



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

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



ТЕРМОДИНАМИЧЕСКИЙ ВЗГЛЯД НА МАШИННОЕ ОБУЧЕНИЕ

Koltcov S.

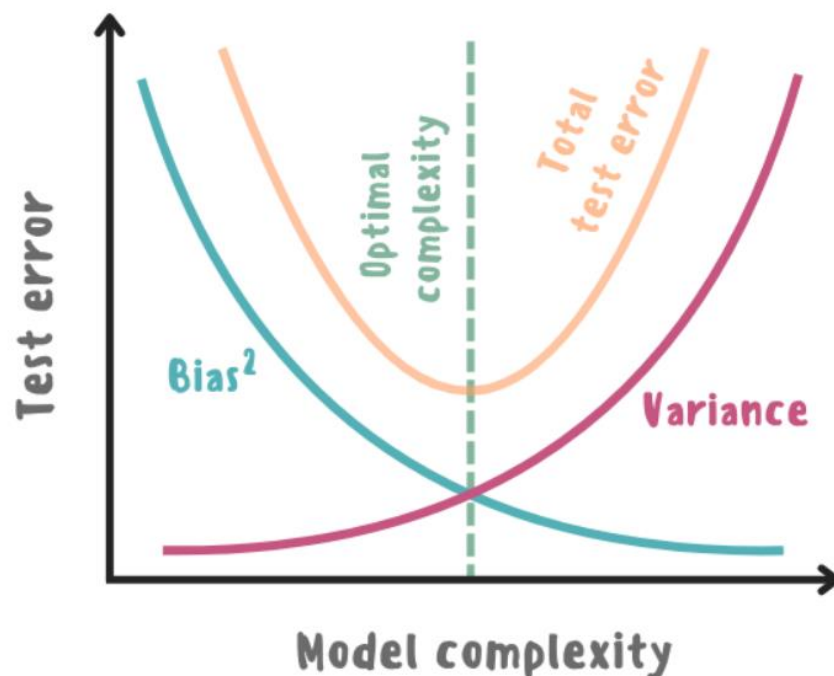
Saint Petersburg, 2025

U-образная кривая MSE (Mean Squared Error)

Когда мы строим график среднеквадратичной ошибки (MSE) на обучающей выборке и тестовой выборке в зависимости от сложности модели, получается характерная U-образная форма:

$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

Общая ошибка прогноза модели состоит из квадрата смещения (насколько модель ошибается в среднем), разброса (как предсказания модели различаются при разных данных) и неустраняемой ошибки из-за шума в данных.



Баланс между **переобучением** (высокая точность на тренировочных данных и низкая на тестовых) и **недообучением** (неспособность модели улавливать закономерности в тренировочных данных, что приводит к низкой точности как на них, так и на тестовых).

оси X — сложность модели (например, степень полинома, количество параметров, число деревьев в ансамбле, число кластеров, нейронов и так далее).

По оси Y — значение MSE.

1. Низкая сложность модели (Underfitting)

Модель слишком простая, чтобы уловить закономерности в данных. Обучающая и тестовая ошибки высокие. Модель плохо справляется и с обучением, и с предсказанием.

2. Средняя сложность модели (Хорошая подгонка)

Модель достаточно сложна, чтобы ловить основные паттерны, но не переобучается.

Обе ошибки — низкие.

Это золотая середина: хороший баланс между смещением (bias) и дисперсией (variance).

3. Высокая сложность модели (Overfitting)

Модель "зазубривает" данные, включая шумы и случайные флуктуации.

Обучающая ошибка падает до очень низких значений (почти идеальное совпадение с данными).

Тестовая ошибка начинает расти, потому что модель не обобщает.

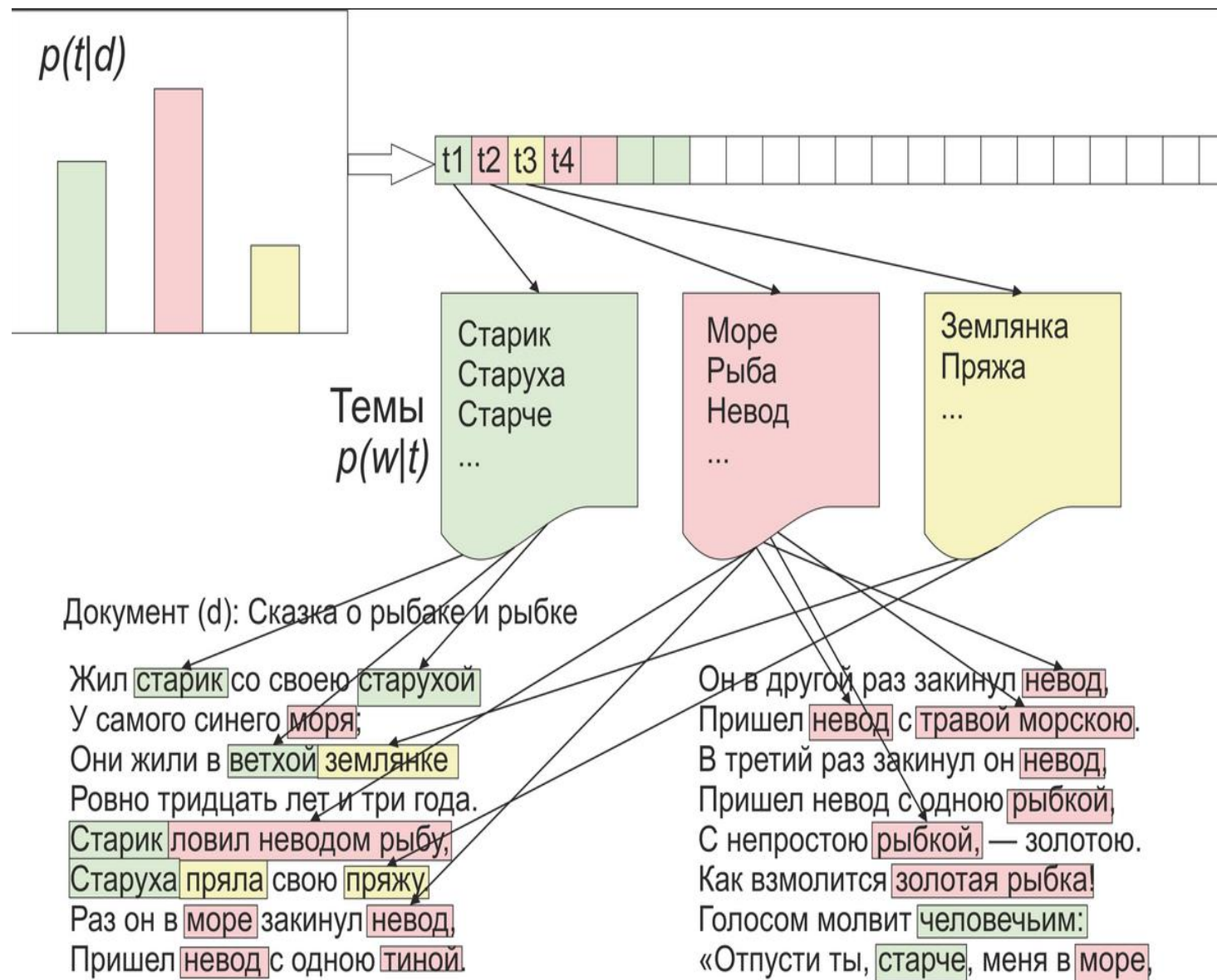


Термодинамический взгляд на проблему

Набор данных, например, коллекция документов является мезоскопической информационной системой, которая состоит из множества элементов (слов, документов, нейронов и так далее), поэтому поведение такой информационной системы можно изучать при помощи методов из статистической физики. Более того, такие информационные системы не являются закрытыми, так как происходит обмен информацией с окружающей средой: например, пользователь может менять число тем/кластеров, варьировать число деревьев и другие гипер-параметры. Соответственно, такая информационная система может не достигать равновесного состояния в смысле максимума энтропии Гиббса–Шеннона, но может находиться в промежуточном состоянии. **Весь вопрос в том как это можно измерять.** Когда мы берем данные мы мало что о них знаем, соответственно энтропия/свободная энергия может быть максимальной или минимальной в зависимости от знака. Применяя тот или иной алгоритм машинного обучения на наших данных, мы меняем состояние информационной системы. Соответственно, поведение информационной системы при вариации гиперпараметров можно сопоставить с поведением той или иной термодинамической функции.

Классическими функциями являются энтропия Шеннона, внутренняя энергия, свободная энергия, энтропия Реньи, энтропия Шарма – Миттала. В рамках данного доклада будут рассмотрены две задачи: 1. **тематическое моделирование**. 2. **Pruning NN**

Probabilistic problem statement for topic modeling



Topic model — модель коллекции текстовых документов, определяющая, к каким темам относится каждый документ в коллекции. Алгоритм построения тематической модели: на вход модели поступает коллекция текстовых документов. На выходе для каждого документа получается числовой вектор, состоящий из оценок степени принадлежности этого документа к каждой из тем.



Probabilistic problem statement for topic modeling

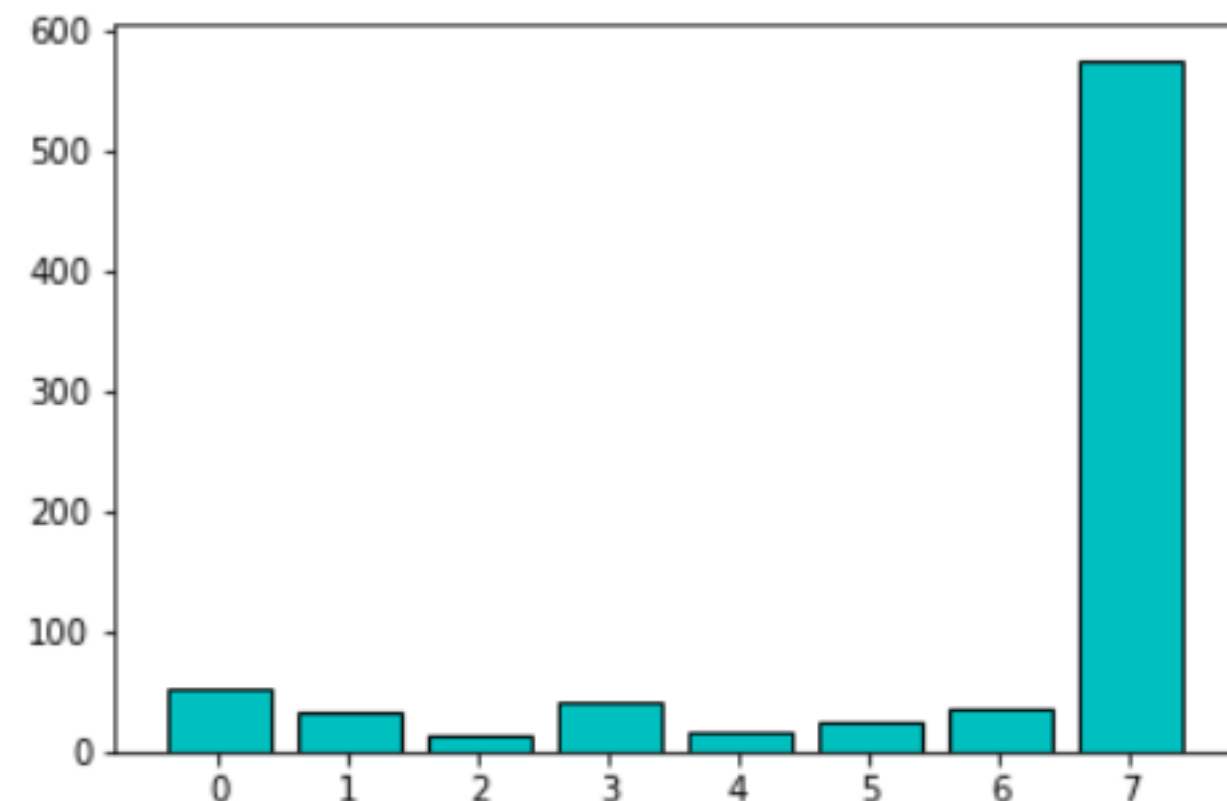
Пусть D - набор текстовых документов, а W - множество (словарь) всех уникальных слов. Каждый документ $d \in D$ - это последовательность терминов $w_1, \dots, w_d$ из словаря (W). Мы также предполагаем, что существует конечное множество тем T , и каждое вхождение слова w в документ d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как случайная и независимая выборка троек (w_i, d_i, t_i) , $i = 1, \dots, n$ из дискретного распределения $p(w, d, t)$ на конечном вероятностном пространстве $W \times D \times T$. Слова w и документы d - наблюдаемые переменные, субъект $t \in T$ - латентная (скрытая) переменная. **Предполагается, что порядок терминов в документах не важен для идентификации субъекта (мешок слов). Порядок документов в коллекции также не имеет значения.**



Matrix - topics distribution in documents

[8]:

	Topic0	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	dominant_topic
Doc0	0.010000	0.010000	0.920000	0.010000	0.010000	0.010000	0.010000	2
Doc1	0.010000	0.010000	0.010000	0.010000	0.670000	0.010000	0.280000	4
Doc2	0.010000	0.010000	0.790000	0.010000	0.010000	0.010000	0.140000	2
Doc3	0.020000	0.020000	0.020000	0.020000	0.020000	0.020000	0.900000	6
Doc4	0.940000	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0
Doc5	0.010000	0.010000	0.010000	0.010000	0.010000	0.010000	0.910000	6
Doc6	0.010000	0.010000	0.010000	0.010000	0.640000			
Doc7	0.010000	0.010000	0.010000	0.010000	0.010000			
Doc8	0.020000	0.020000	0.020000	0.020000	0.020000			
Doc9	0.010000	0.630000	0.010000	0.010000	0.010000			
Doc10	0.010000	0.010000	0.010000	0.010000	0.010000			
Doc11	0.020000	0.020000	0.020000	0.020000	0.020000			
Doc12	0.010000	0.010000	0.770000	0.010000	0.010000			
Doc13	0.020000	0.020000	0.020000	0.020000	0.910000			
Doc14	0.010000	0.010000	0.010000	0.930000	0.010000			





Matrix Φ – word distribution in topics

[75]:

	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Word 0	что - 0.0004698091106239771	турции - 0.0003081567456290563	что - 0.0005595929984206749	кино - 0.0008377494184893066	миронова - 0.00038785660283128723	тамара - 0.0002899678330191077
Word 1	было - 0.00045069188824187873	новости - 0.00026218325858764395	для - 0.0004909673222097849	актёров - 0.0006882717680612737	что - 0.0003523639008715388	эвра - 0.0002814785833549309
Word 2	сама - 0.0003799606292472247	риа - 0.000260257133978759	только - 0.000415914812551397	кто - 0.0006281071291426114	платт - 0.0003287332341951211	меньшов - 0.00027926902209925825
Word 3	гостями - 0.0003733567480631264	евстигнеев - 0.0002398973477305586	больше - 0.00038533637021255035	программа - 0.0006129727273055995	михайлов - 0.00032093511409124026	прохоренко - 0.00027534367118077073
Word 4	становятся - 0.00037335455688781603	россиян - 0.00023919681240274778	крымского - 0.00036362016932275297	жизни - 0.0006116706036697333	теличкина - 0.0003084961624621251	её - 0.000271491447661209
Word 5	люди - 0.00036664935894809535	nacional - 0.0002301310751217433	волчек - 0.0003462529584878737	что - 0.0005900216353031027	георгий - 0.00026916693703453746	пирог - 0.0002561178331886546
Word 6	обществе - 0.0003566117509430164	el - 0.00023012959928839767	момента - 0.0003418114960347403	тех - 0.000554808273682613	павел - 0.00026774550827918497	зелёная - 0.0002560817262403854
Word 7	чьа - 0.0003565912094529661	игил - 0.00021421495451066618	уно - 0.0003418045940863241	нашего - 0.000541571646685365	никогда - 0.00026303024040720085	рина - 0.00025608118940324365
Word 8	личность - 0.00035658720458574473	ван - 0.0002018472458170297	дейр - 0.00033838791544775	ссср - 0.0005348414317714289	большая - 0.00025981700043984133	рины - 0.0002560801109106136
Word 9	вызывают - 0.0003565816472557962	болдуина - 0.00019822969960403715	борисов - 0.00029495745286122865	авторской - 0.0004785189503699558	мало - 0.00025897457522792506	сёмина - 0.00022792832325523838
Word 10	его - 0.00033809876867681556	алеппо - 0.00019102189597331812	украины - 0.00029426321946377225	людях - 0.0004717950638932046	был - 0.0002484538784704299	первого - 0.00022549558660126703
Word 11	ведущих - 0.0003329551412600506	жан - 0.00019073057360025886	мост - 0.00029404605645420325	всегда - 0.00046309333331577866	как - 0.0002483347595859574	алексиевич - 0.00022017162369273883
Word 12	интерес - 0.00032457711727933766	новогоднюю - 0.00018736058971367334	вертолет - 0.00029095614132396076	программы - 0.0004629812043686788	героя - 0.00024431115075935903	нефти - 0.0002183796027773021



Mathematical model for topic modeling

Вероятность $p(w|d)$ появления терминов w в документах d может быть выражена как произведение распределений $p(w|t)$ и $p(t|d)$. Согласно формуле полной вероятности и гипотезе условной независимости, мы имеем следующее выражение:

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

где $p(w|t)$ - распределение слов по темам, а $p(t|d)$ - распределение документов по темам. Построение тематической модели данных означает решение обратной задачи, в которой необходимо найти множество скрытых тем T , то есть множество одномерных условных распределений $p(w|t) \equiv \phi(w, t)$ для каждой темы t , составляющих матрицу Φ (распределение слов по темам), и множество одномерных распределений $p(t|d) \equiv \theta(t, d)$ (матрица θ , распределение документов по темам) для каждого документа d на основе наблюдаемых переменных d и W .



The main directions of topic modeling

1. Модели, основанные на методе максимизации логарифмического правдоподобия. Математическая формулировка этого семейства моделей заключается в том, что матрица распределения слов для документов F представляется как произведение матрицы распределения слов для тем Φ и матрицы распределения тем для документов Θ . Таким образом, задача $F = \Phi\Theta$ сводится к поиску стохастического разложения матрицы, которое выполняется с помощью алгоритма Е-М. Для нахождения приближенного решения максимизирует логарифм правдоподобия.

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

2. Модели, основанные на методе Монте-Карло. При таком подходе вероятность распределения слов и документов можно оценить с помощью математического ожидания, если знать подынтегральные функции. В качестве подынтегральных функций используются мультиномиальные функции и функции Дирихле. Скрытые распределения определяются с помощью процедуры сэмплирования Гиббса.

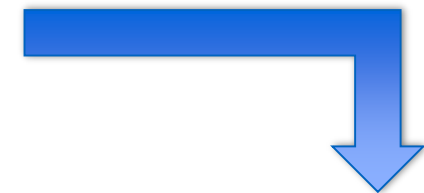
$$p(w | z, \beta) = \int p(w | z, \Phi) p(\Phi | \beta) d\Phi \quad p(z | \alpha) = \int p(z | \Theta) p(\Theta | \alpha) d\Theta$$



Recovery of matrices Φ and Θ by maximizing the likelihood function using the E-M algorithm (PLSA).

Задача максимального правдоподобия формулируется следующим образом. Заменяя $p(w|t)$ и $p(t|d)$ матрицами ϕ_{wt} , θ_{td} и прологарифмировав выражение, чтобы избавиться от произведения, мы получим выражение, в котором нужно подобрать значения матриц так, чтобы это выражение было максимальным.

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



$$\sum_{w \in W} \phi_{wt} = 1$$

$$L(\phi, \theta) = \sum_{d \in D} \sum_{w \in d} n_{wd} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max$$

Поиск максимального значения функции $L(\phi, \theta)$ осуществляется с помощью EM-алгоритма.



