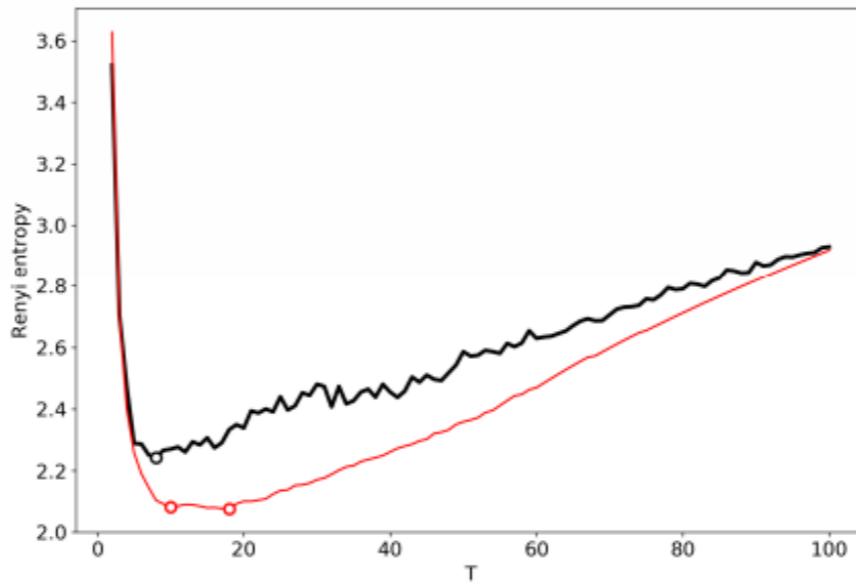
Объединение случайных тем



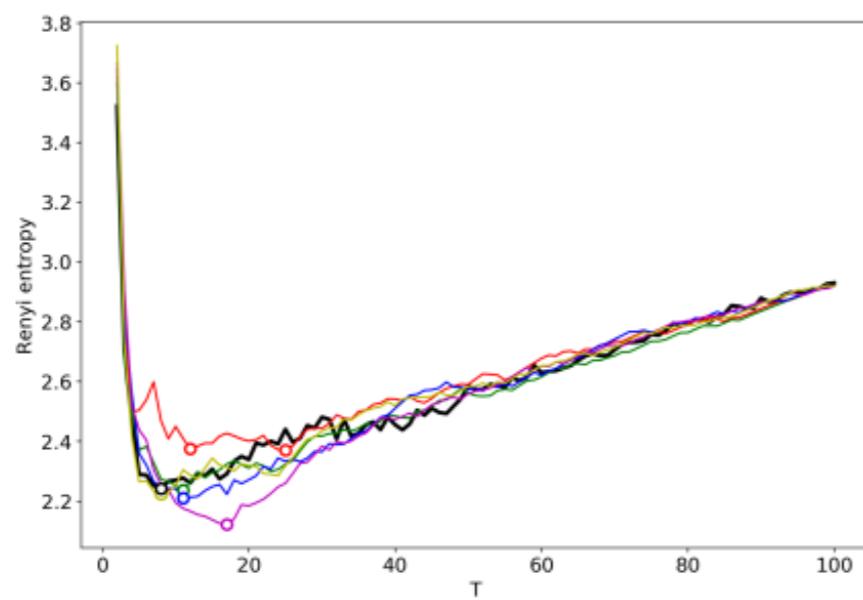
Объединение схожих тем (КЛ).



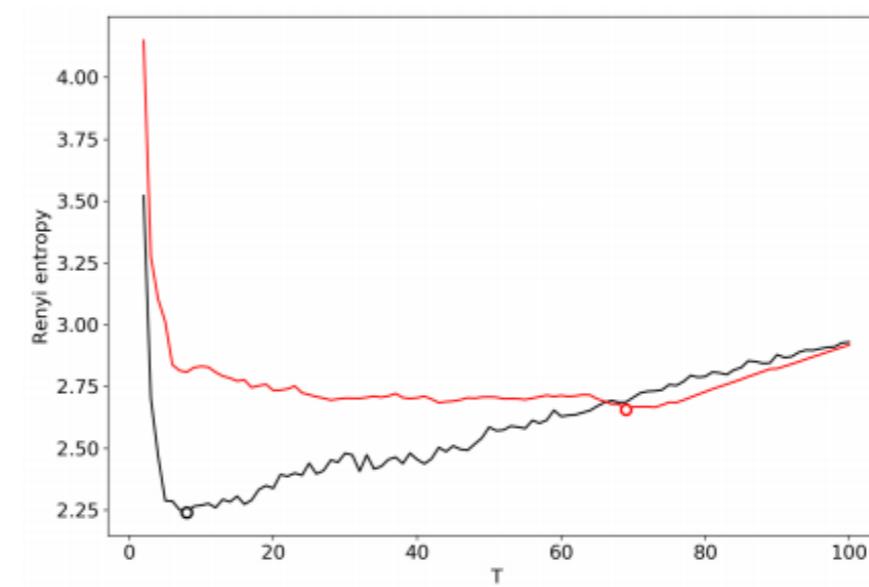
## Ренормализация модели LDA Gibbs sampling (lenta)



Объединение тем с наименьшими локальными энтропиями Ренъи.



Объединение случайных тем.



Объединение схожих тем (КЛ).



## **Renormalization of topic modeling based on variational inference (VLDA).**

Алгоритм ренормализации состоит из следующих шагов:

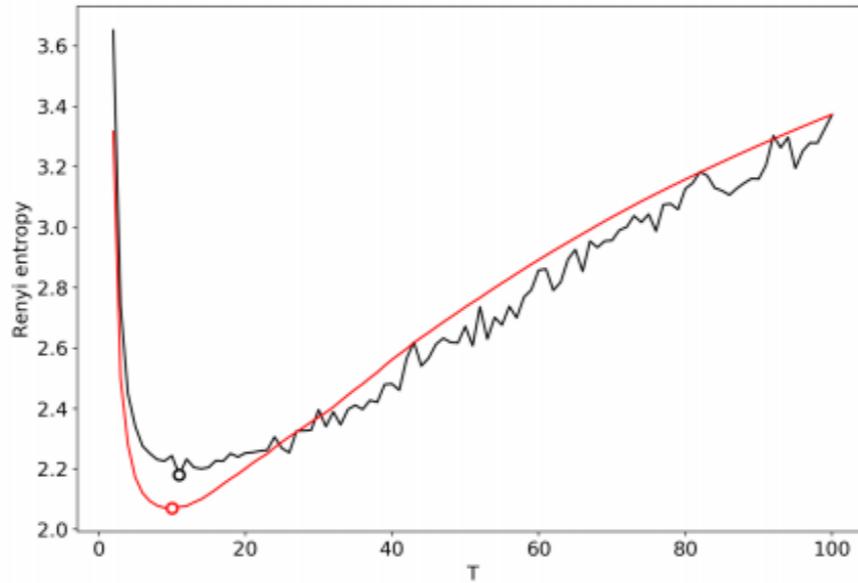
1. Выбор пары тем (колонок) для склейки. В рамках ренормализации мы сначала выбираем две темы для склеивания  $t_1$  и  $t_2$ .
2. Склейивание выбранных тем. Теперь значения распределения “новой” темы, полученной при склеивании  $t_1$  и  $t_2$ , мы записываем в столбец  $\phi_{\cdot t_1}$  :  $\phi_{wt_1} := \phi_{wt_1} \cdot \exp(\psi(\alpha_{t_1})) + \phi_{wt_2} \cdot \exp(\psi(\alpha_{t_2}))$ ,

где  $\psi(\alpha_{t_1})$  – дигамма-функция. После этого мы нормируем новый столбец  $\phi_{\cdot t_1}$  так, чтобы  $\sum_{w \in W} \phi_{wt_1} = 1$ . Также мы записываем новое значение  $\alpha_{t_1} := \alpha_{t_1} + \alpha_{t_2}$ , соответствующее “новой” теме. Затем удаляем столбец  $\phi_{\cdot t_2}$  из матрицы  $\Phi$  и элемент  $\alpha_{t_2}$  из вектора  $\alpha$ . Далее, новые значения вектора  $\alpha$  нормализуются, так что бы сумма компонент вектора = 1.