Latent Dirichlet Allocation (LDA)

Идея Блея заключается в следующем. Он предположил, что столбцы матрицы ϕ_{wt} и строки θ_{td} в модели PLSA генерируются распределениями Дирихле. Каждое распределение характеризуется своим параметром α (для слов) и β (для документов), т. е. все параметры α и β различны. В модели также предполагается, что вероятность встречи с каждой темой соответствует мультиномиальному распределению:

Multinomial distribution	$P(x p) = \frac{n!}{\prod_{i=1}^K x_i!} \prod_{i=1}^K p_i^{x_i}$	Dirichlet distribution	$P(p; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^K p_i^{\alpha_i - 1}$
-----------------------------	--	---------------------------	---

Эти предположения значительно упрощают байесовский вывод модели благодаря свойствам сопряженности распределений Дирихле с мультиномиальным распределением. Модель PLSA с функциями Дирихле преобразуется в следующую модель LDA:

$$\theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}$$

$$\phi_{tw} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}$$



LDA на основе Gibbs sampling

Термины: D - пространство документов, W - пространство слов, Z - пространство тем. Темы - это скрытые параметры, которые необходимо найти. И оценка основывается на двух вещах: 1. Оценка проводится в соответствии с ожиданиями. 2. Используются мультиномиальные функции и функции Дирихле.

$$p(w | z, \beta) = \int p(w | z, \Phi) p(\Phi | \beta) d\Phi$$

$$p(z | \alpha) = \int p(z | \Theta) p(\Theta | \alpha) d\Theta$$

Θ, Φ : the matrix of document – topic matrix and the word – topic

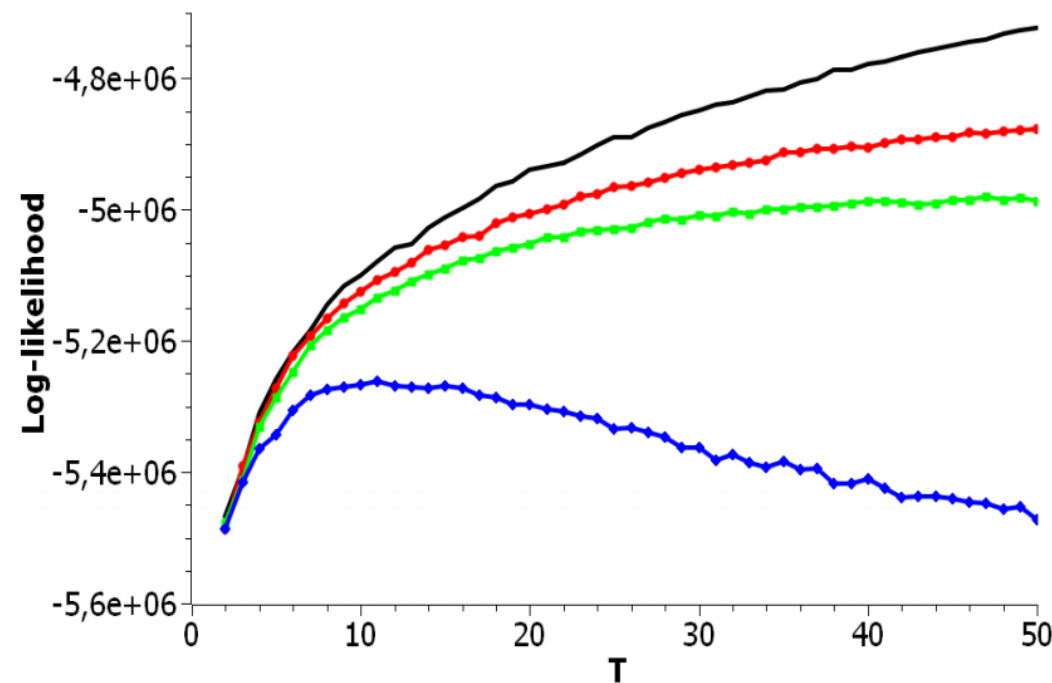
Распределение вероятностей слов и документов можно оценить с помощью мат. ожиданий, если знать подынтегральные функции.

$$P(z_i = j | w_i = m, z_{-i}, w_{-i}) \approx \frac{C_{m,j}^{WT} + \beta}{\sum_{m'} C_{m',j}^{WT} + V\beta} \cdot \frac{C_{d,j}^{DT} + \alpha}{C_{d,j'}^{DT} + \alpha T}$$

How to determine optimal LDA parameters

logarithm of likelihood – логарифм правдоподобия

$$\ln(P(\hat{D}|\Phi, \Theta)) = \sum_{d=1}^D \sum_{w=1}^W n_d^w \ln\left(\sum_{t=1}^T \phi_{wt} \theta_{td}\right).$$



The greater the
value of likelihood
is the better

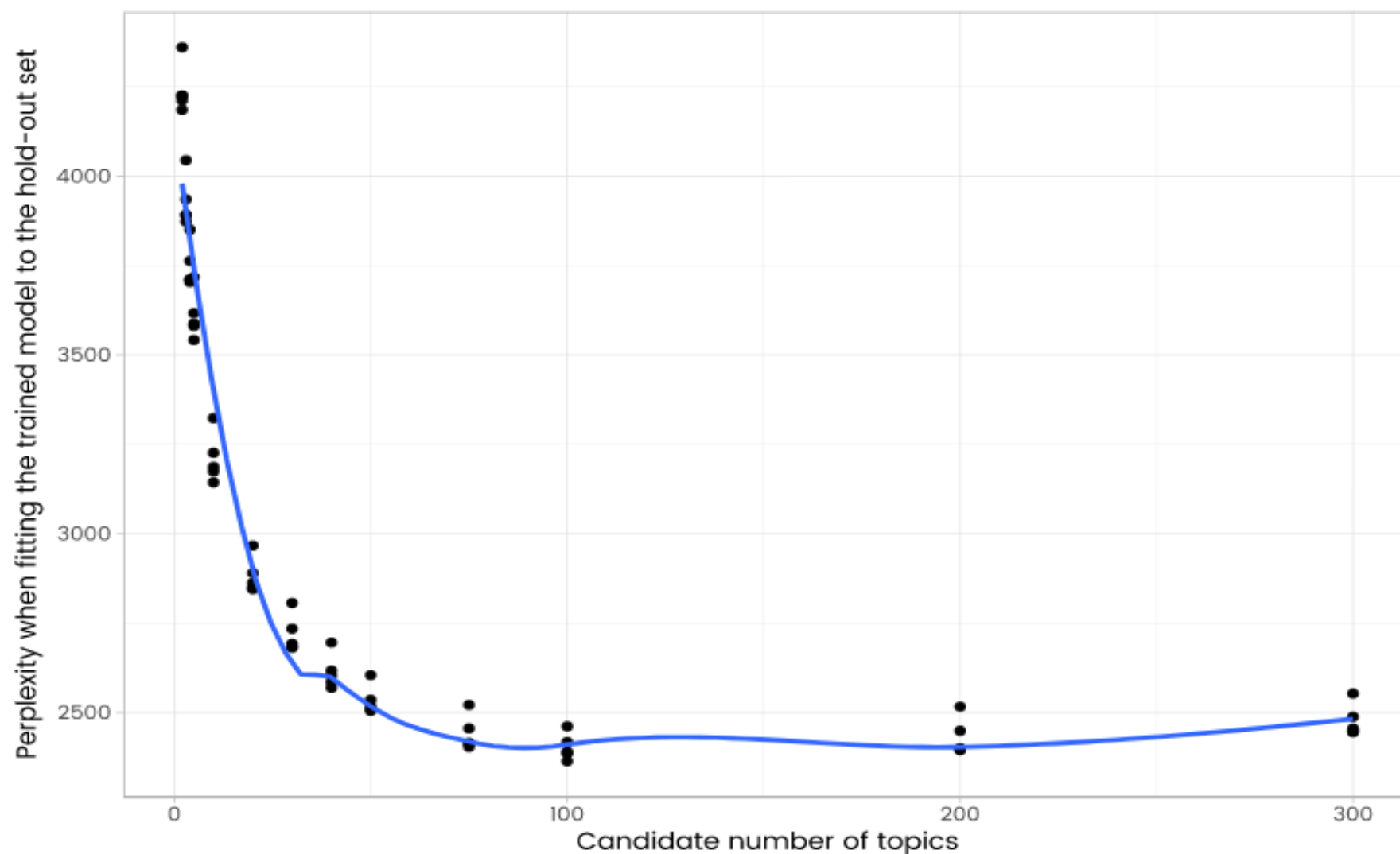
Figure Log-likelihood distribution over T for different α and β (Russian dataset).
pLSA - black, LDA GS ($\alpha=0.1$, $\beta=0.1$) - red, LDA GS ($\alpha=0.5$, $\beta=0.1$) - green, LDA GS ($\alpha=1$, $\beta=1$) - blue.



How to determine optimal LDA parameters

Perplexity

$$(D_{\text{test}}) = \exp \left(-\frac{\sum_{d=1}^M \log p(d)}{\sum_{d=1}^M N_d} \right) = \exp \left(-\frac{\sum_{d=1}^M \sum_{w=1}^W n_d^w \log(\sum_{t=1}^T \phi_{wt} \theta_{td})}{\sum_{d=1}^M N_d} \right)$$



The smaller the
value is the
better result



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 Реньи, то в дальнейшем будем анализировать поведение энтропии Реньи.
2. Выбор энтропию Реньи 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) = $1/T$.



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: **$S = \ln(\check{\rho}(T))$** ,

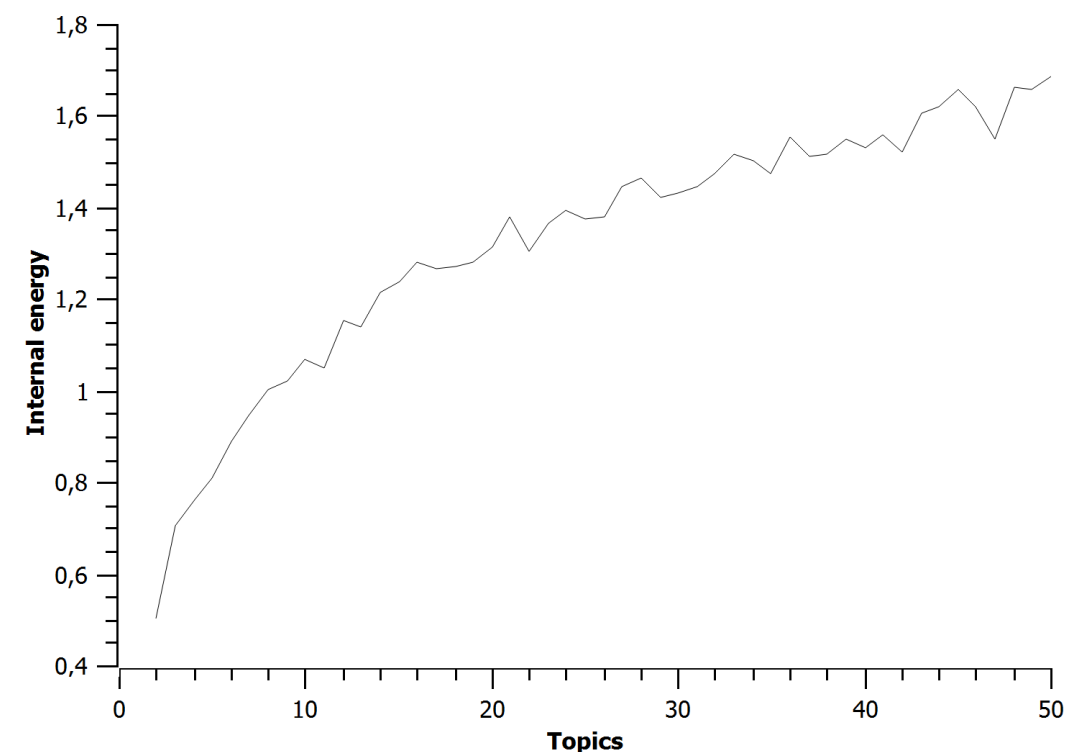
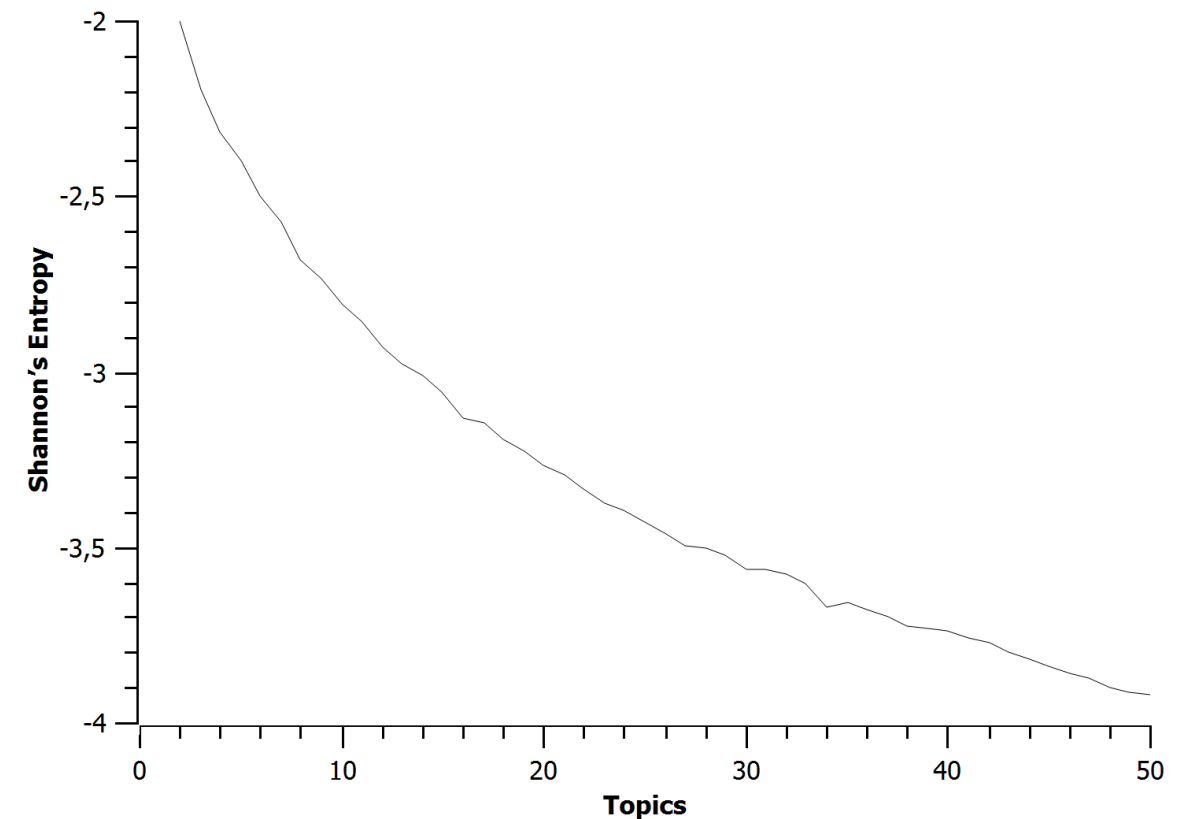
$\check{\rho} = \frac{\sum_{t=1}^T \sum_{n=1}^W N_{tn}}{WT}$, where $N(E)_{tn}$ – number of states with energy E, n,t - summation over all words and topics.

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.

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

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

$$\begin{aligned} \Lambda_F &= F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T \\ &= -\ln(\check{P}) - T \cdot \ln(\check{\rho}) \end{aligned}$$





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 датассете 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 датассете. 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 Реньи рассчитывалась на основе усредненных значений плотности состояний и внутренней энергии.

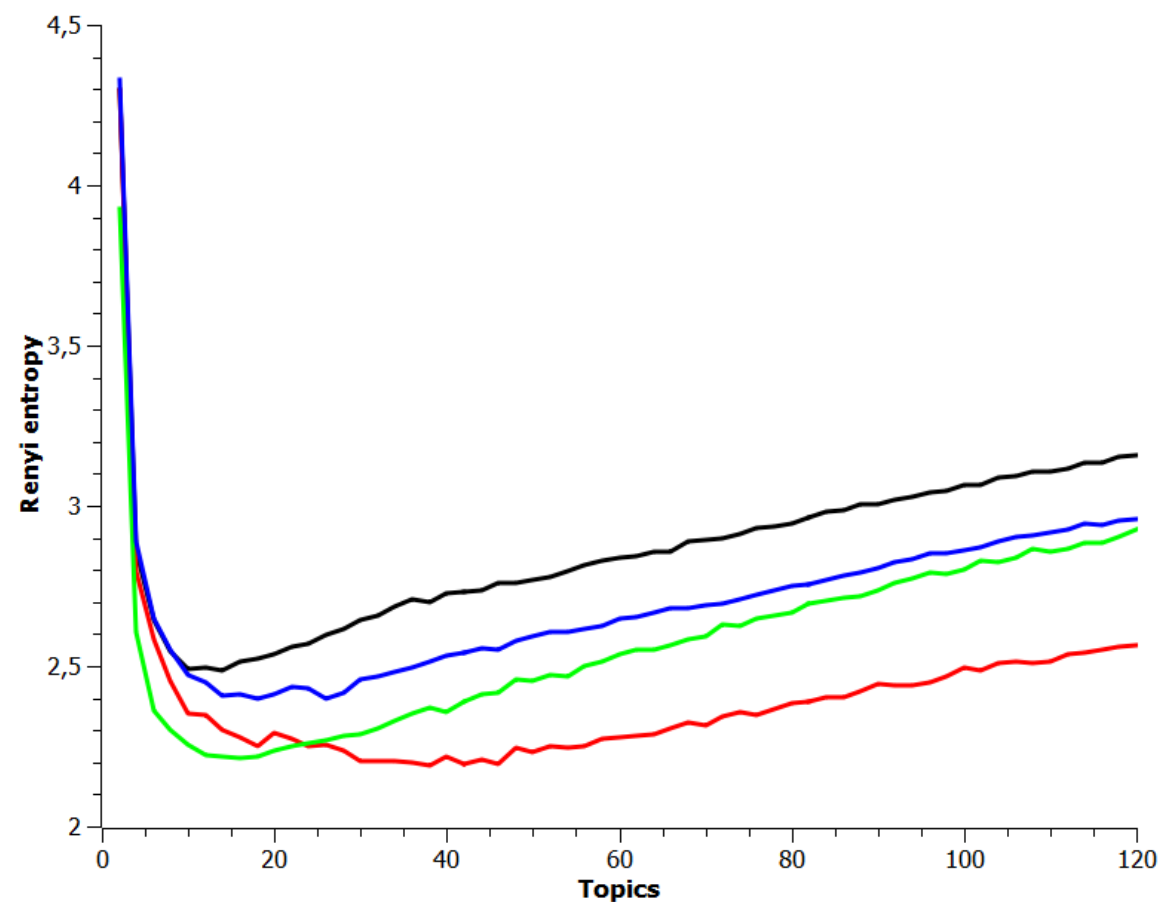


Renyi entropy

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.

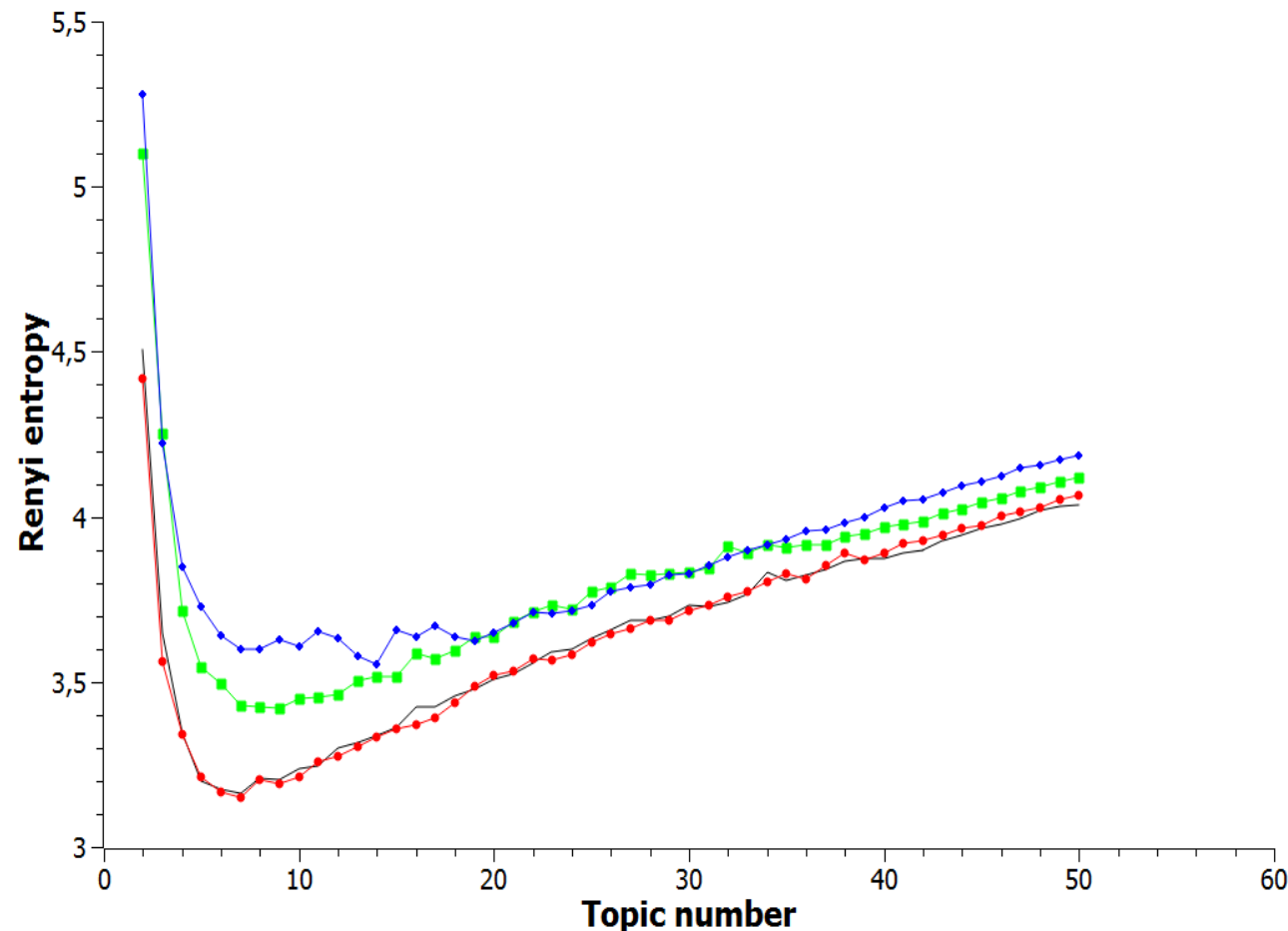
$$S_{q=1/T}^R = \frac{\Lambda_F}{T-1}$$

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

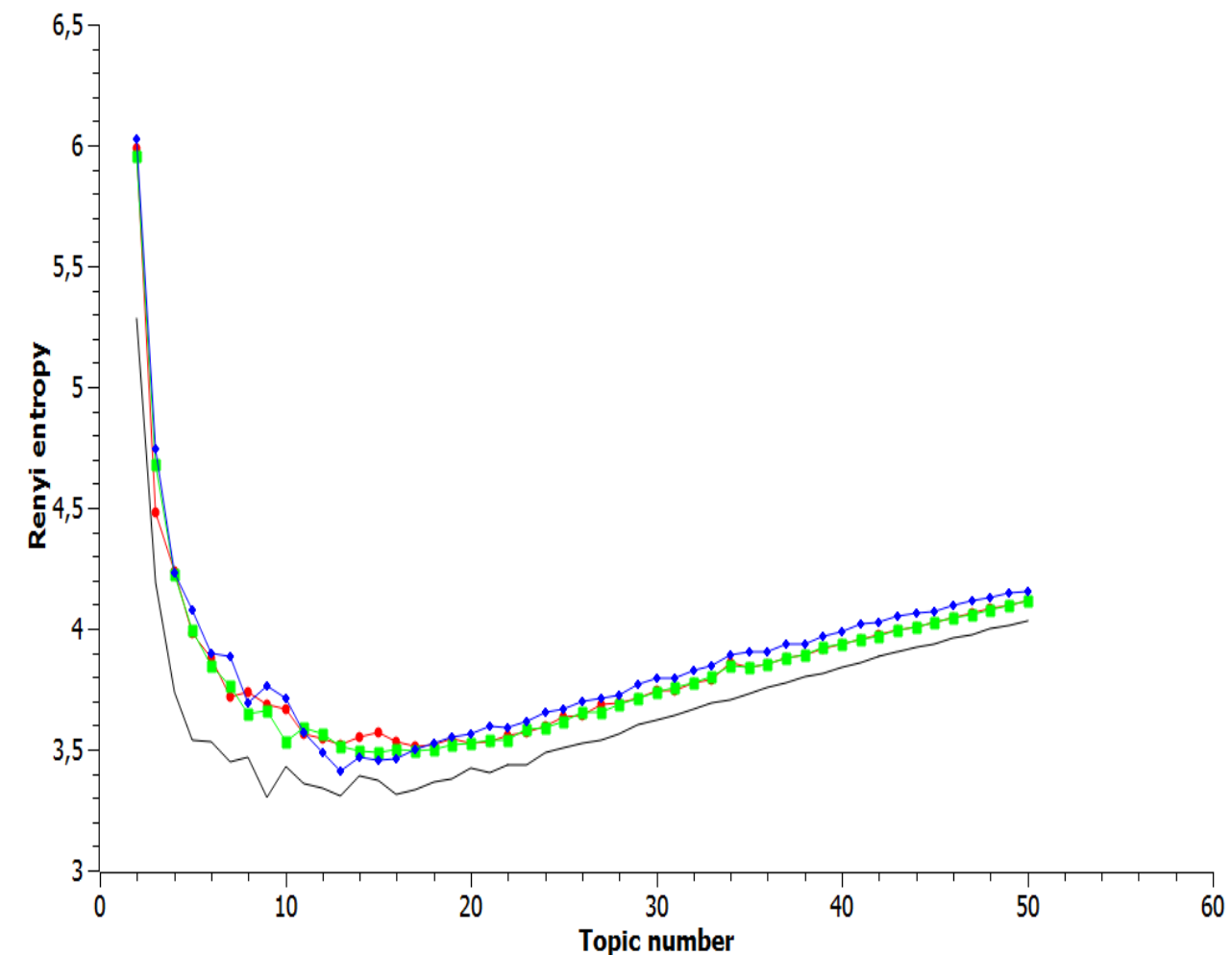


LDA (Gibbs sampling) – black color, **GLDA** (Gibbs sampling) – red color,
PLSA (E-M algorithm) – blue color, **VLDA** (E-M algorithm) – green color.

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.



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).

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



Sharma-Mittal Entropy in topic modeling

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

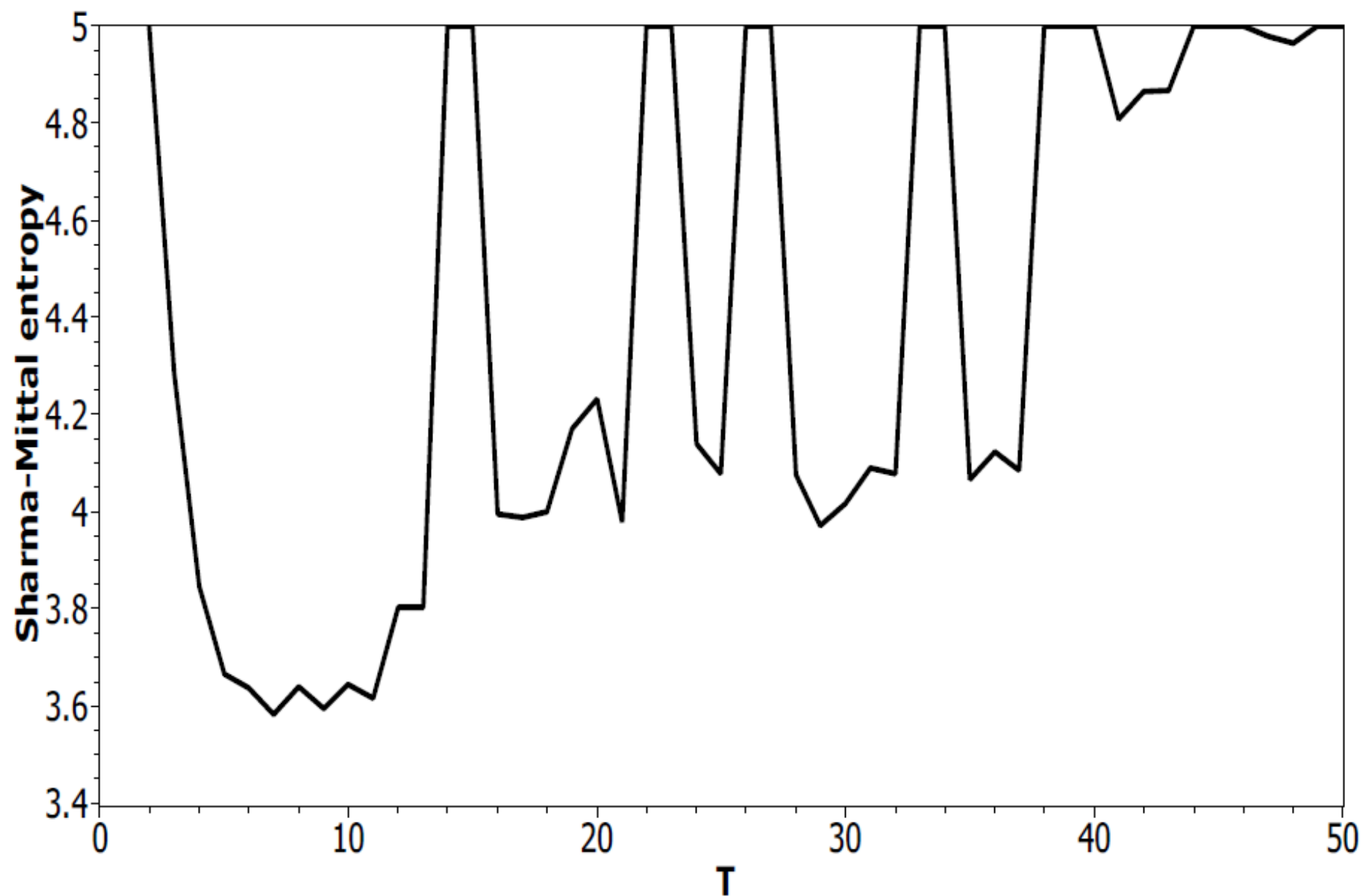
Получаем, что предел энтропии Шарма-Митталь при $r \rightarrow 0$ равен экспоненте от энтропии Реньи без единицы, что можно рассмотреть как деформированную перплексию. Таким образом, когда $r=0$, то энтропия Шарма-Митталь выглядит следующим образом: $S_{S,M} = e^{S_q^R} - 1$, то есть энтропия становится 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.

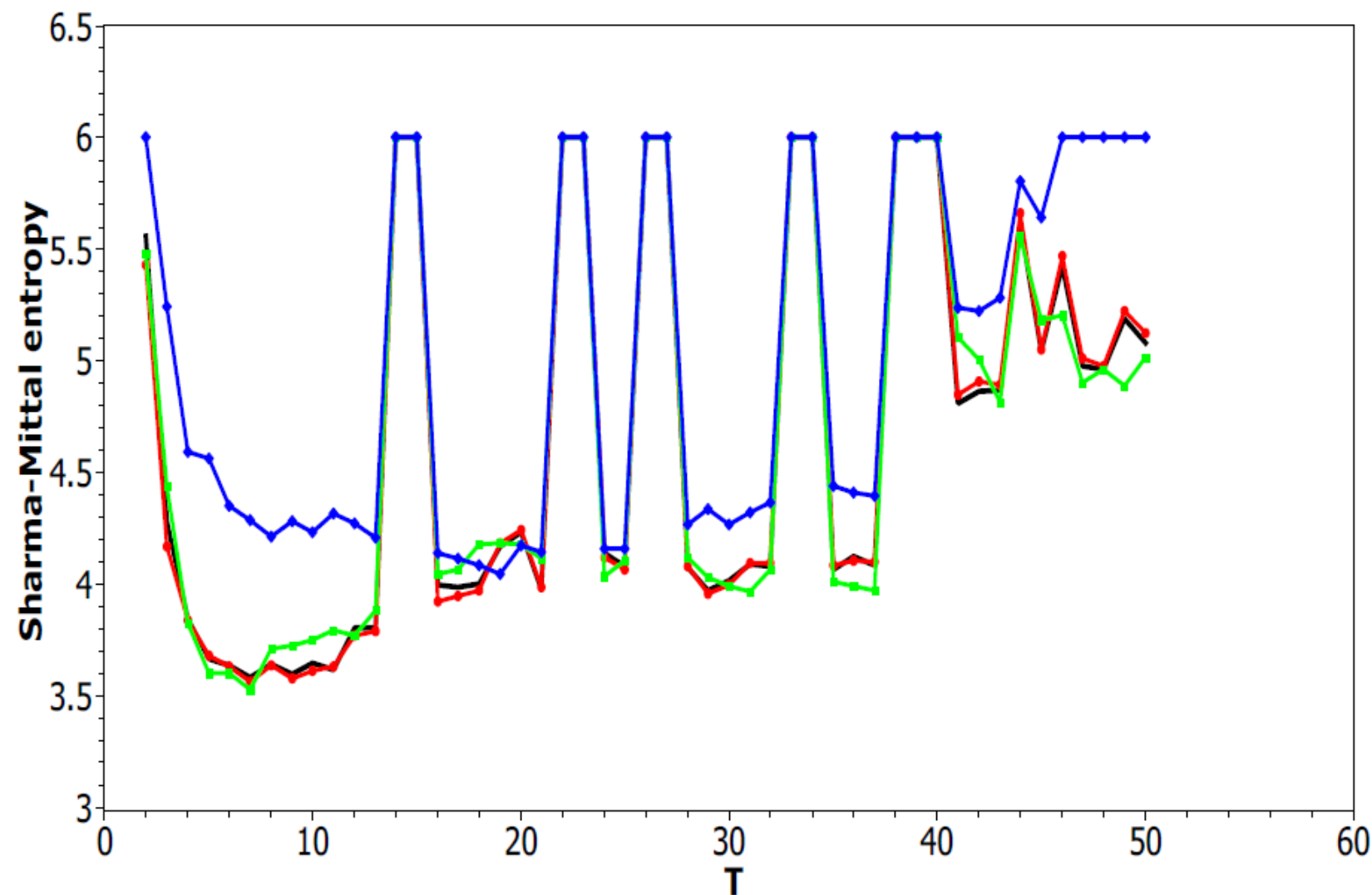


Sharma-Mittal entropy при вариации числа тем



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

Sharma-Mittal entropy при вариации T и гипер - параметров



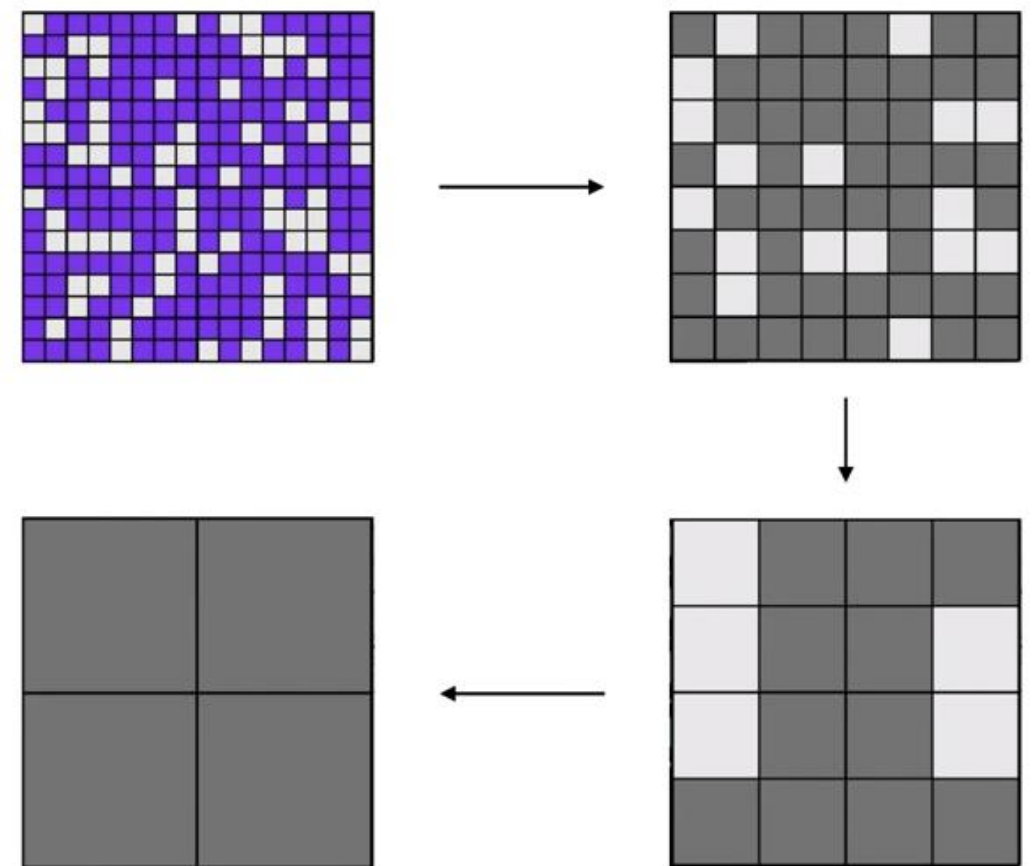
Sharma-Mittal entropy distribution over topics (Russian dataset). pLSA – black, LDA ($a = 0.1$, $b = 0.1$) – red, LDA ($a = 0.5$, $b = 0.1$) – green, LDA ($a = 1$, $b = 1$) – blue.



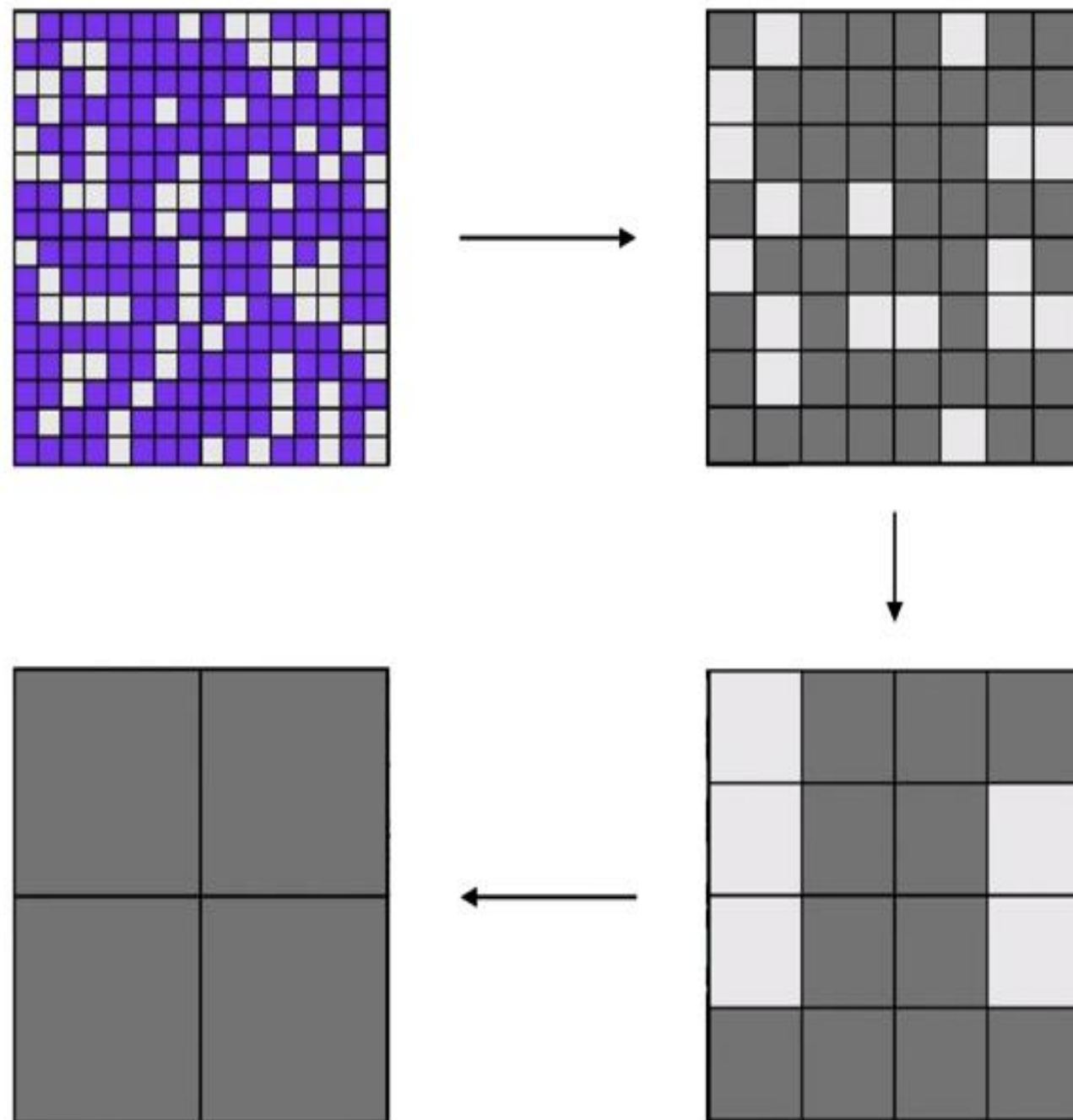
Введение в теорию ренормализации

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.