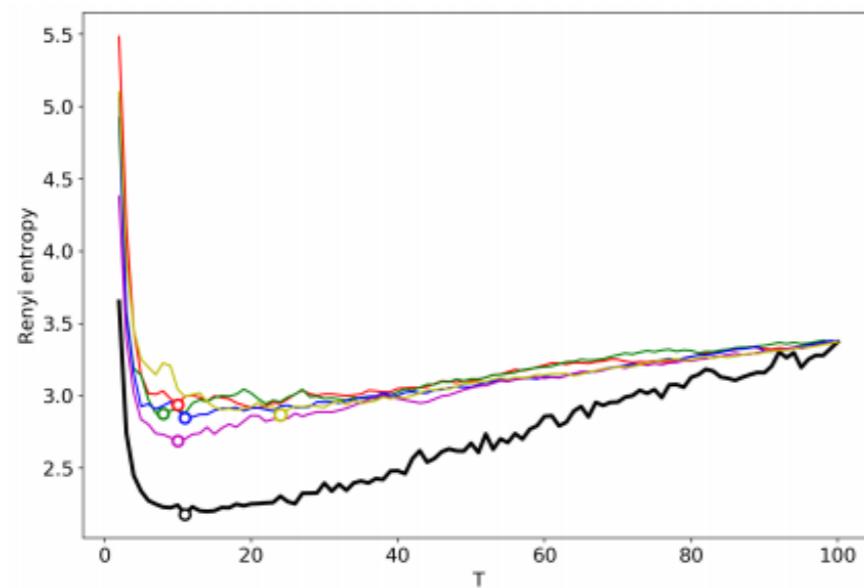
3. Расчет итогового значения глобальной величины энтропии Ренни. После того, как сформировано новое тематическое решение, рассчитывается глобальная энтропия Ренни. Шаги 1, 2, 3 повторяются до тех пор, пока не останется только 2 темы. Далее, по итогам ренормализации строится кривая глобальной энтропии Ренни, как функция от параметра ренормализации, то есть от числа тем.



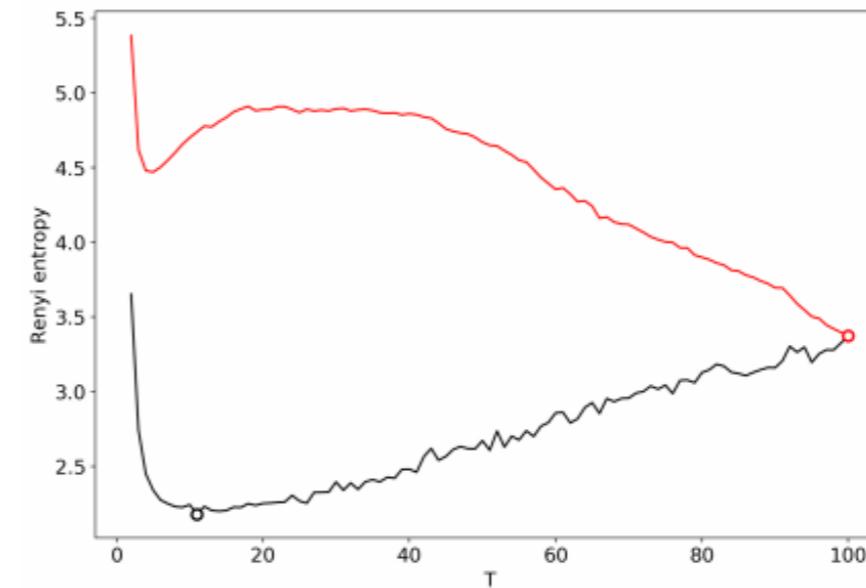
## Ренормализация модели VLDA (lenta)



Объединение тем с наименьшими локальными энтропиями Ренъи.



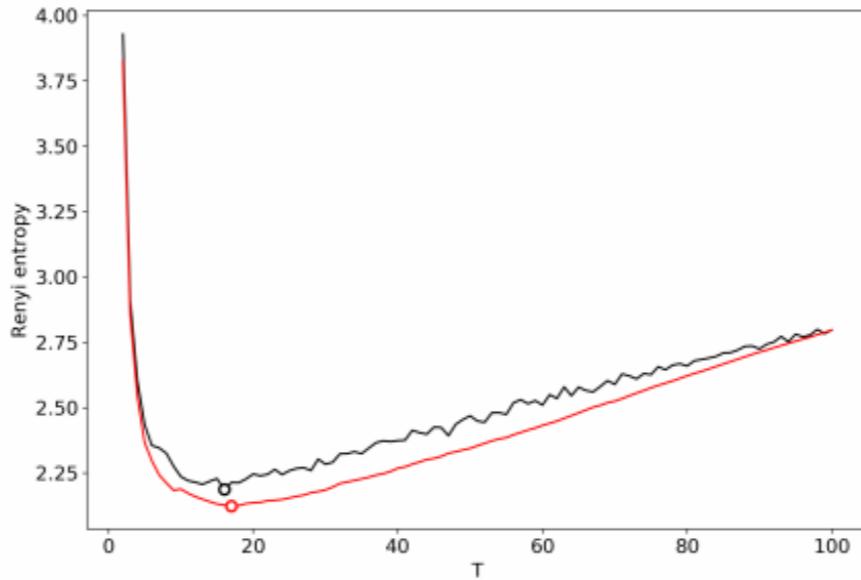
Объединение случайных тем.



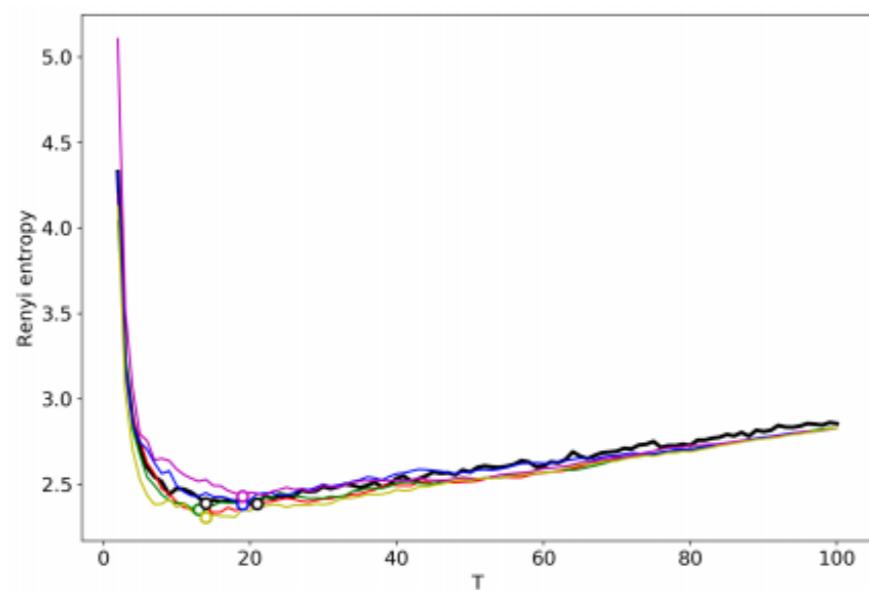
Объединение схожих тем (КЛ).