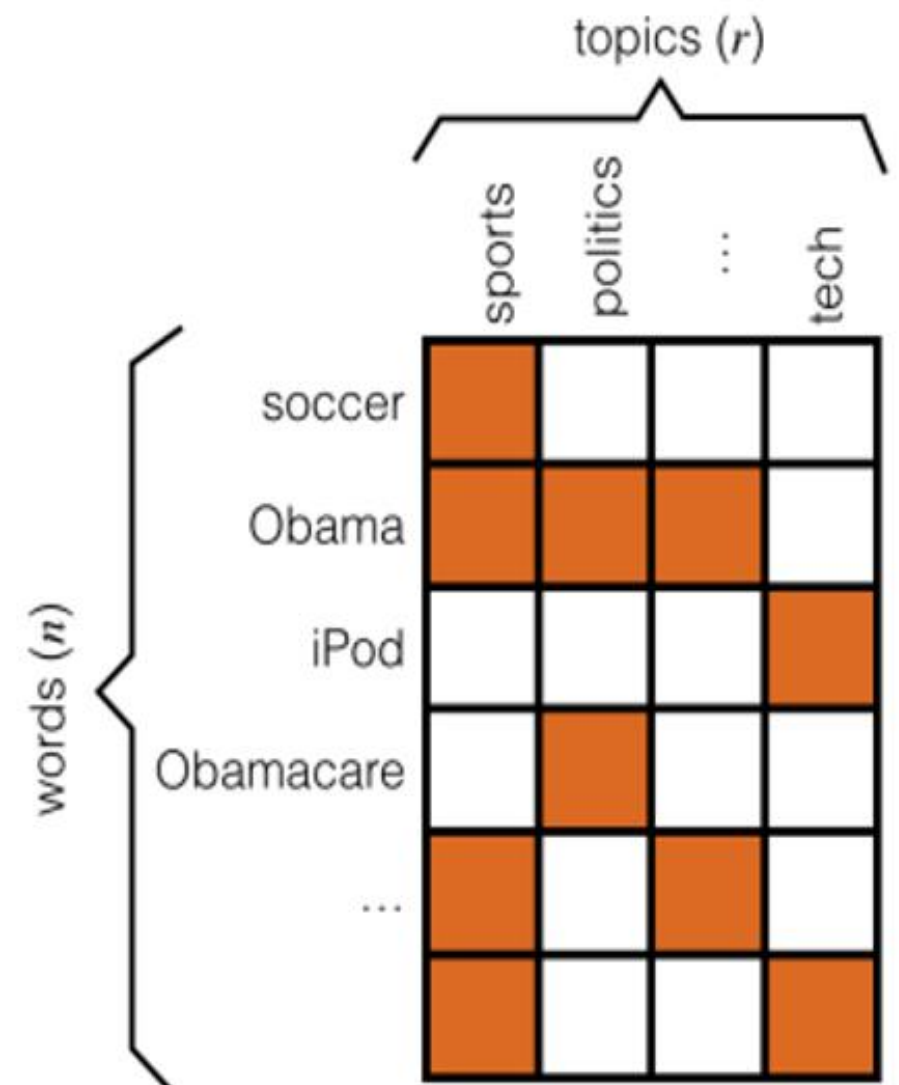
Суть процедуры ренормализации заключается в следующем. Рассмотрим поверхность состоящих из совокупности узлов. Каждый из узлов, характеризуется направлением спина. В свою очередь, спин может занимать определение направление, количество которых зависит от задачи. Узлы с одинаковыми спинами составляют 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 всей 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 ϕ_{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 Реньи 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 ϕ_{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 Реньи.
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 Реньи.** В данном подходе рассчитывается энтропия Реньи для каждой темы по отдельности. Далее производится сортировка всех тем по величине энтропии Реньи. 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.



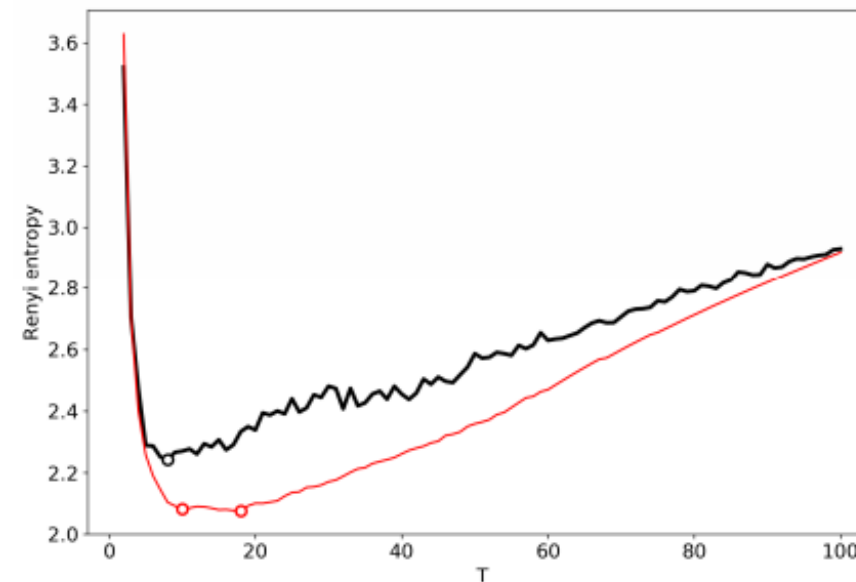
Ренормализации в модели LDA Gibbs sampling

Алгоритм ренормализации для процедуры сэмплирования Гиббса состоит из следующих шагов:

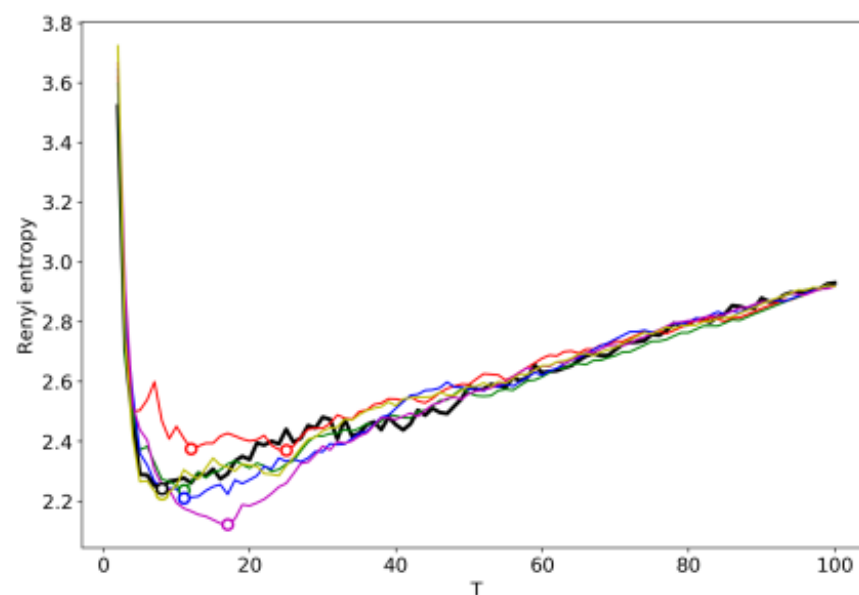
- 1. Выбор пары тем для склейки.** В рамках ренормализации мы сначала выбираем 2 темы для склеивания разными способами: Обозначим 2 выбранные темы одним из перечисленных способов за t_1 и t_2 .
- 2. Склеивание выбранных тем.** Склеивание выбранных заключается в суммировании частот слов c_{wt} двух выбранных тем. Затем, на основании новых значений счетчика производится расчет матрицы ϕ_{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 $\phi_{.t_1}$ уже удовлетворяет свойству: $\sum_{w \in W} \phi_{wt_1} = 1$. Далее мы удаляем столбец $\phi_{.t_2}$ из матрицы Φ . 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 Реньи, как функция от параметра ренормализации, то есть от числа тем.

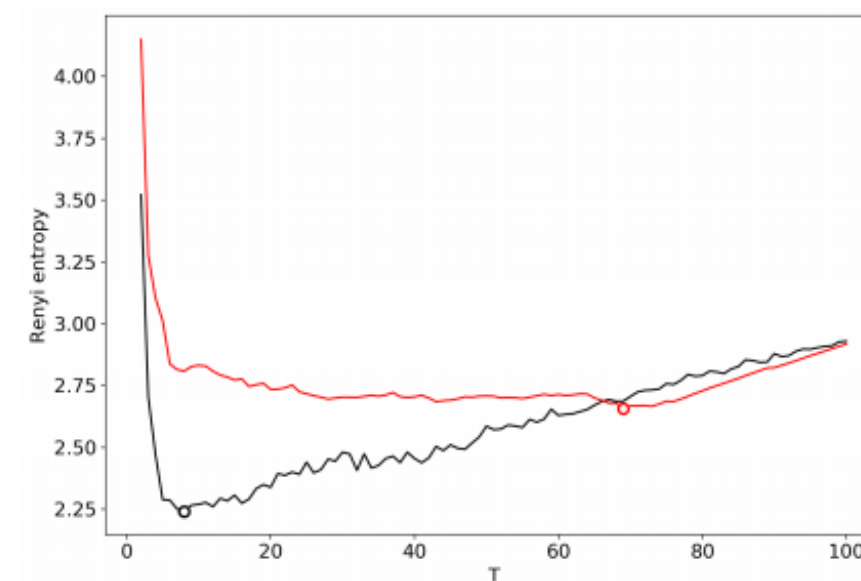
Ренормализация модели LDA Gibbs sampling (lenta)



Объединение тем с наименьшими локальными энтропиями Реньи.

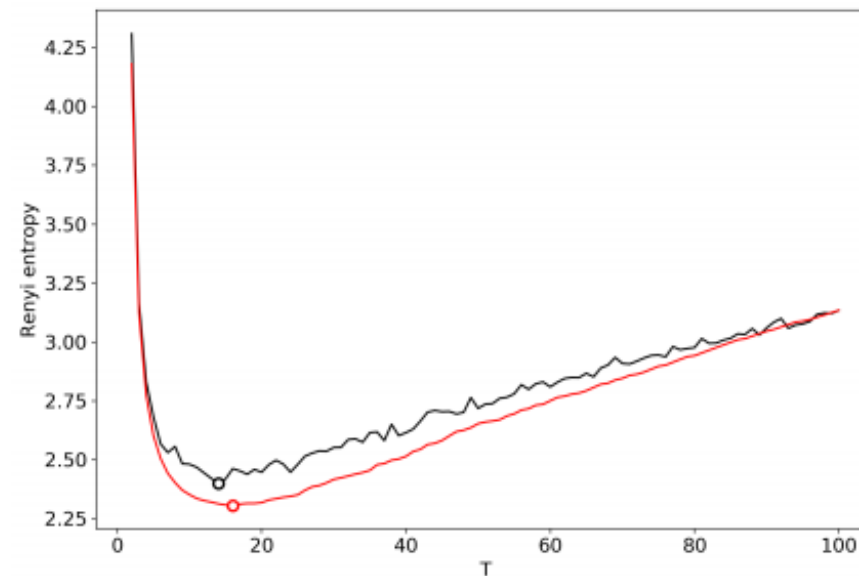


Объединение случайных тем.

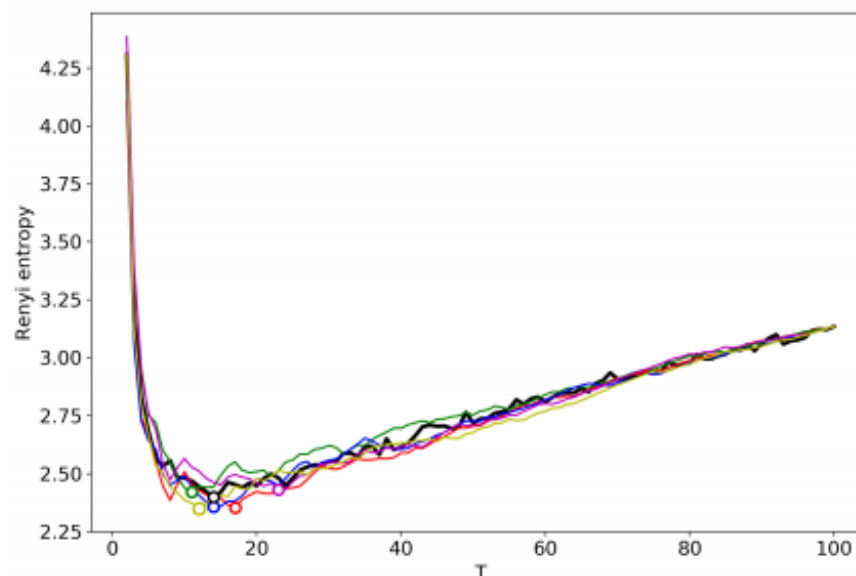


Объединение схожих тем (КЛ).

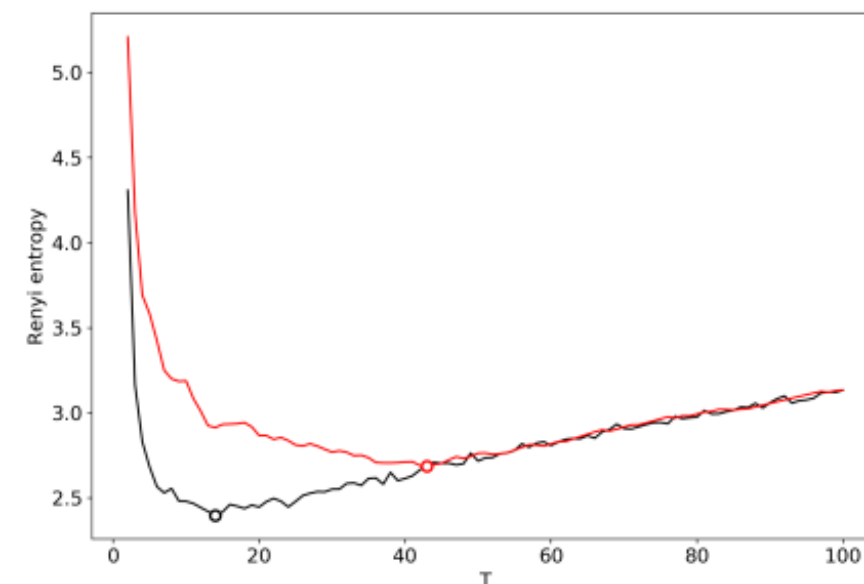
Ренормализация модели LDA Gibbs sampling (20Newsgroups)



Объединение тем с наименьшими локальными энтропиями Реньи.

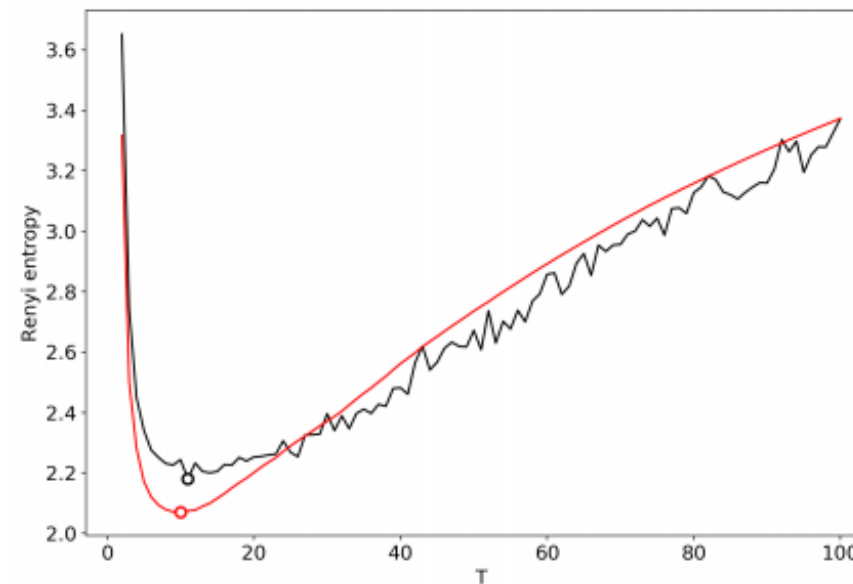


Объединение случайных тем

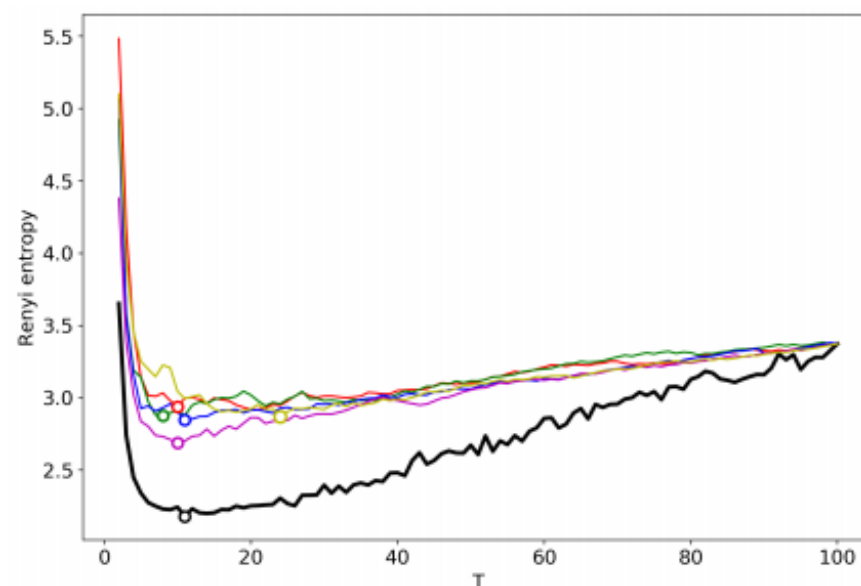


Объединение схожих тем (КЛ).

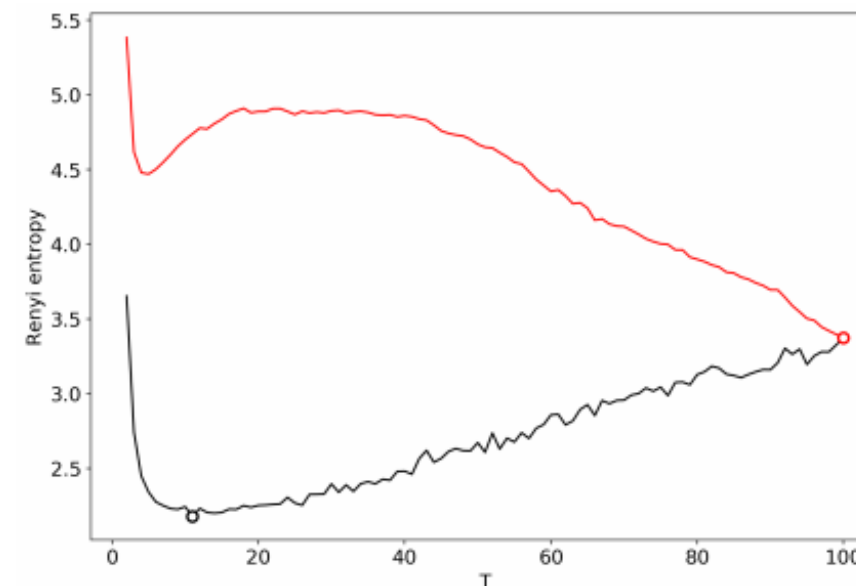
Ренормализация модели VLDA (lenta)



Объединение тем с наименьшими локальными энтропиями Реньи.

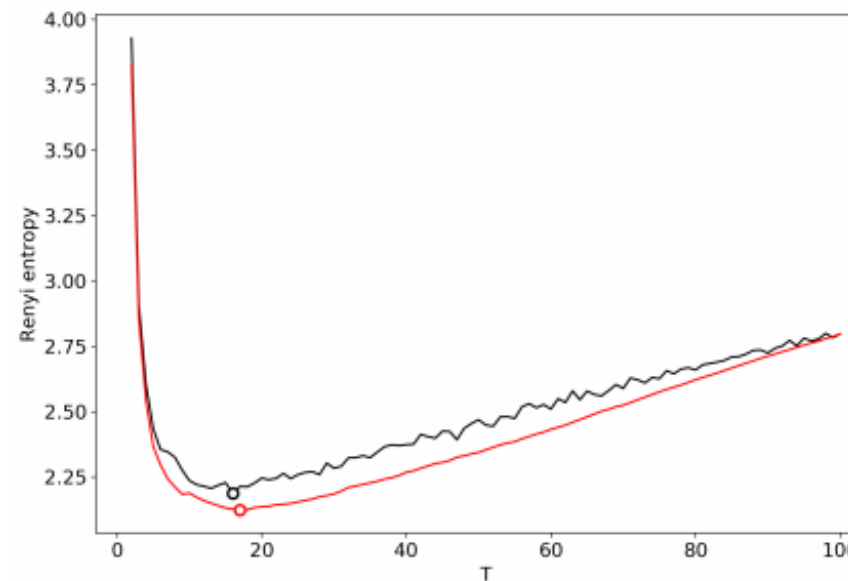


Объединение случайных тем.

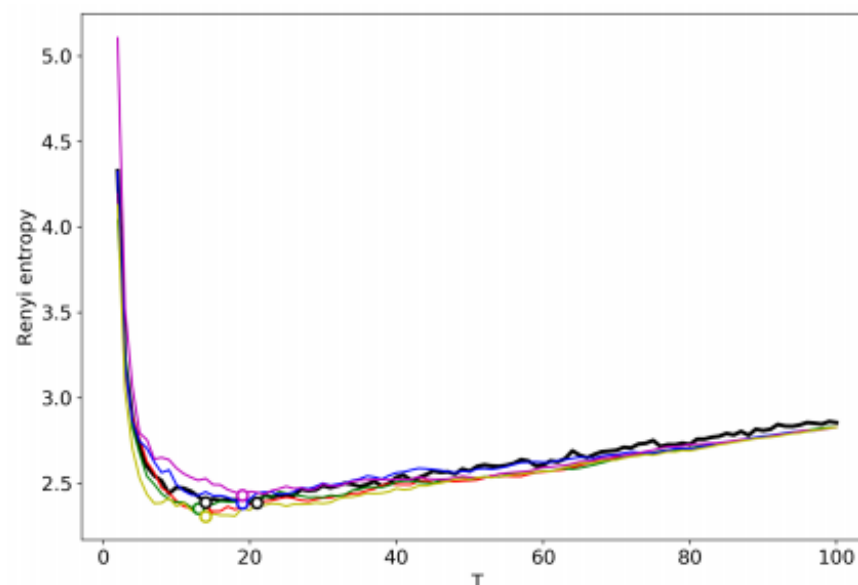


Объединение схожих тем (КЛ).

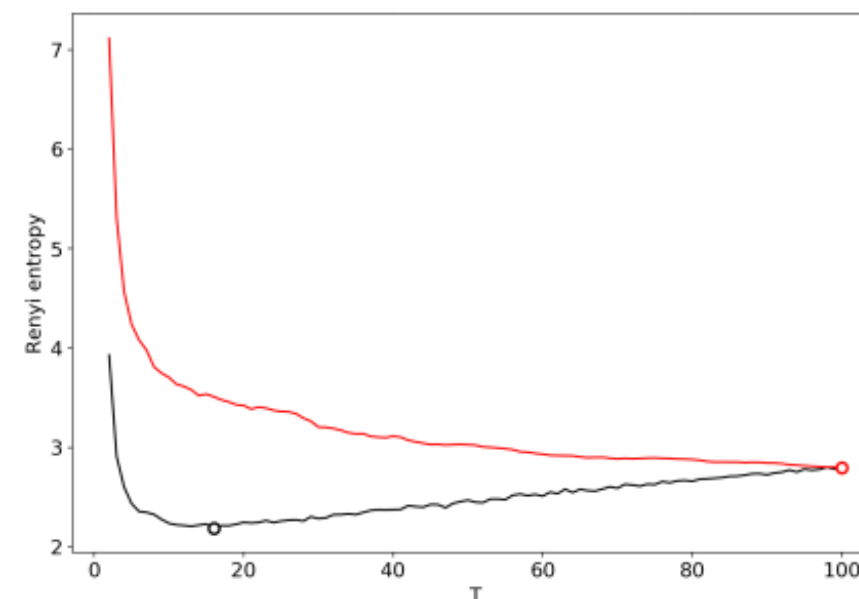
Ренормализация модели VLDA (20Newsgroups)



Объединение тем с наименьшими локальными энтропиями Реньи.



Объединение случайных тем



Объединение схожих тем (КЛ).



Время расчета разных 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 Реньи ведут к удовлетворительным результатам с точки зрения точности получаемых аппроксимаций и дают существенный выигрыш по времени расчетов (20 часов для pLSA на коллекции 20Newsgroups).



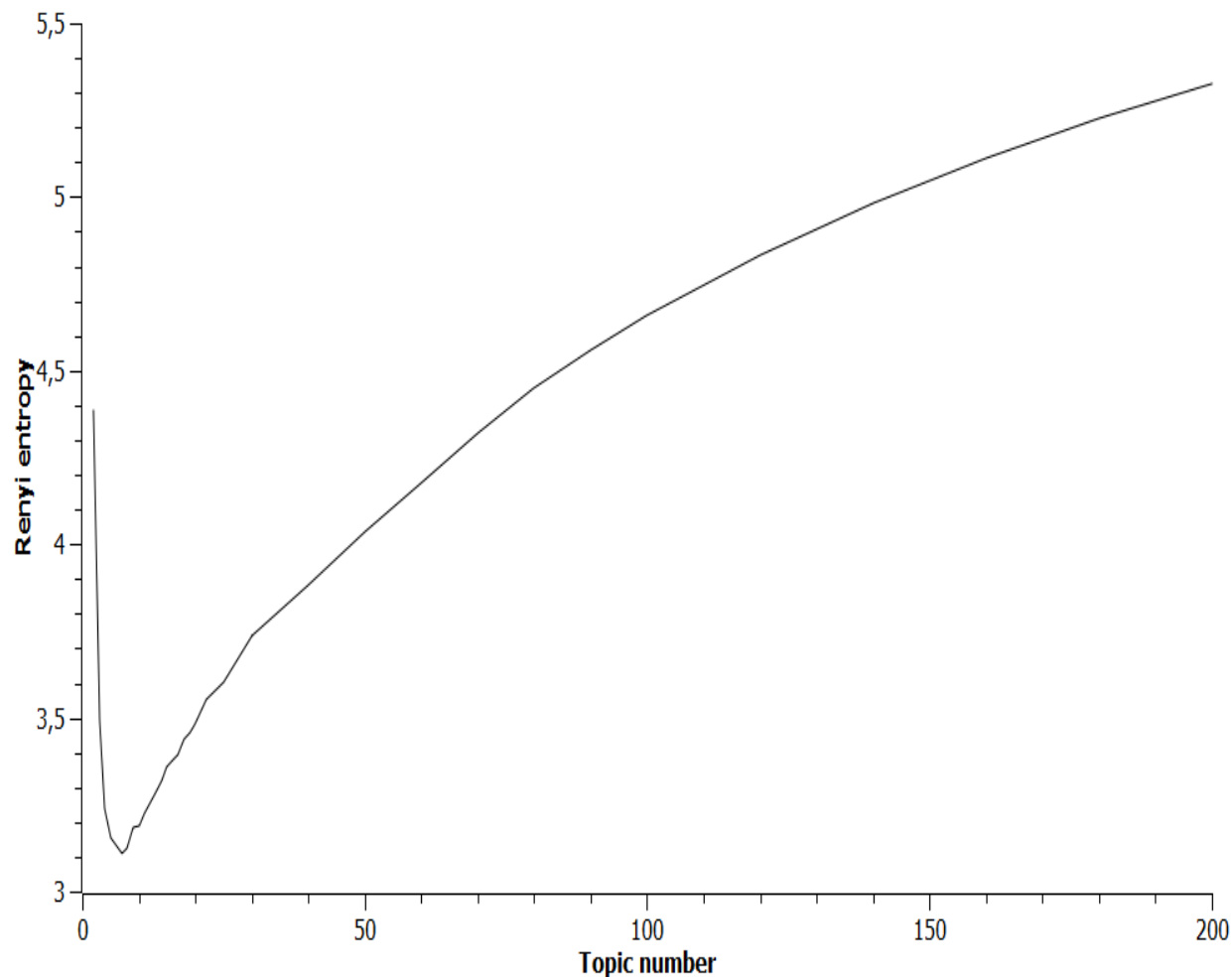
Модель hARTM – это иерархическая версия подхода аддитивной регуляризации тематических моделей (ARTM)

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.

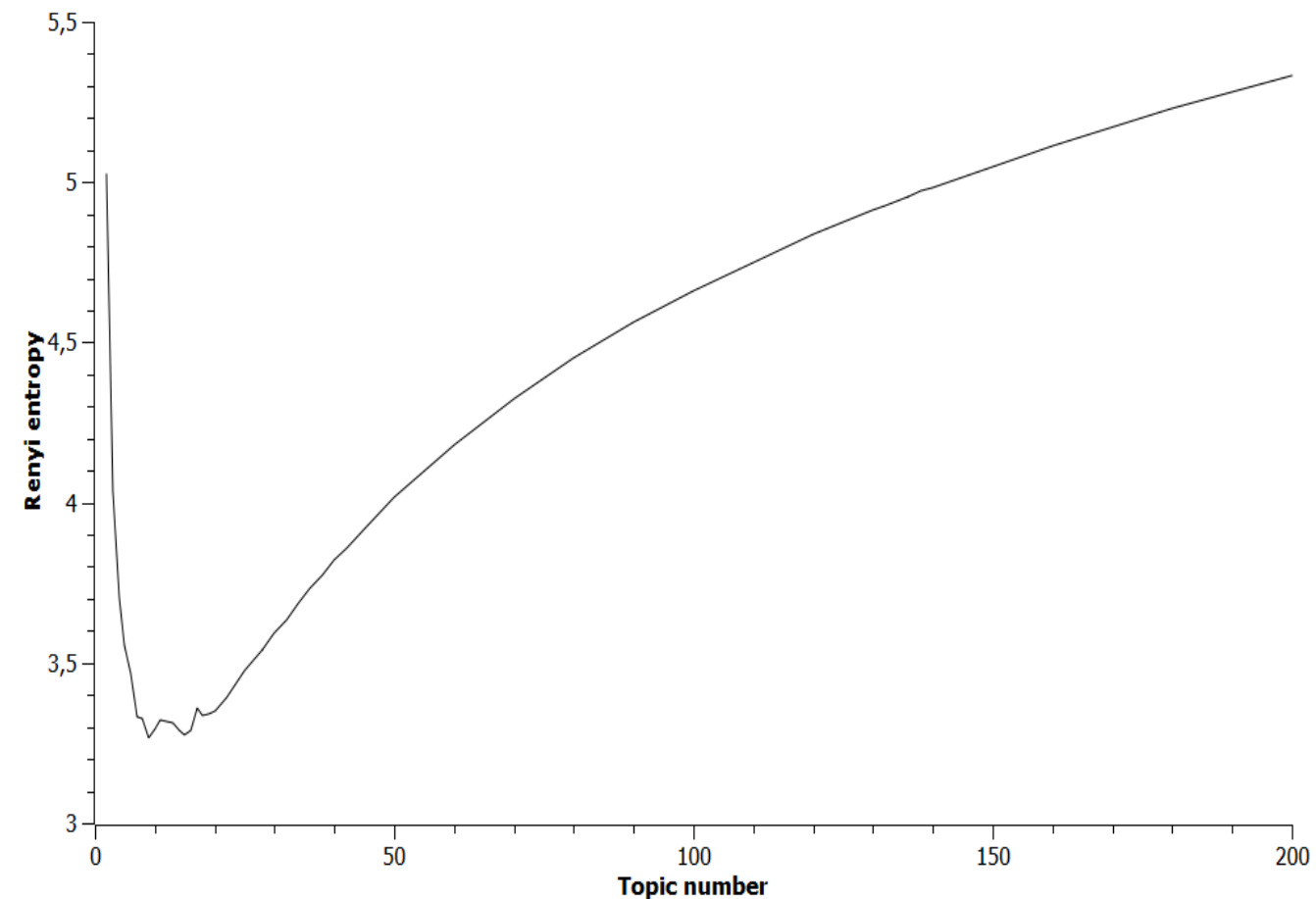
В рамках исследования данной модели проводилось ТМ на 4 датасетах. При этом варьировалось число тем на втором и третьем уровне при разных параметрах инициализации (seed).



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



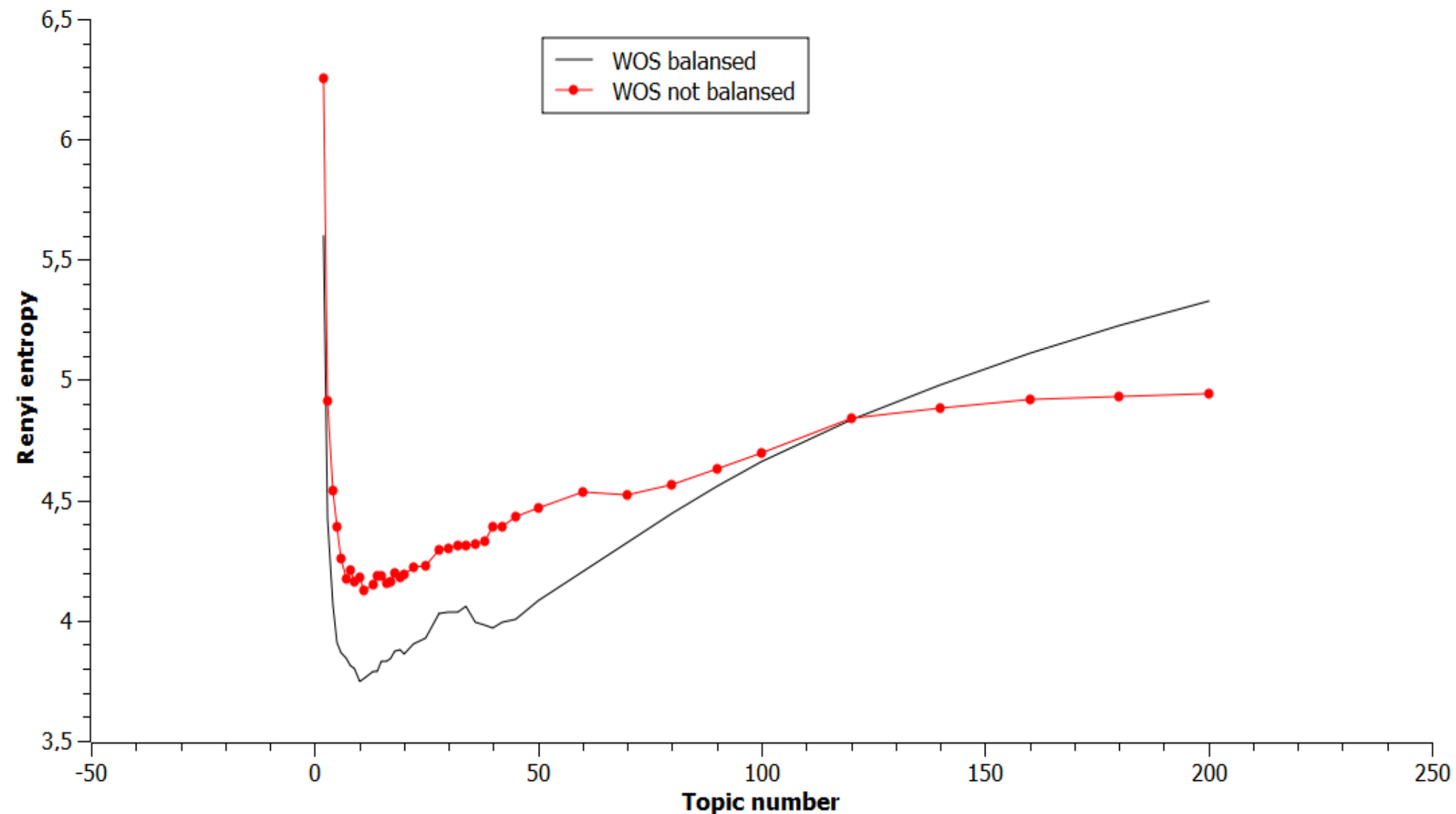
Зависимость энтропии Реньи от числа тем на первом уровне иерархии в модели hARTM ('lenta')



Зависимость энтропии Реньи от числа тем на втором уровне иерархии в модели hARTM ('20Newsgroups')

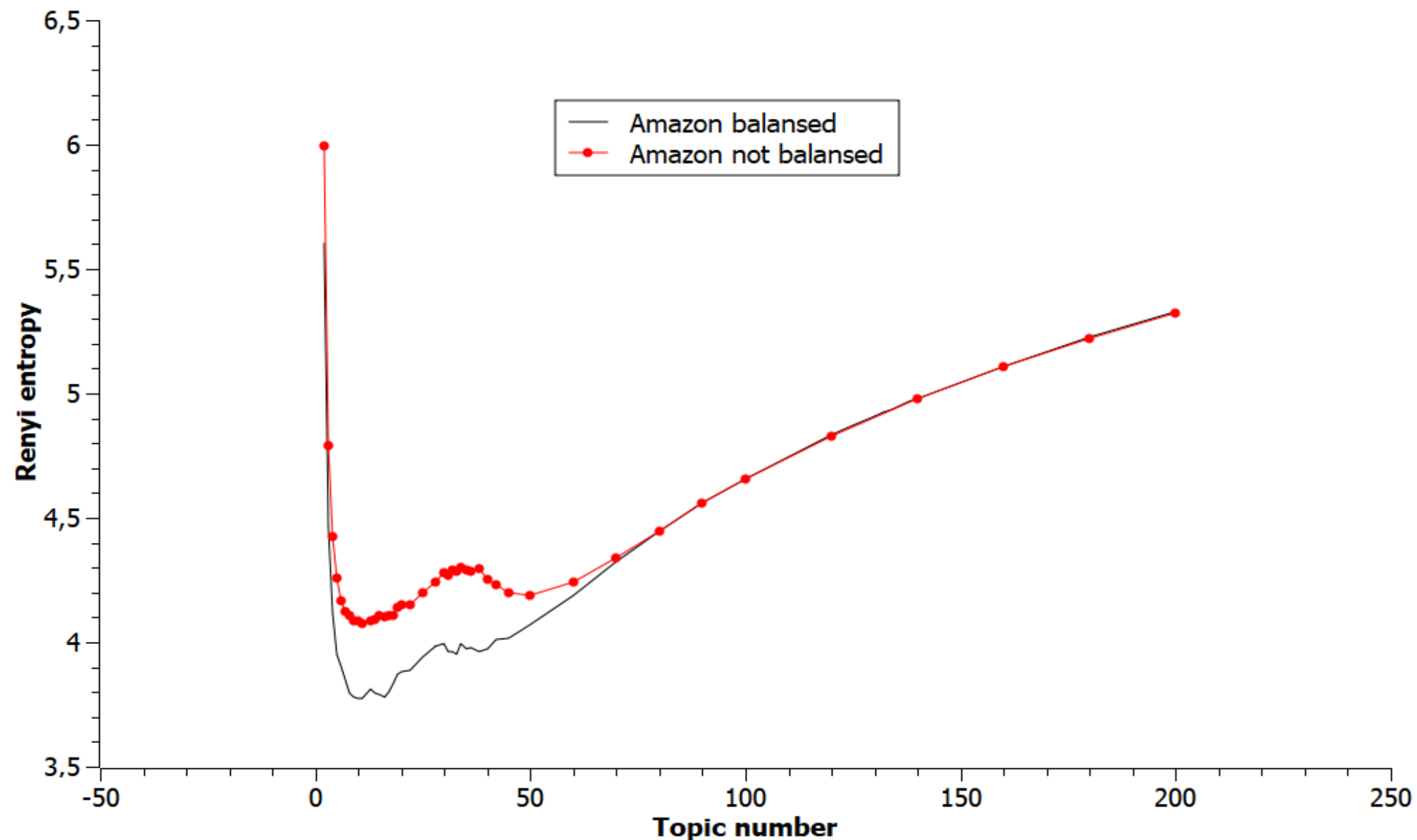
Модель hARTM хорошо видит плоскую структуру датасетов.

Модель hARTM ('WoS balanced' и 'WoS')



Кривые энтропии Ренья для сбалансированного и не сбалансированного датасета ('WOS') на первом уровне. Черный цвет — сбалансированный датасет, красный цвет — несбалансированный датасет.

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



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



Что такое Pruning (разреживание)?

Зачем нужен pruning?

- Современные нейросети имеют миллионы и миллиарды параметров.
- Это требует много памяти и времени на обучение/инференс.
- **Pruning** позволяет удалить наименее значимые веса, сохранив точность.

Проблема:

- Неизвестно, **какой уровень разреживания** выбрать заранее.
- Часто это делается эмпирически, с перебором, что занимает много времени и ресурсов.

Weight pruning (разреживание весов)

Удаляется веса между нейронами
Пример: $W = [0.4, -0.01, 0.003]$ при
threshold 0.01 $\rightarrow W = [0.4, 0, 0]$

Activation pruning (разреживание активаций)

Обнуляется выходы нейронов модели.

Structured Pruning — это подход, при котором удаляются целые структурные блоки нейросети, а не отдельные веса.



Идея: нейросеть как термодинамическая система

Нейросеть рассматривается как информационная система с N элементами (весами). У такой информационной системы можно измерять различные ‘термодинамические’ параметры: энтропия Шеннона, температура, внутренняя энергия, свободная энергия и так далее.

Энтропия Шеннона $S(T)$

$$S(T) = \ln(\rho)$$

где ρ — относительная плотность весов нейронной сети которая выражается следующим образом:

$$\rho = \frac{N_{non\_zero}}{N}$$

N_{non_zero} — количество не нулевых весов нейронной сети при заданной величине pruning, N — общее количество весов в нейронной сети до проведения процедуры разреживания. Плотность ρ изменяется при изменении порога прореживания (**threshold**), что позволяет анализировать влияние процедуры pruning на энтропию системы по отношению к начальному состоянию информационной системы.