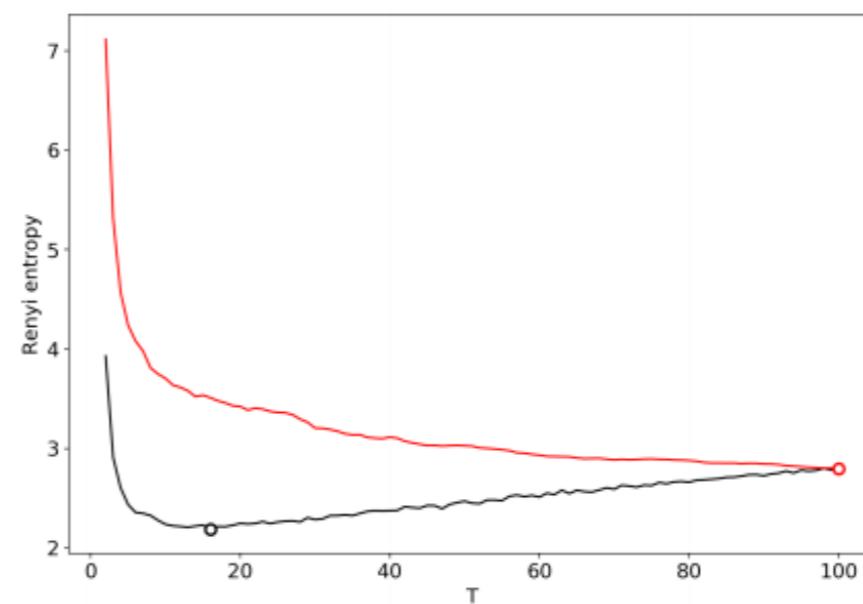
## Ренормализация модели VLDA (20Newsgroups)



Объединение тем с наименьшими локальными энтропиями Ренъи.



Объединение случайных тем



Объединение схожих тем (КЛ).



## Время расчета различных тематических моделей

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

Время расчетов для разных типов ренормализации и для последовательного тематического моделирования

1. Ренормализация ТМ может использоваться как метод быстрого приближенного поиска оптимального числа тем в текстовых коллекциях.
2. Принцип выбора тем для объединения существенно влияет на конечные результаты.
3. Показано, что ренормализация с объединением случайно выбранных тем и с объединением тем с минимальными значениями локальной энтропии Ренси ведут к удовлетворительным результатам с точки зрения точности получаемых аппроксимаций и дают существенный выигрыш по времени расчетов (20 часов для pLSA на коллекции 20Newsgroups).



# **Применение энтропийного подхода к в иерархическом тематическом моделировании**

В иерархических тематических моделях также существует проблема настройки гипер-параметров, включая число тем на разных уровнях иерархии. Кроме того, в неразмеченных датасетах неизвестно наличие плоской или иерархической структуры.



## Энтропийный подход в иерархическом тематическом моделировании

В иерархическом тематическом моделировании, при переходе от уровня к уровню происходит изменение доли слов с вероятностями выше величины  $1/W$ . Уровень определяется параметрами:

1. Число тем  $T_i$  на  $i$ -ом уровне.
2. Количество слов  $N_i = \sum_k N_{ik} (\Phi_{wik} > \frac{1}{W})$ , вероятность которых выше порога.
3. Сумма вероятностей слов  $\tilde{P} = \sum_{k=1}^{K_i^N} \Phi_{wik} (\Phi_{wik} > \frac{1}{W})$ , где каждая из вероятностей выше порога  $1/W$ . Энергия уровня:  $E_i = -\ln(\tilde{P}/T_i)$ , Энтропия уровня  $S_i = \ln(\frac{N_i}{W T_i})$ . Свободная энергия иерархического уровня выражается следующим образом:  $F_i = E_i - T_i \cdot S_i$ . Энтропия Ренни  $i$ -го уровня:  $S_i^R = \frac{F_i}{1-q}$ , где  $q = 1/T_i$  – параметр деформации, характеризующий каждый уровень иерархии.

Таким образом, измеряя величину энтропии на каждом уровне при вариации числа тем и иных гипер-параметров для конкретного датасета, можно оценить процесс построения иерархической модели с точки зрения поведения  $S_i^R$  при переходе от уровня к уровню, то есть, оценить зависимость энтропии от числа тем и значений гипер-параметров.



## Описание датасетов

Датасет **“WoS”** – это датасет, имеющий иерархическую разметку на два уровня. Доступен по ссылке <https://data.mendeley.com/datasets/9rw3vkcfy4/1>. Данный датасет содержит 46.985 аннотаций опубликованных статей, доступных в базе *Web of Science*. Первый уровень иерархической разметки содержит **7 категорий**. Второй уровень иерархической разметки содержит **134 темы** (подкатегории), каждая из которых принадлежит к одной из категорий первого уровня. **“WoS balanced”** - Балансировка датасета заключалась в удалении из коллекции тем, которые содержат менее 260 документов, то есть удалялись плохо представленные темы. Число тем: первый уровень иерархической разметки содержит 7 категорий, второй уровень – 33 темы.

Датасет **“Amazon”** – это датасет, имеющий иерархическую разметку на три уровня. Доступен по ссылке <https://www.kaggle.com/kashnitsky/hierarchical-text-classification/version/1>. Первый уровень иерархической разметки содержит **6 категорий**, второй уровень – **64 подкатегории** и третий уровень – **510 классов**. Так как третий уровень содержит пустые метки, в данной работе рассматриваются только первые два уровня иерархической разметки. Сбалансированный датасет **“Amazon balanced”** – подмножество датасета “Amazon”, которое содержит только подкатегории с числом документов более 500. Таким образом, первый уровень содержит **6 категорий**, а второй уровень – **27 подкатегорий**.



## Модель hLDA

Датасет ‘**lenta**’  
10 запусков

eta	Число тем (второй уровень)	Число тем (третий уровень)
0.001	6 - 11	31 - 67
0.01	6 - 11	13 - 30
0.2	2-3	5- 7
0.3	2	2- 4
0.5	2	2-3
0.7	2	2-3
1	3	2-3

Датасет ‘**20Newsgroups**’  
10 запусков

eta	Число тем (второй уровень)	Число тем (третий уровень)
0,001	288-358	911-1402
0,01	81-11	274-334
0,2	6 - 11	14-18
0,3	4 - 9	7- 11
0,5	3 - 5	5 - 9
0,7	3-4	3-7
1	2-4	3-6



## Модель hLDA

Датасет ‘**WoS balanced**’,  
10 запусков

eta	Число тем (второй уровень)	Число тем (третий уровень)
0,001	482- 652	1751- 2242
0,01	68-93	325 - 453
0,2	2-5	6- 13
0,3	2-3	3-7
0,5	2	2-3
0,7	2	2-3
1	2	2-3

Датасет ‘**Amazon balans**’  
10 запусков

eta	Число тем (второй уровень)	Число тем (третий уровень)
0,001	108-148	561 - 654
0,01	23-36	108-122
0,2	3	5-6
0,3	2 -3	3- 4
0,5	2-3	3-4
0,7	2	2-4
1	2	2-3