Идея: нейросеть как термодинамическая система

Температура системы

В данной работе, процесс модификации сети заключается в проведении процедуры разреживания. Исходя из этого определим температуру информационной системы T в виде обратной величины порога прореживания:

$$T = 1 / \text{threshold}$$

Температура T характеризует "состояние" системы: чем выше температура, тем ниже уровень прореживания (меньше весов обнуляется), и на оборот, понижение температуры (увеличение уровня pruning) приводит увеличению количества нулевых весов в нейронной сети.

Внутренняя энергия

Внутренняя энергия E нейронной сети может быть определена в виде отношения суммы вероятностей весов при заданном уровне разреживания, и суммы вероятностей весов сети до проведения разреживания.



Идея: нейросеть как термодинамическая система

Внутренняя энергия

Вероятность конкретного веса в сети — это отношение количества его вхождений к общему числу весов в модели.

$$P_w = \frac{C_w}{\sum_w C_w}$$

Далее рассчитывались суммы вероятностей весов при заданном пороге разреживания. Таким образом, для заданного уровня прореживания внутренняя энергия нейронной сети выглядит следующим образом:

$$E(T) = \ln\left(\frac{\sum |P_i|}{\sum |P_i^0|}\right)$$

$\sum |P_i|$ — сумма весов нейронной сети при заданной величине температуры (то есть при заданном уровне pruning), $\sum |P_i^0|$ — сумма весов при нулевой величине pruning.



Идея: нейросеть как термодинамическая система

Свободная энергия

Свободная энергия F нейронной сети объединяет внутреннюю энергию и энтропию Шеннона следующим образом:

$$F = E - T \cdot S,$$

где E — внутренняя энергия, S — энтропия Шеннона. В итоге получим свободную энергию, которая выражается через наблюдаемые величины:

$$F = \ln\left(\frac{\sum |P_i|}{\sum |P|}\right) - T \cdot \ln\left(\frac{N_{non\_zero}}{N}\right)$$

Таким образом, можно проводить варьирование порога разреживания (то есть температуры) в нейронной сети, при этом вычислять точность работы сети (например в задаче классификации) и вычислять термодинамические параметры, например свободную энергию.



Pruning renormalization

Суть процедуры ренормализации заключается в следующем. Берется исходное распределение весов нейронной сети (вектор всех весов). Далее организуется цикл, в котором варьируется порог отсечения, и на каждой итерации цикла обнуляются веса, которые ниже текущего порога отсечения, то есть общее количество весов не меняется, но меняется количество весов с абсолютными величинами больше нуля. Далее, на каждом уровне отсечения вычисляются **энтропия Шеннона, внутренняя энергия и свободная энергия** по выше приведенным формулам. Отличие процедуры ренормализации от процесса прунинга состоит в том, что мы ренормализуем только исходное распределение весов сети.

Результаты компьютерных экспериментов приведены на рисунках графики 'А' демонстрируют зависимости точности нейронной сети, графики 'В' показывают поведение свободной энергии (два варианта расчета. Соответственно графики 'С' показывают зоны информационных фазовых переходов, которые характеризуются резкими скачками производной свободной энергии в зависимости от уровня обрезания.



Нейронные сети и датасеты

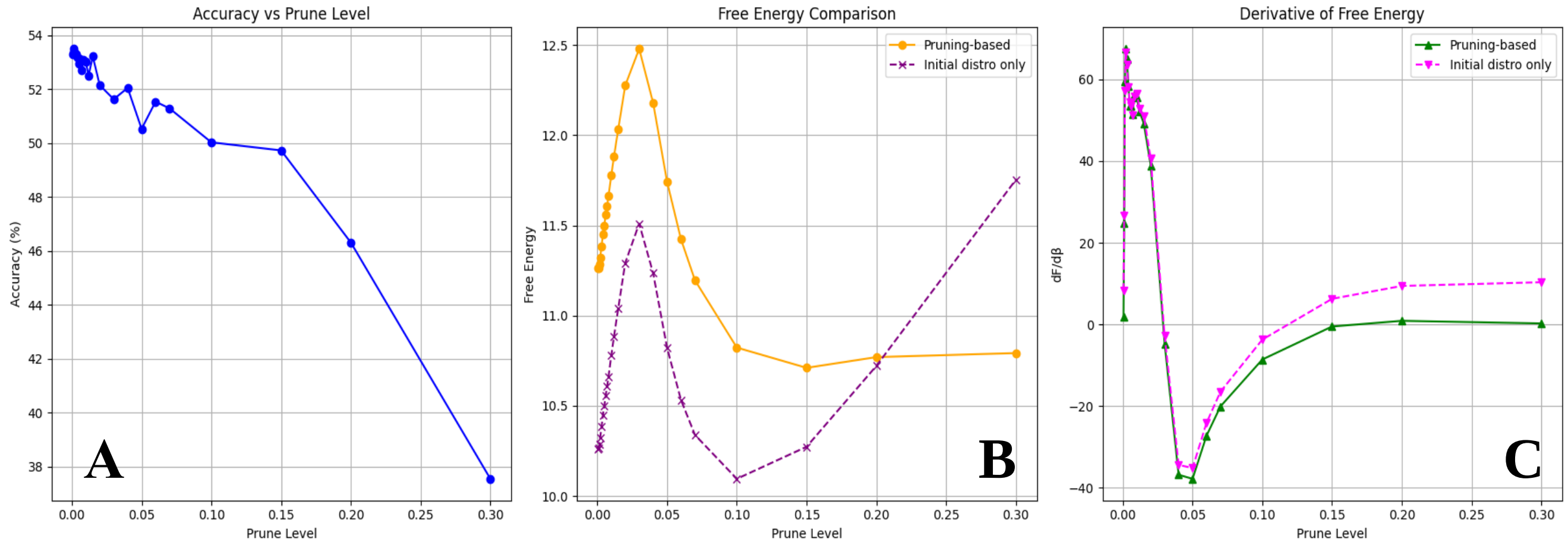
В рамках данной работы были использованы два разнотипных датасета,

1. Усеченный датасет Reviews dataset, размера 28423 строк,
2. Датасет изображений Cifar10 dataset, размером 50000 изображений.

В рамках компьютерных экспериментов мы использовали три различных архитектуры;

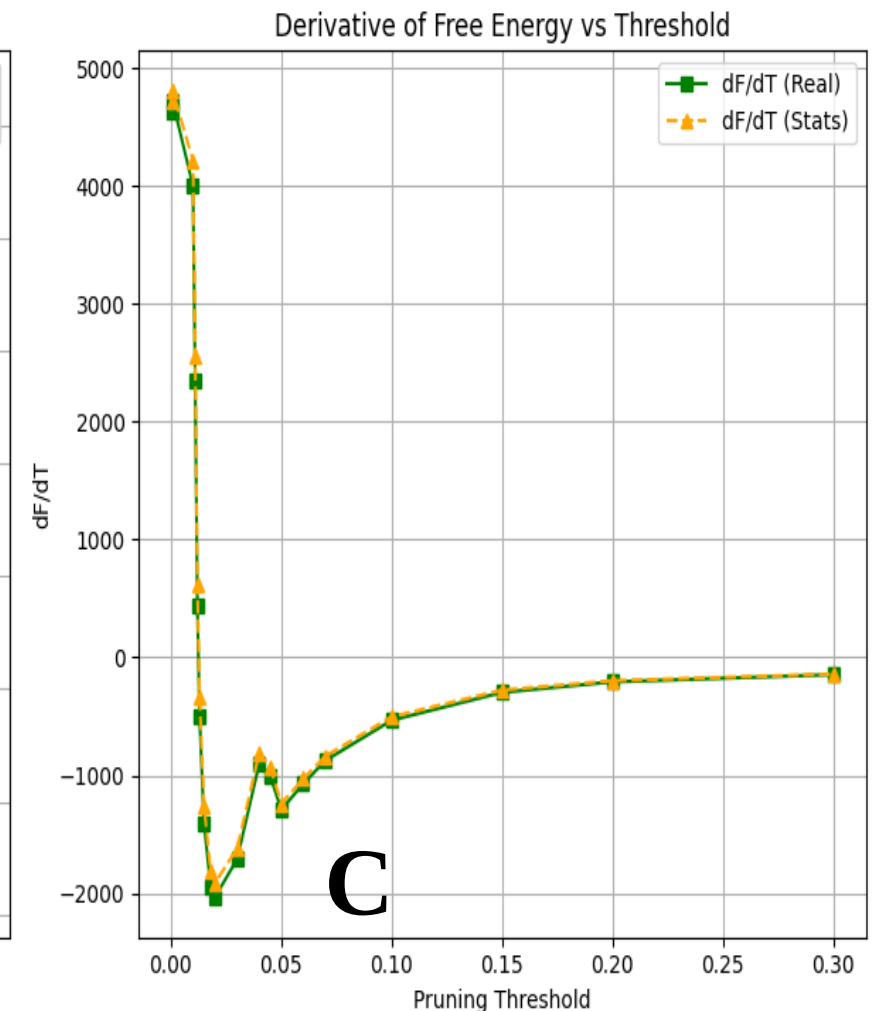
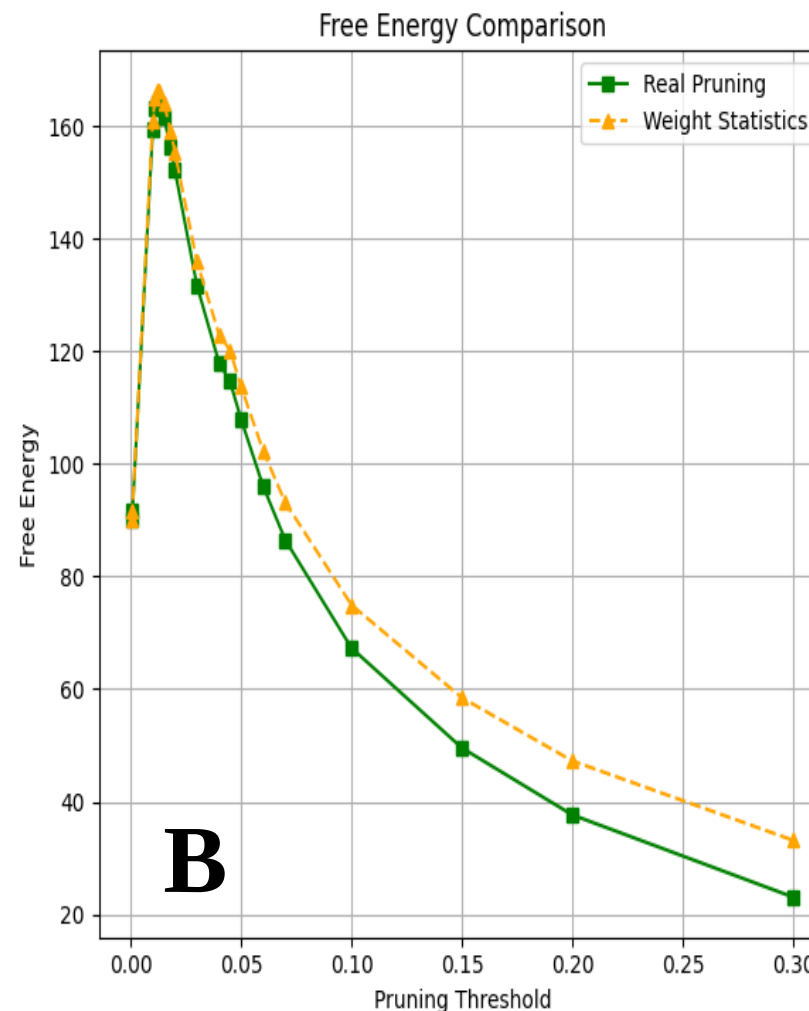
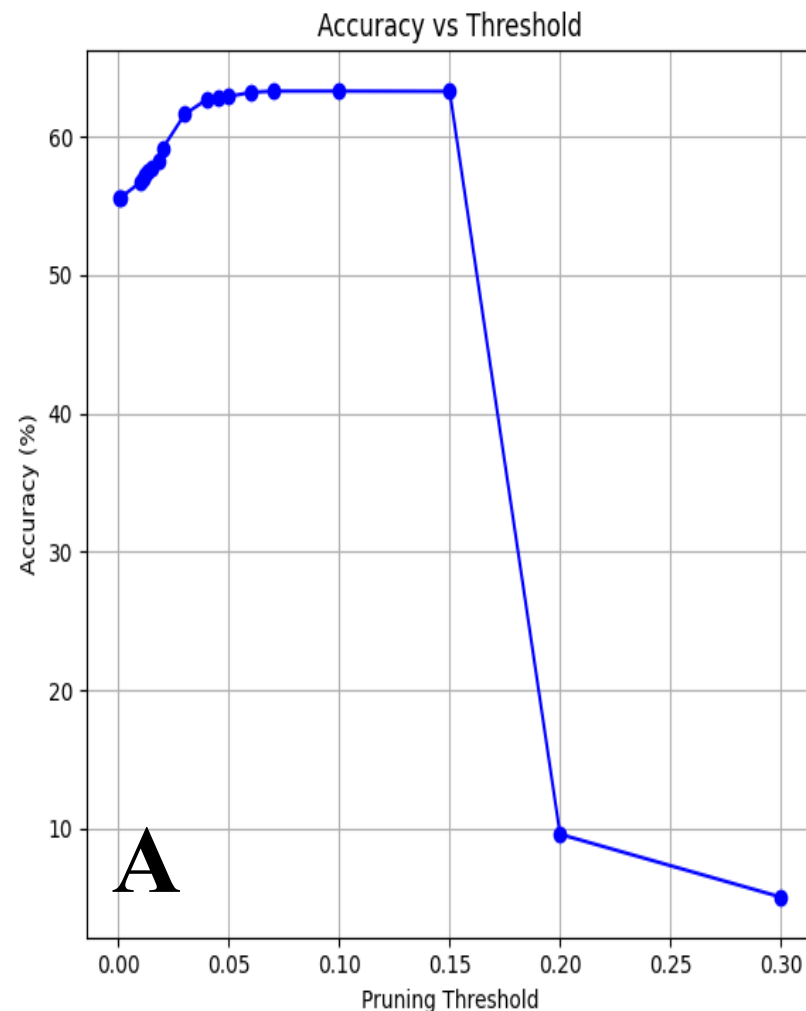
1. **Полносвязанная нейронная сеть.** Первая полносвязанная сеть для обработки изображений представляет собой следующую архитектуру: Первый слой : `nn.Linear(3 * 32 * 32, 1024)` — преобразует плоский вектор изображения (размером $3 \times 32 \times 32 = 3072$) в вектор длиной 1024. Второй и третий слои : `nn.Linear(1024, 1024)` — два последовательных слоя, обрабатывающих векторы длиной 1024. Финальный слой: `nn.Linear(1024, 10)` — преобразует вектор длиной 1024 в выходной вектор из 10 классов. **Вторая полносвязанная сеть** для текстовых данных представляет собой следующую архитектуру: первый линейный слой преобразует входной вектор размером `input_dim` (размер TF-IDF-вектора) в скрытый слой из 512 нейронов. Второй слой уменьшает размерность до 256 нейронов, извлекая более абстрактные признаки. Финальный слой выводит логиты для 5 классов (например, оценки от 1 до 5).
2. **Архитектура ResNet-18** - сверточная нейросеть. Данная архитектура использовалась для классификации изображений.
3. В качестве третьей архитектуры была использована **TinyBERT**, которая представляет собой усеченную версию BERT .

Результаты экспериментов



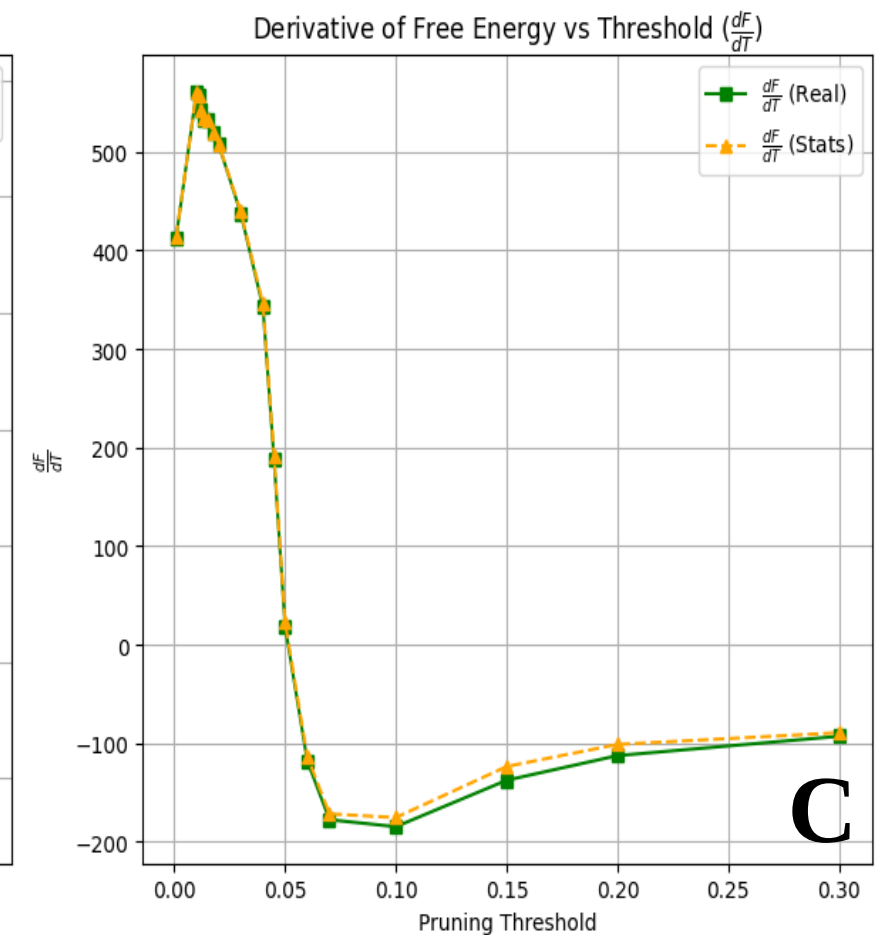
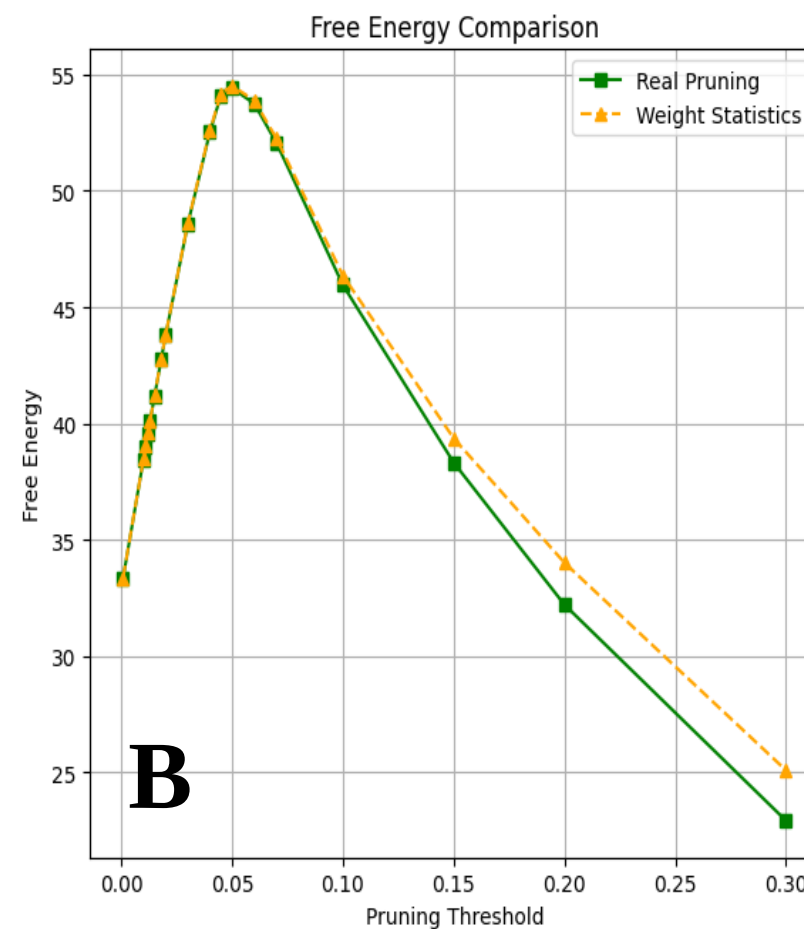
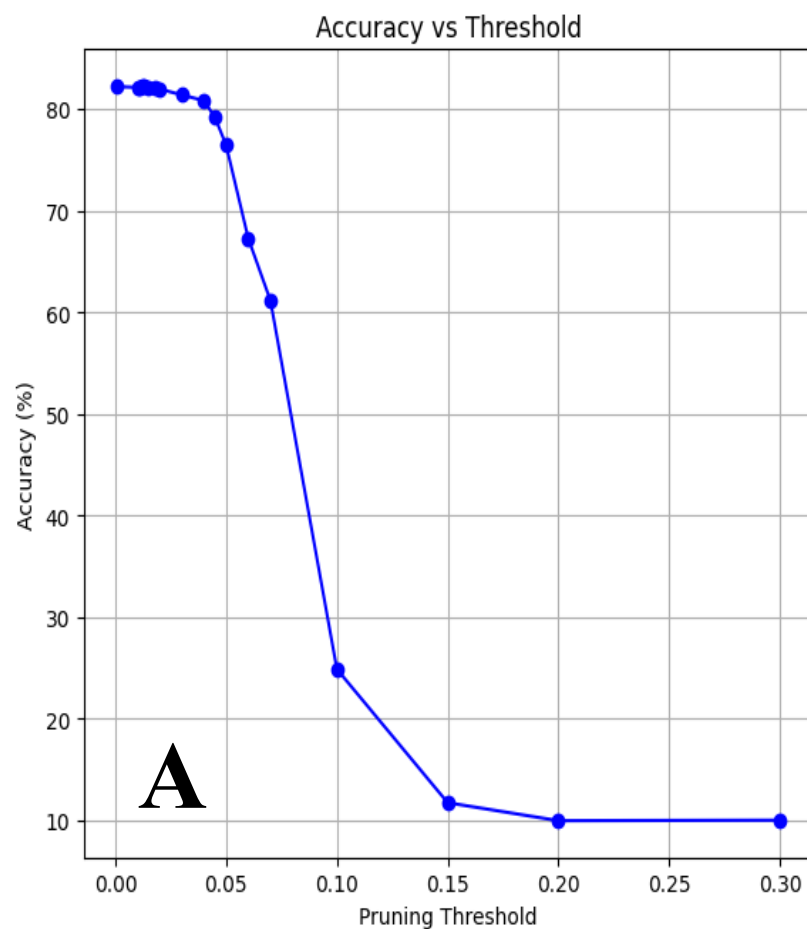
Кривые точности, свободной энергии и производной свободной энергии в зависимости от уровня обрезания в полносвязанной нейронной сети на изображениях.

Результаты экспериментов



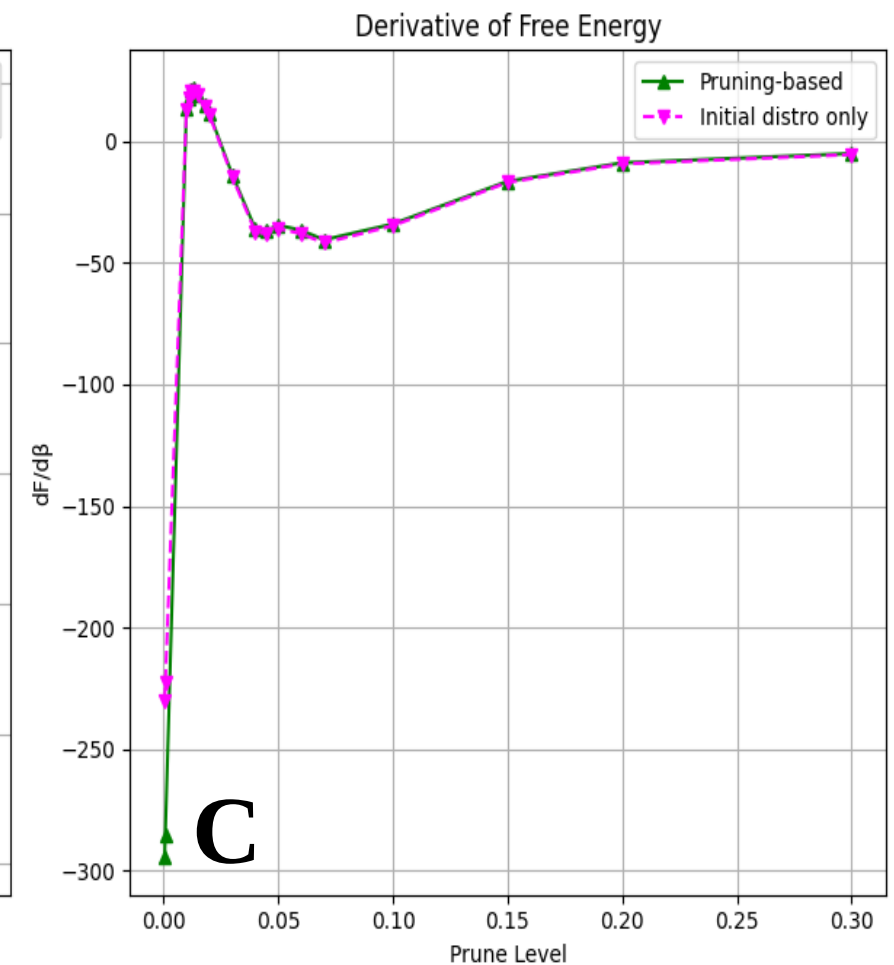
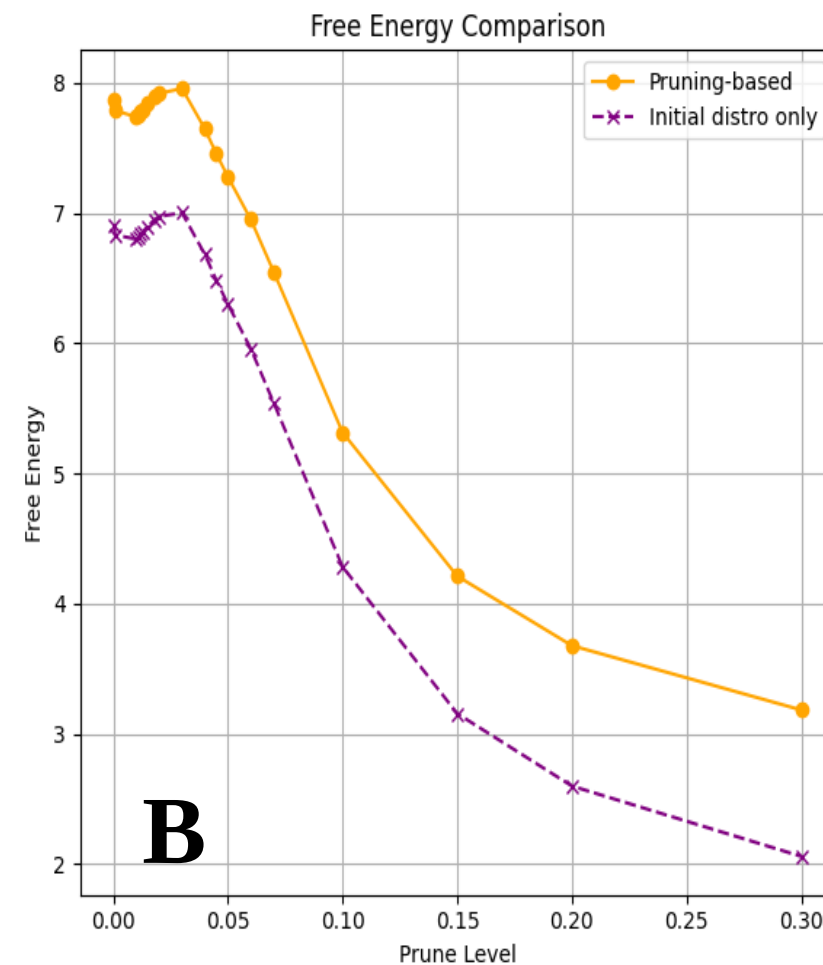
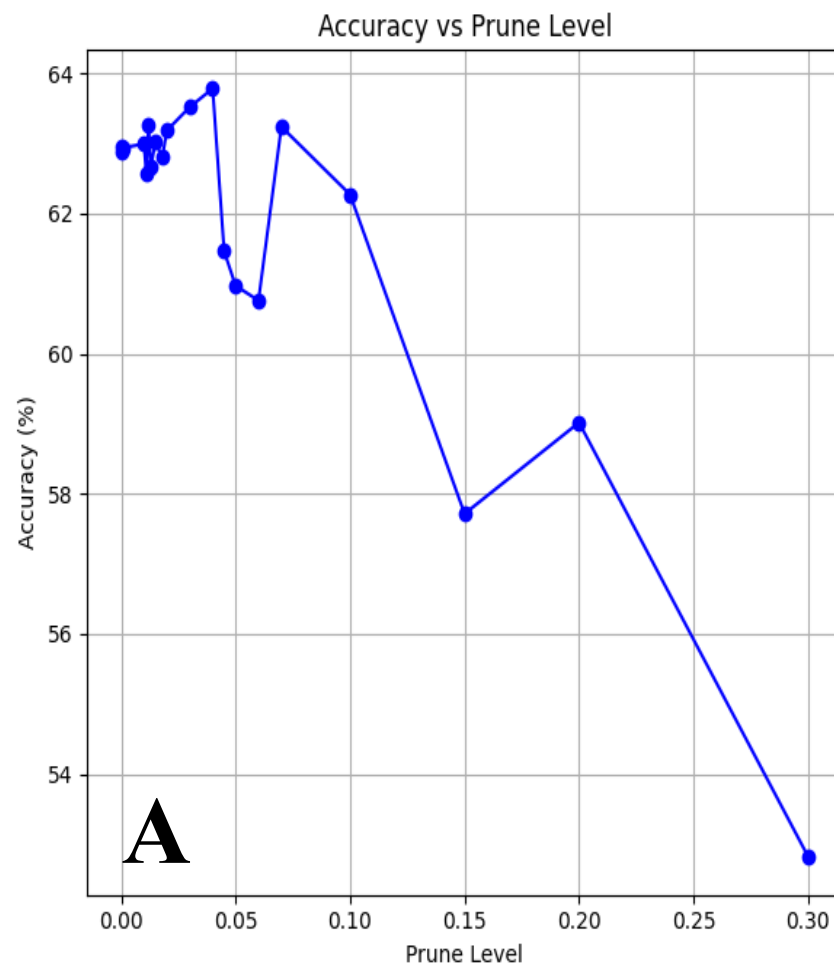
Кривые точности, свободной энергии и производной свободной энергии в зависимости от уровня обрезания в полносвязанной нейронной сети на текстовых данных.

Результаты экспериментов



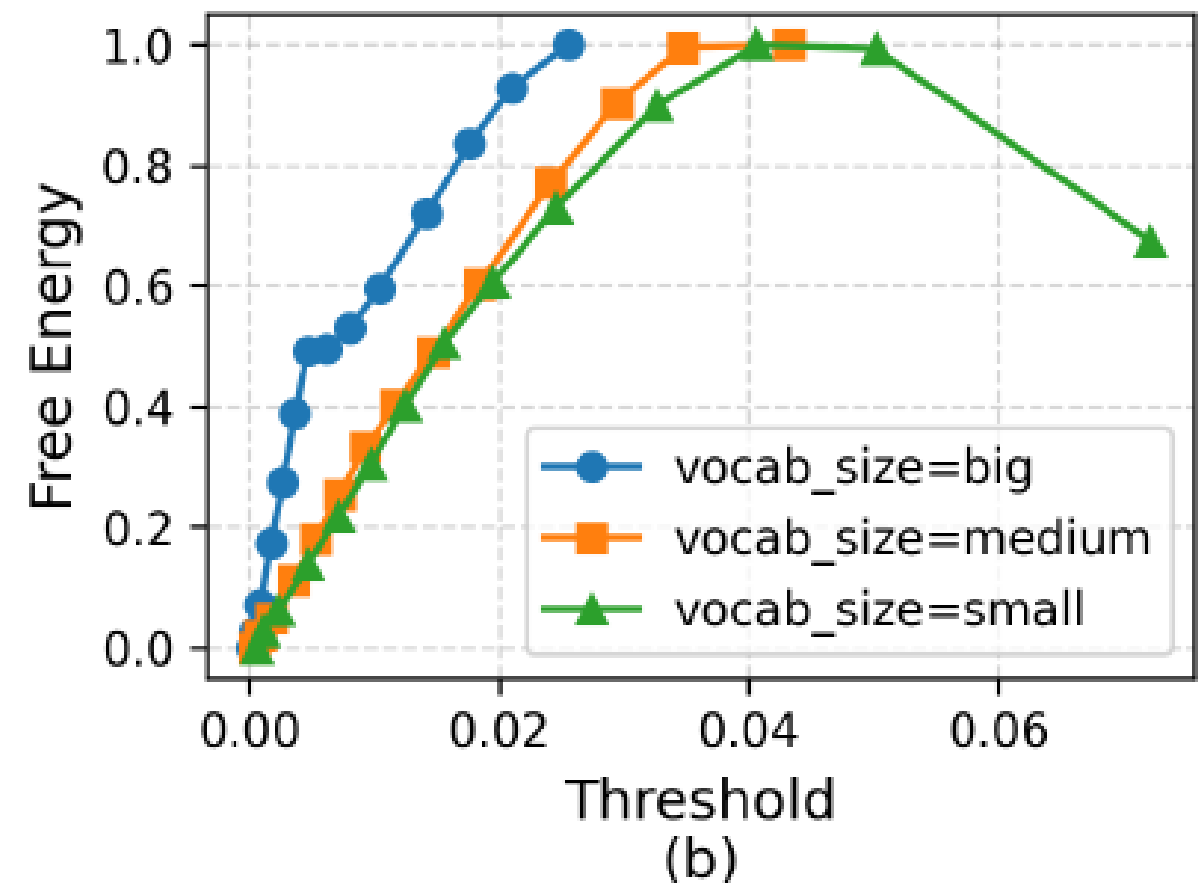
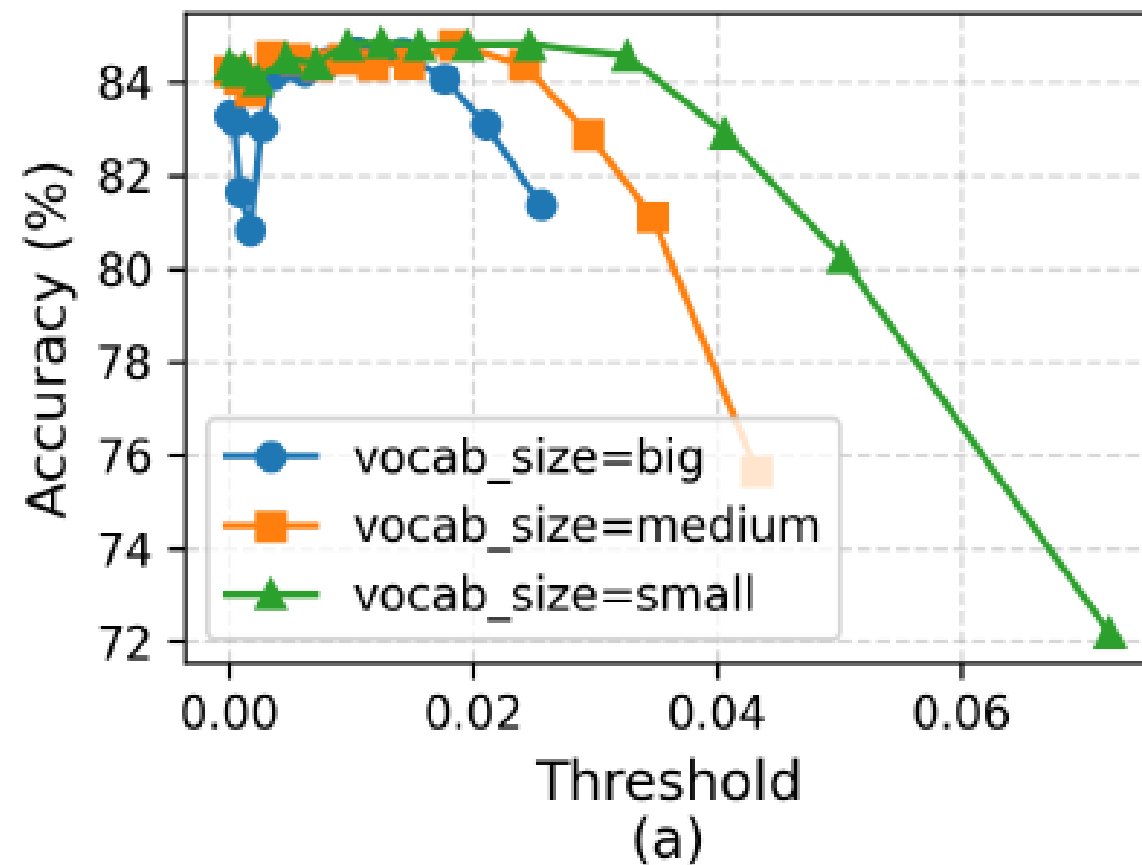
Кривые точности, свободной энергии и производной свободной энергии в зависимости от уровня обрезания в ResNet18 сети на изображениях.

Результаты экспериментов



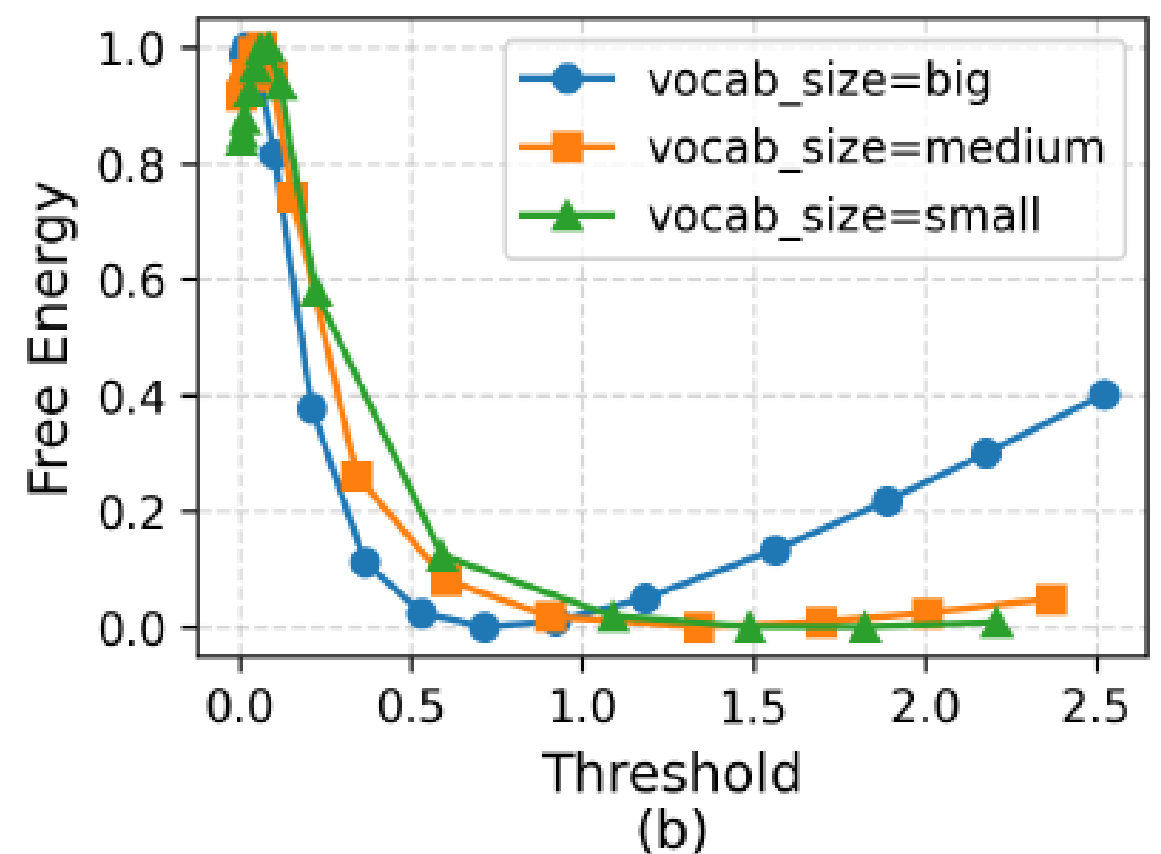
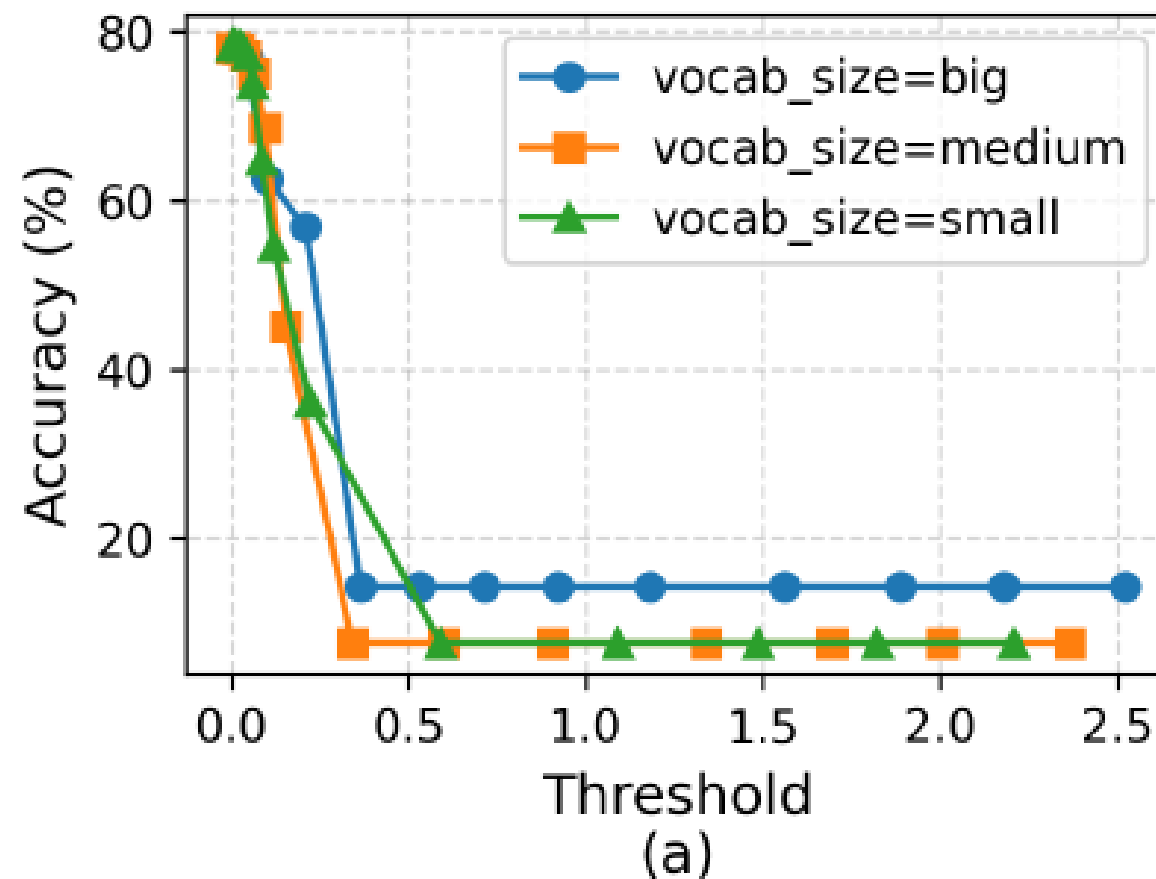
Кривые точности, свободной энергии и производной свободной энергии в зависимости от уровня обрезания в TinyBERT сети на текстах.