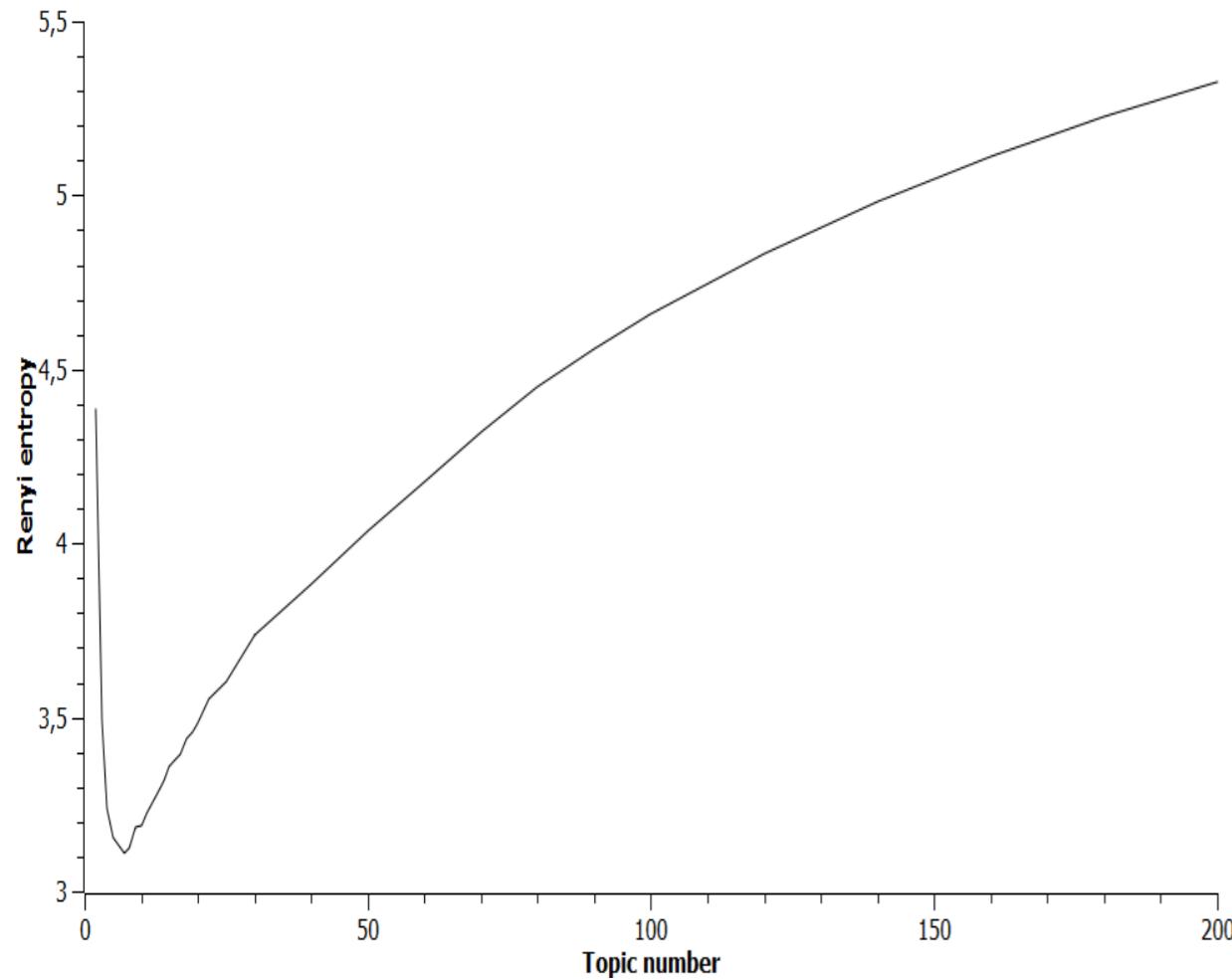
## **Модель hARTM – это иерархическая версия подхода аддитивной регуляризации тематических моделей (ARTM)**

Модель hARTM – это иерархическая версия подхода аддитивной регуляризации тематических моделей (ARTM). В рамках hARTM тема может иметь несколько родительских тем. Число тем на каждом уровне иерархии должно быть задано пользователем. Для построения иерархии обучаются несколько плоских тематических моделей, а затем они связываются через регуляризацию.

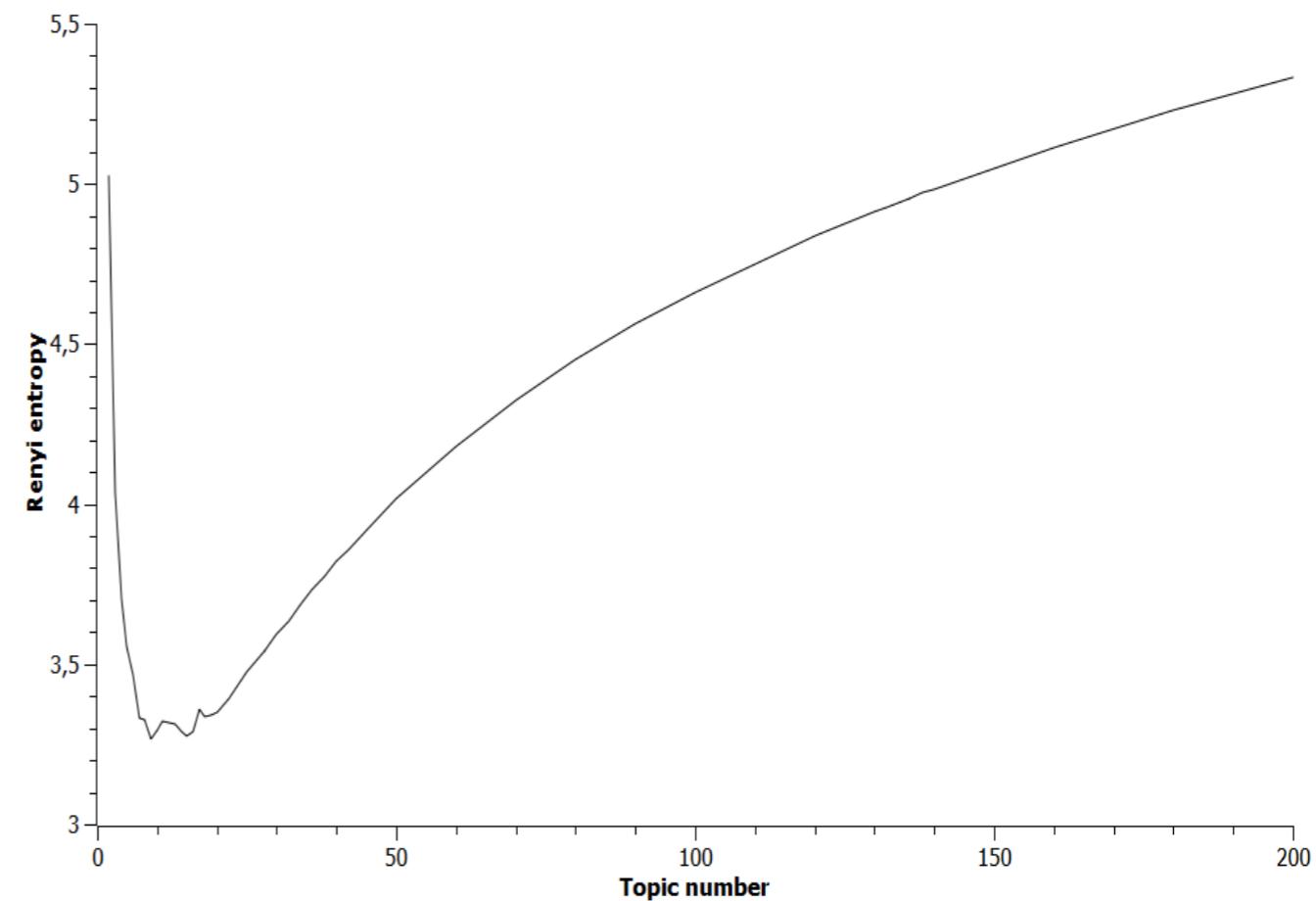
В рамках исследования данной модели проводилось ТМ на 4 датасетах. При этом варьировалось число тем на втором и третьем уровне при разных параметрах инициализации (seed).



## Модель hARTM ('lenta' и '20Newsgroups')



Зависимость энтропии Ренъи от числа тем на  
первом уровне иерархии в модели hARTM  
(‘lenta’)

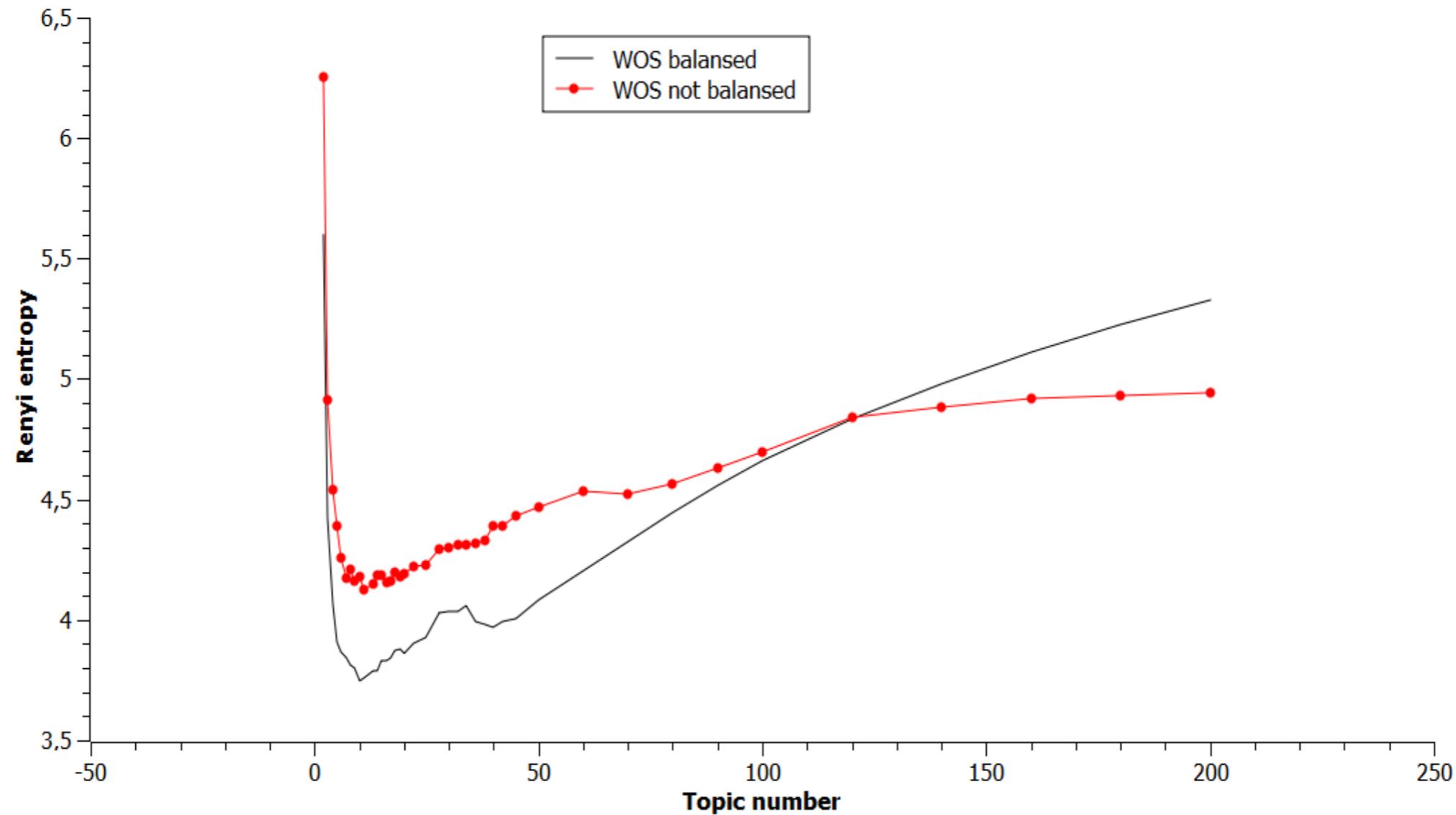


Зависимость энтропии Ренъи от числа тем на  
втором уровне иерархии в модели hARTM  
(‘20Newsgroups’)

Модель hARTM хорошо видит плоскую структуру датасетов.



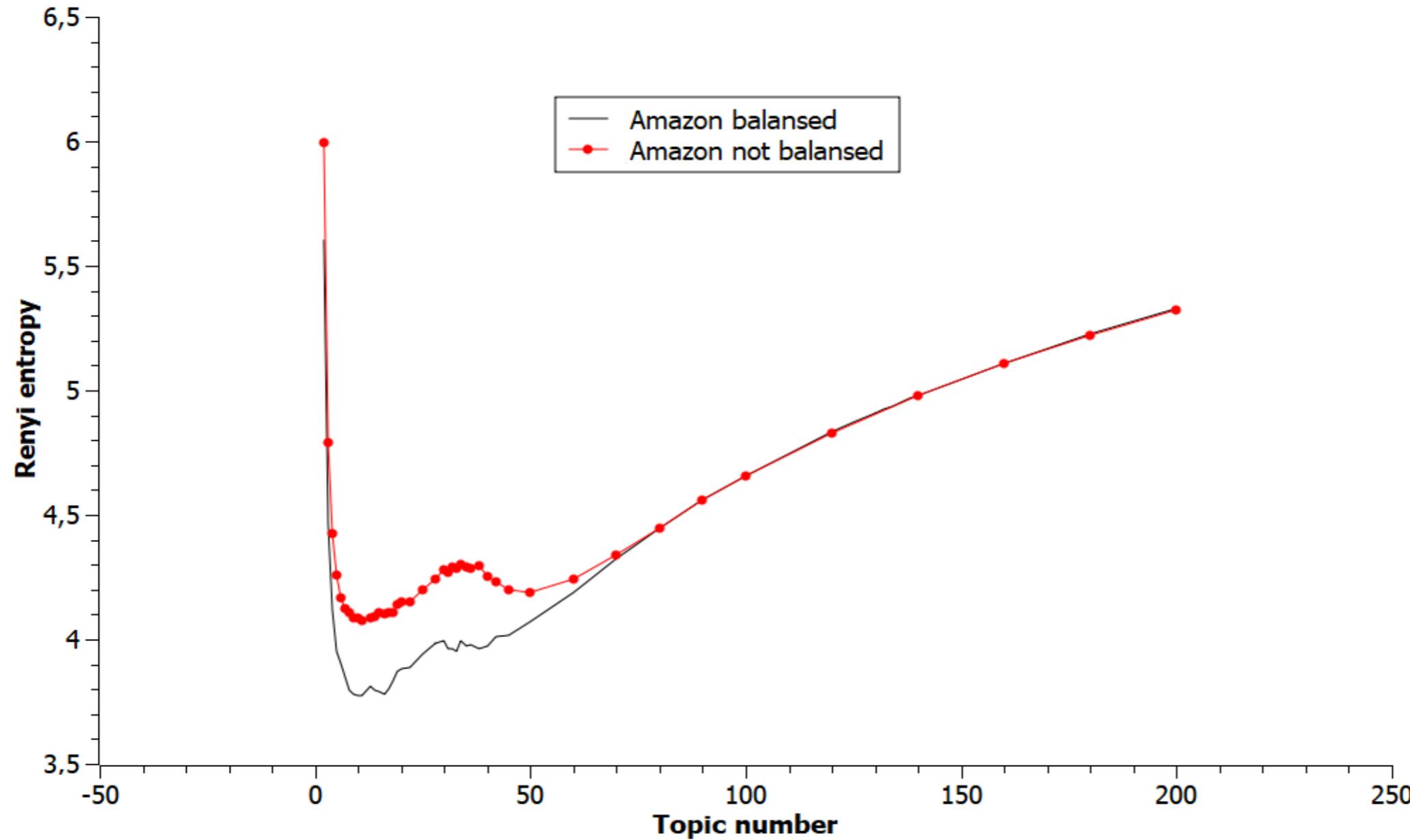
## Модель hARTM ('WoS balanced' и 'WoS')



Кривые энтропии Ренъя для сбалансированного и не сбалансированного датасета ('WOS') на первом уровне. Чёрный цвет – сбалансированный датасет, красный цвет – несбалансированный датасет.



## Модель hARTM ('Amazon balans' и 'Amazon')



Кривые энтропии Ренъи для сбалансированного и не сбалансированного датасета 'Amazon'. Красный цвет – не сбалансированный датасет, черный цвет – сбалансированный датасет.



## Granulated topic model with word embeddings

The GLDAW model is realized with Gibbs sampling algorithm as follows. There are three stages of the algorithm.