Дополнительные результаты экспериментов



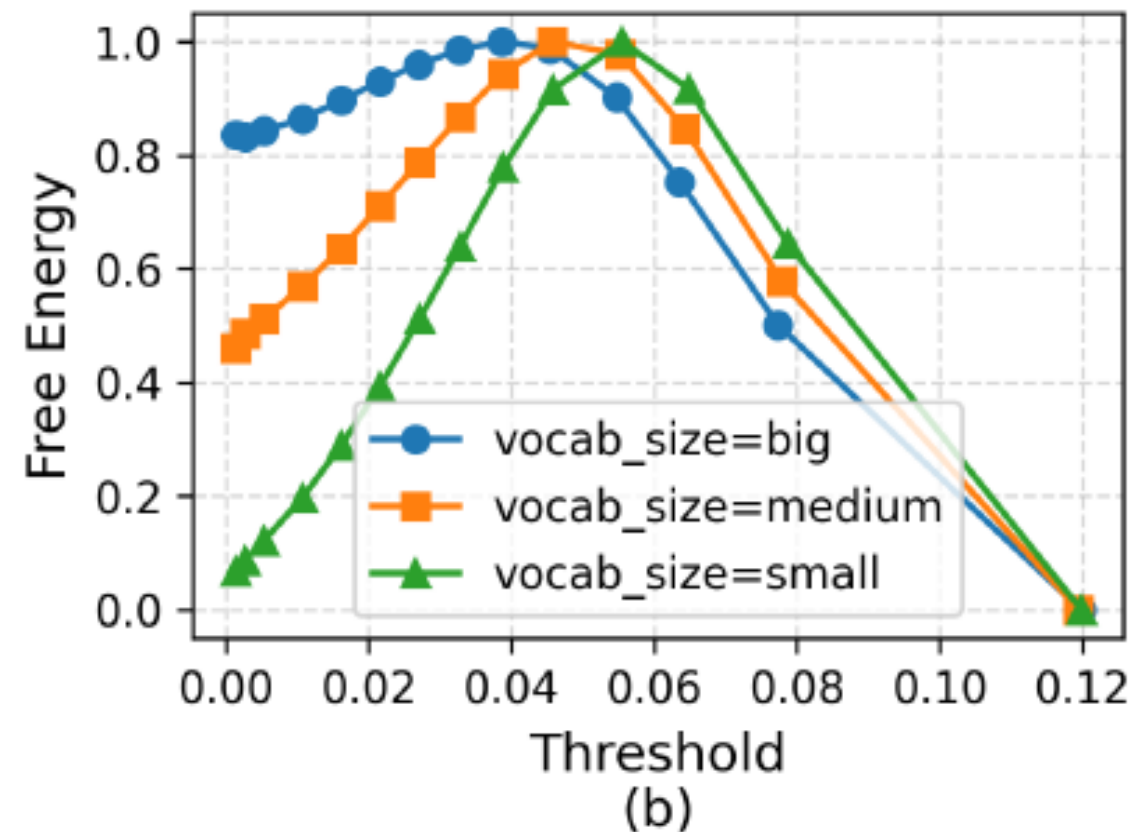
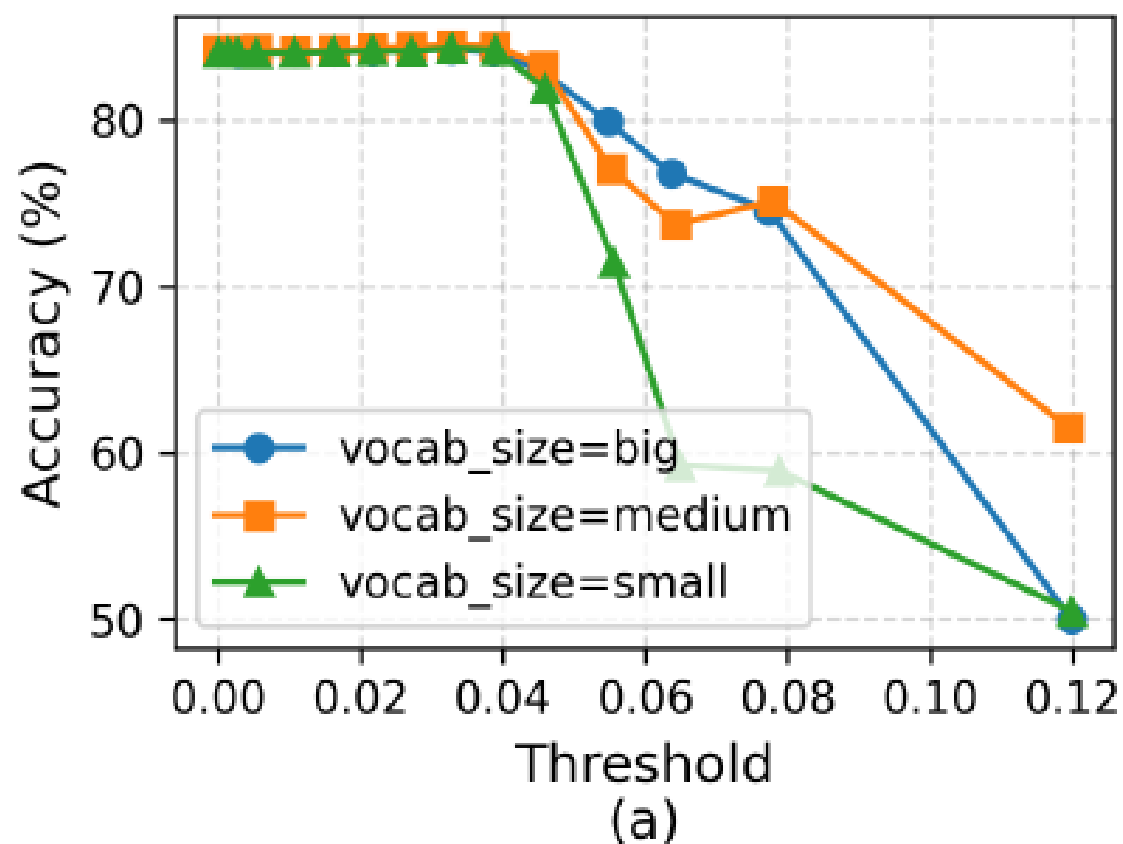
Многослойная нейронная сеть прямого распространения. IMDB dataset

Дополнительные результаты экспериментов



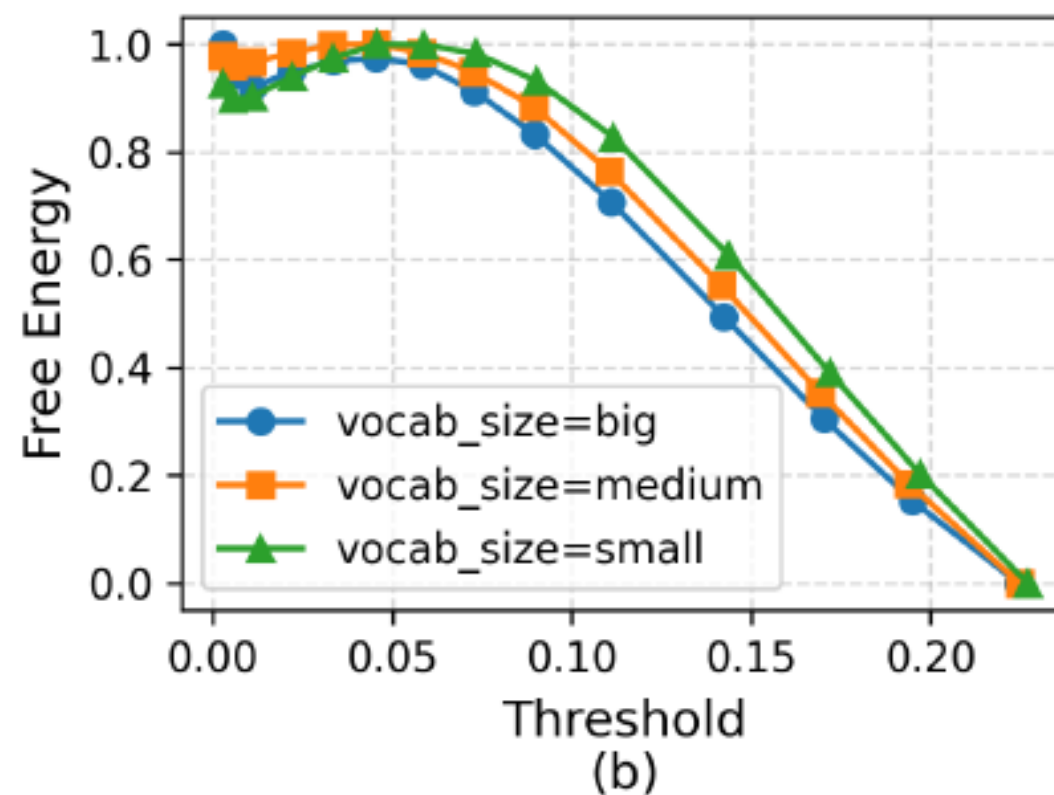
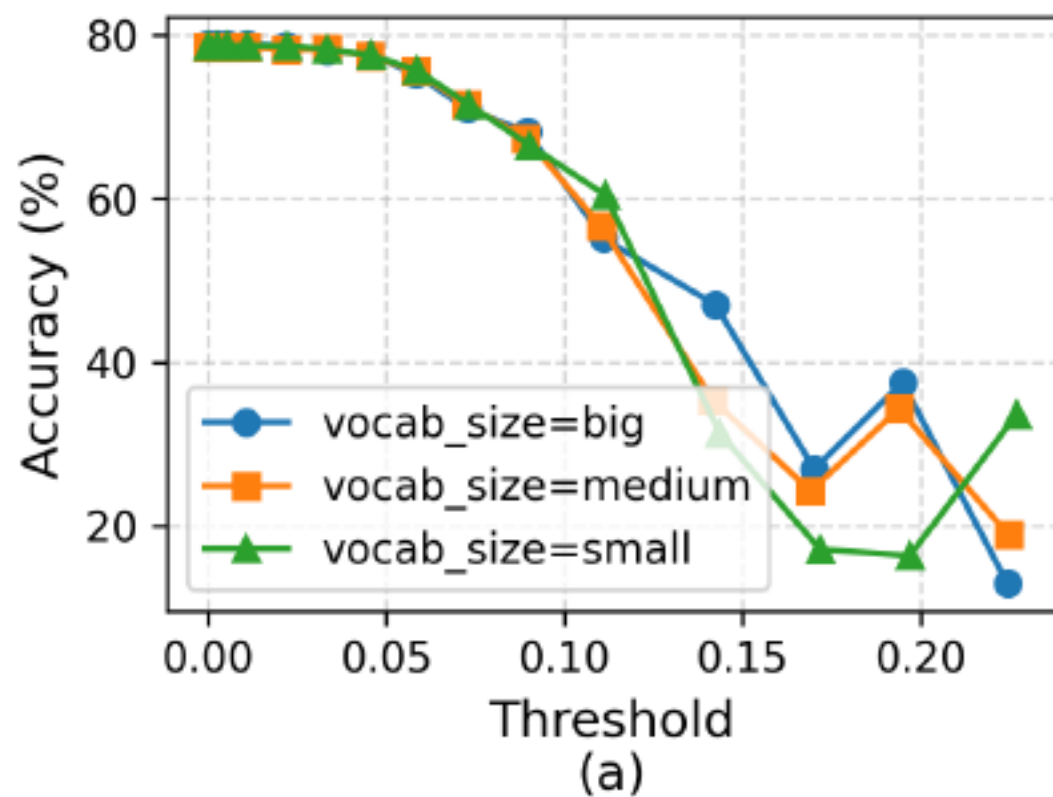
Encoder-decoder model with embedding pruning for Reviews dataset

Дополнительные результаты экспериментов



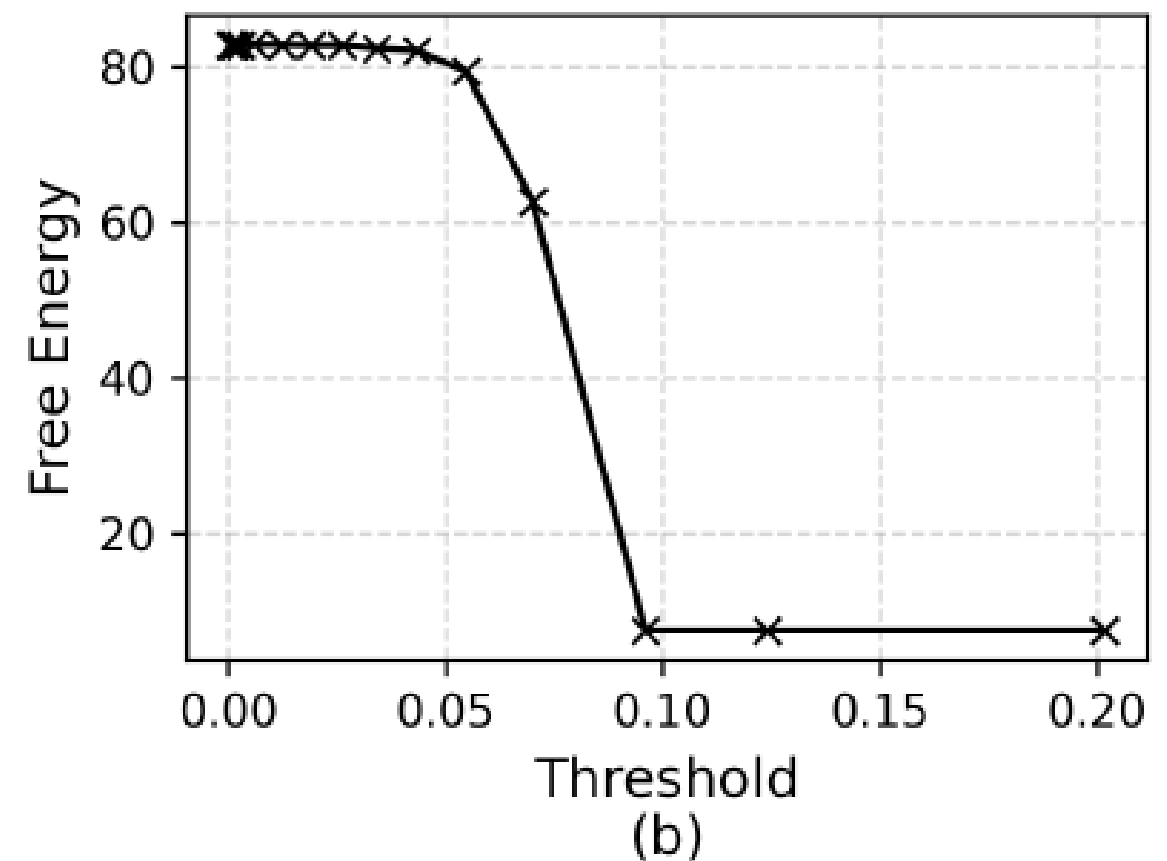
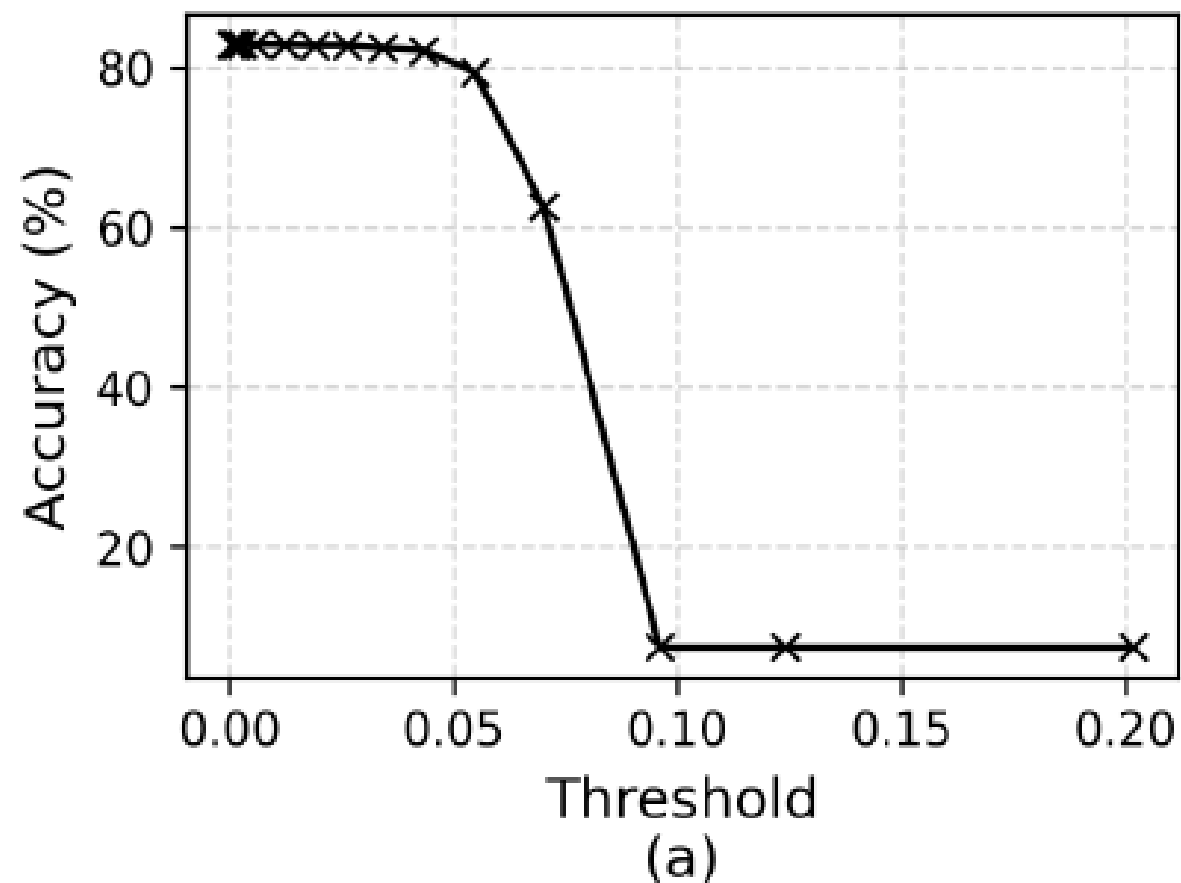
Encoder-decoder model for IMDB dataset

Дополнительные результаты экспериментов