1. We form a matrix of the nearest words by given word embeddings.

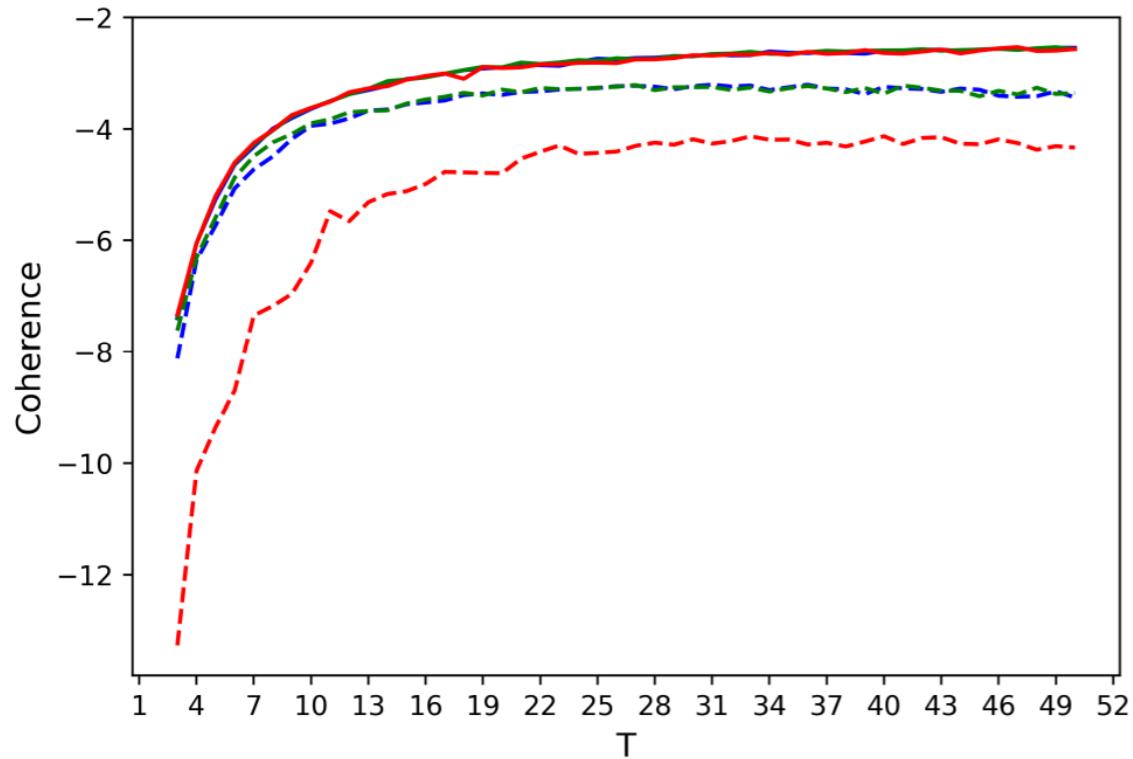
At this stage, the algorithm checks if the vocabularies of the dataset and word embeddings match. If a word from the given set of word embeddings is missing from the dataset's vocabulary, then its embedding is deleted. This procedure reduces the size of the set of word embeddings and speeds up the computation at the second stage. The number of the nearest words is set manually by the “window” parameter.

2. the computation is similar to Granulated LDA Gibbs sampling with the choice of an anchor word from the text and the attachment of this word to a topic. The topic is computed based on the counters. in this algorithm, the counters of the words corresponding to the nearest embeddings are increased.

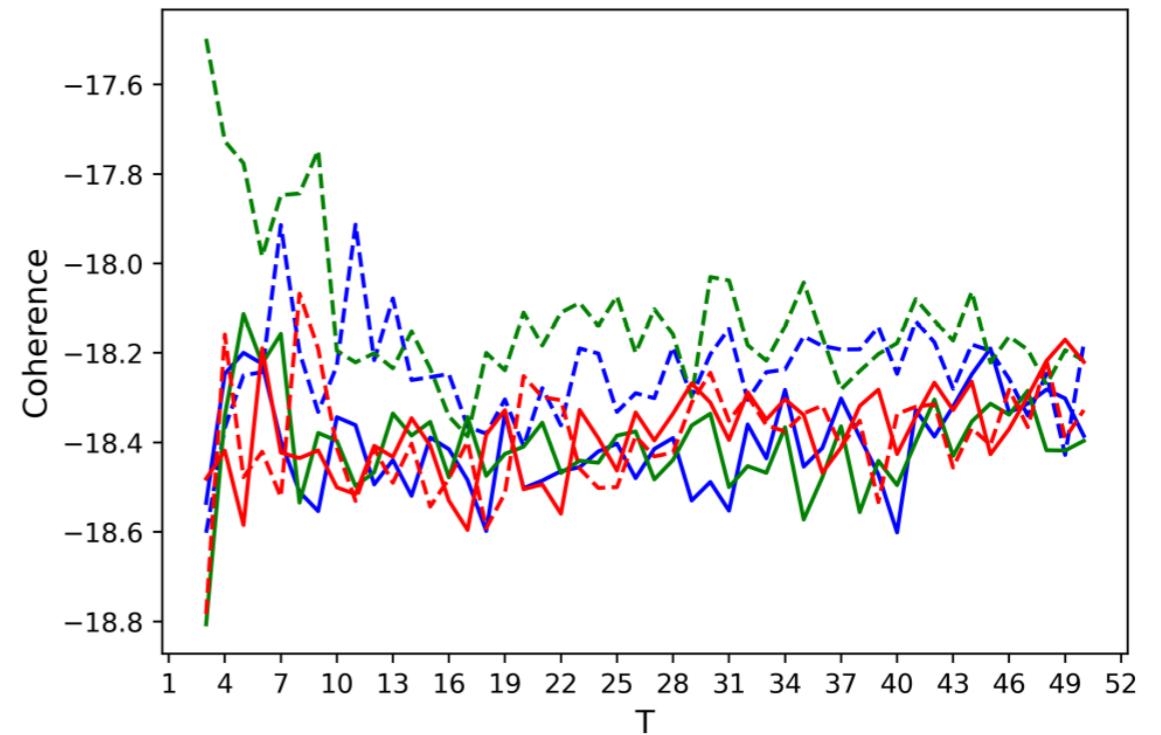
3. the resulting matrix of counters is used to compute the probabilities of all words as in the standard LDA Gibbs sampling:

$$\phi_{wt} = \frac{n_{wt} + \beta}{n_t + \beta W}, \quad \theta_{td} = \frac{n_{td} + \alpha}{n_d + \alpha T},$$

## Coherence ETM model

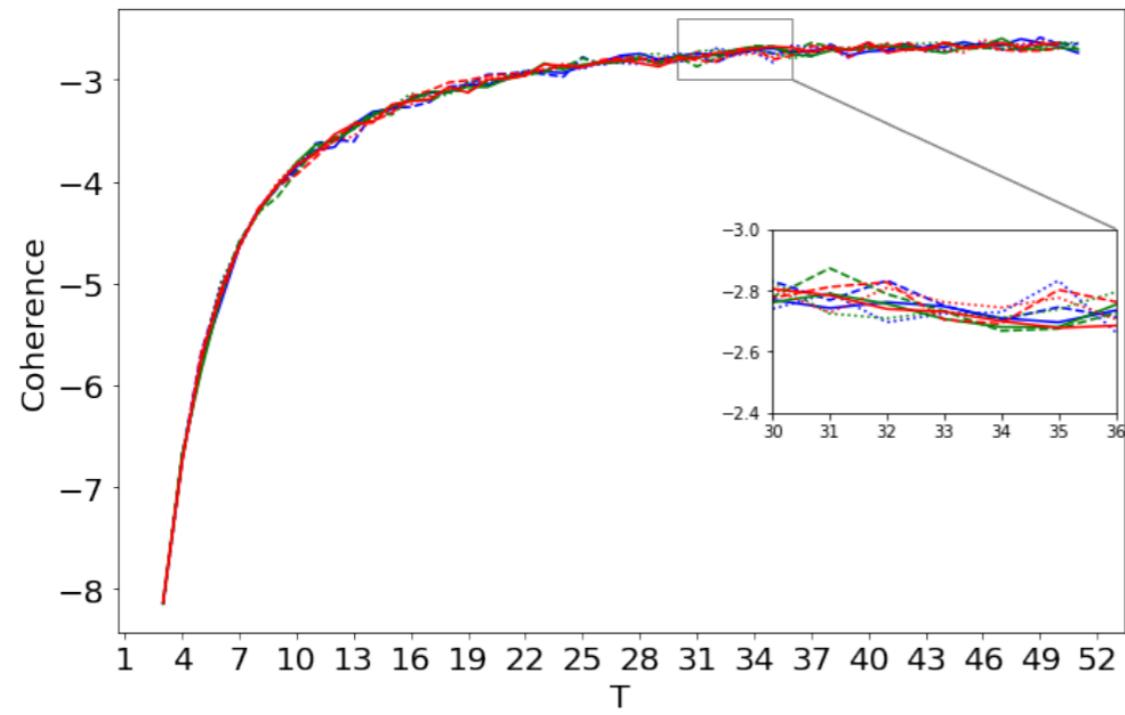


(a) Lenta dataset (first level of pre-processing).  
rus\_vectors\_embeddings - red, Navec - green,  
300\_wiki\_embeddings - blue. Pre-trained em-  
beddings - dash line; additionally trained em-  
beddings - solid line.

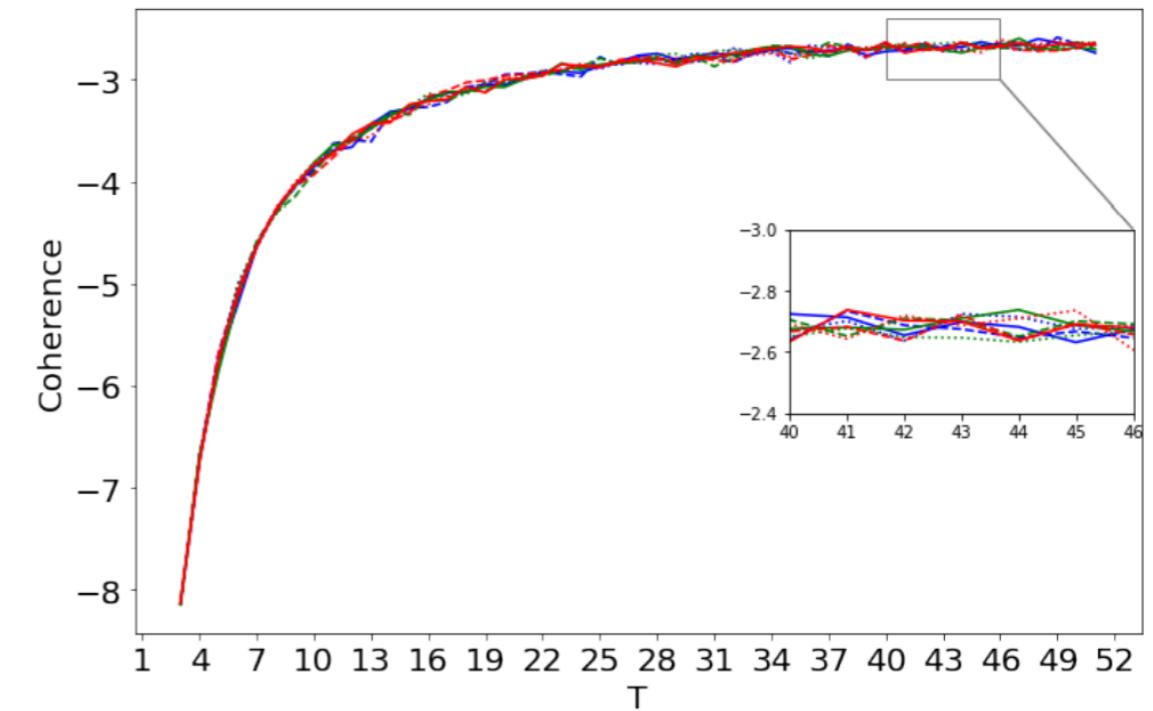


(b) Lenta dataset (second level of pre-  
processing). rus\_vectors\_embeddings - red,  
Navec - green, 300\_wiki\_embeddings - blue. Pre-  
trained embeddings - dash line; additionally  
trained embeddings - solid line.

## Coherence GLDAW model

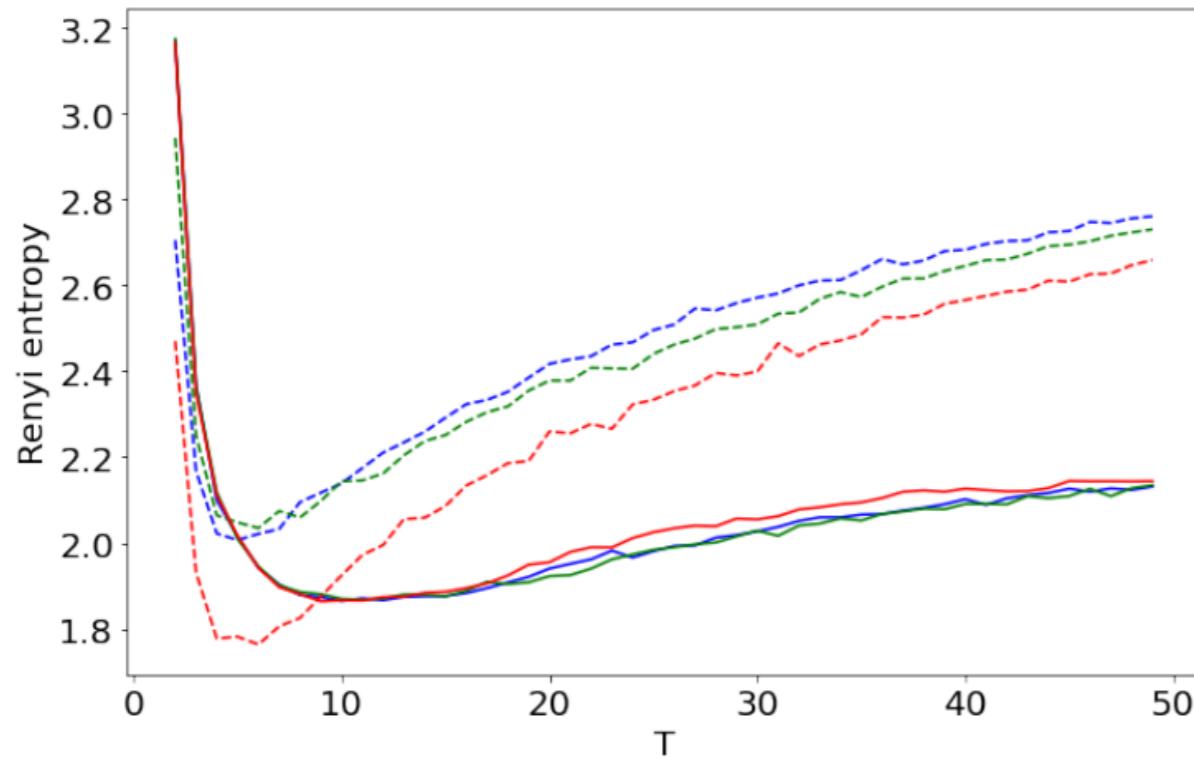


(a) Lenta dataset (first level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Window of 10 words - solid line, window of 50 words - dash line, window of 100 words - dotted line.

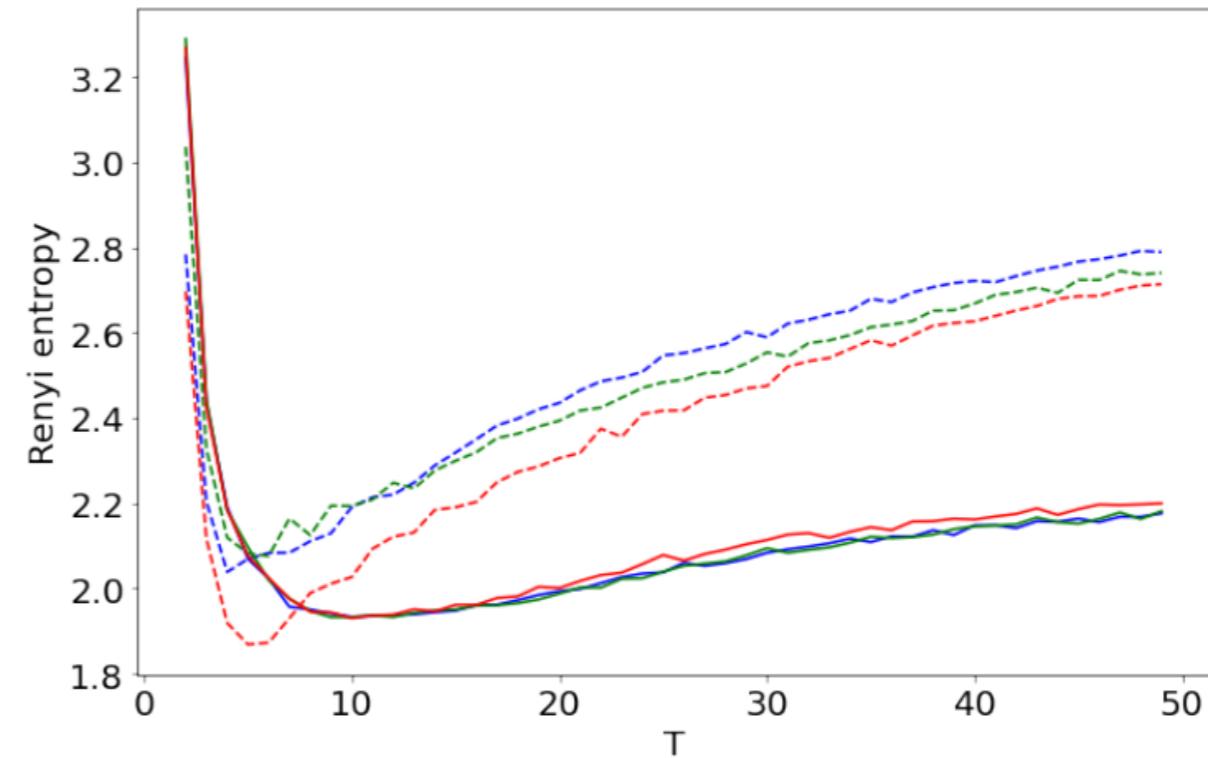


(b) Lenta dataset (second level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Window of 10 words - solid line, window of 50 words - dash line, window of 100 words - dotted line.

## Renyi entropy ETM model



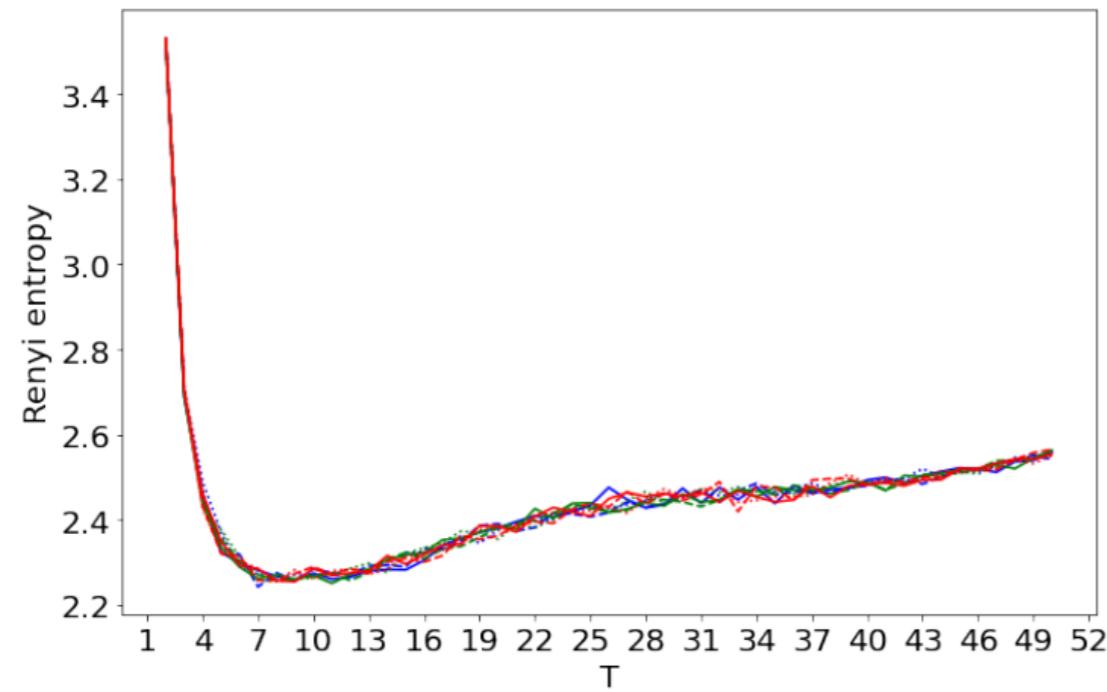
(a) Lenta dataset (first level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Pre-trained embeddings - dash line; additionally trained embeddings - solid line.



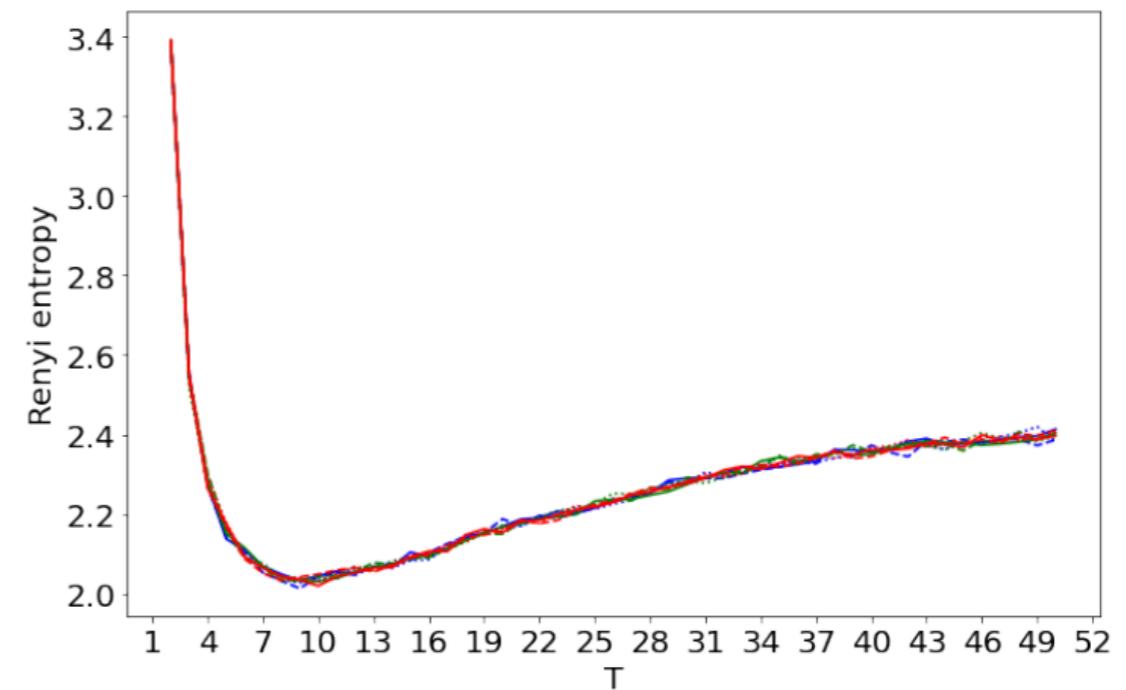
(b) Lenta dataset (second level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Pre-trained embeddings - dash line; additionally trained embeddings - solid line.



## Renyi entropy GLDAW model



(a) Lenta dataset (first level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Window of 10 words - solid line, window of 50 words - dash line, window of 100 words - dotted line.



(b) Lenta dataset (second level of pre-processing). rus\_vectors\_embeddings - red, Navec - green, 300\_wiki\_embeddings - blue. Window of 10 words - solid line, window of 50 words - dash line, window of 100 words - dotted line.



## Stability

Модель	Датасет 'Lenta'. Решение на 10 тем.	Датасет '20 topics news'. Решение на 20 тем.
Embedded topic model (ETM)	6 тем	16 тем
<b>Granulated topic model with word embedding (GLDAW)</b>	<b>8 тем</b>	<b>18 тем</b>
Model Gaussian Softmax distribution (GSM)	2 темы	7 тем
Model WAE with Gaussian Mixture prior with Gaussian Softmax (WTM-GMM)	3 темы	14 тем
Model WAE with Dirichlet prior with Gaussian Softmax (WTM-MMD)	1 тема	12 тем



NATIONAL RESEARCH  
UNIVERSITY

# Thank you for your attention

<https://linis.hse.ru/>

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru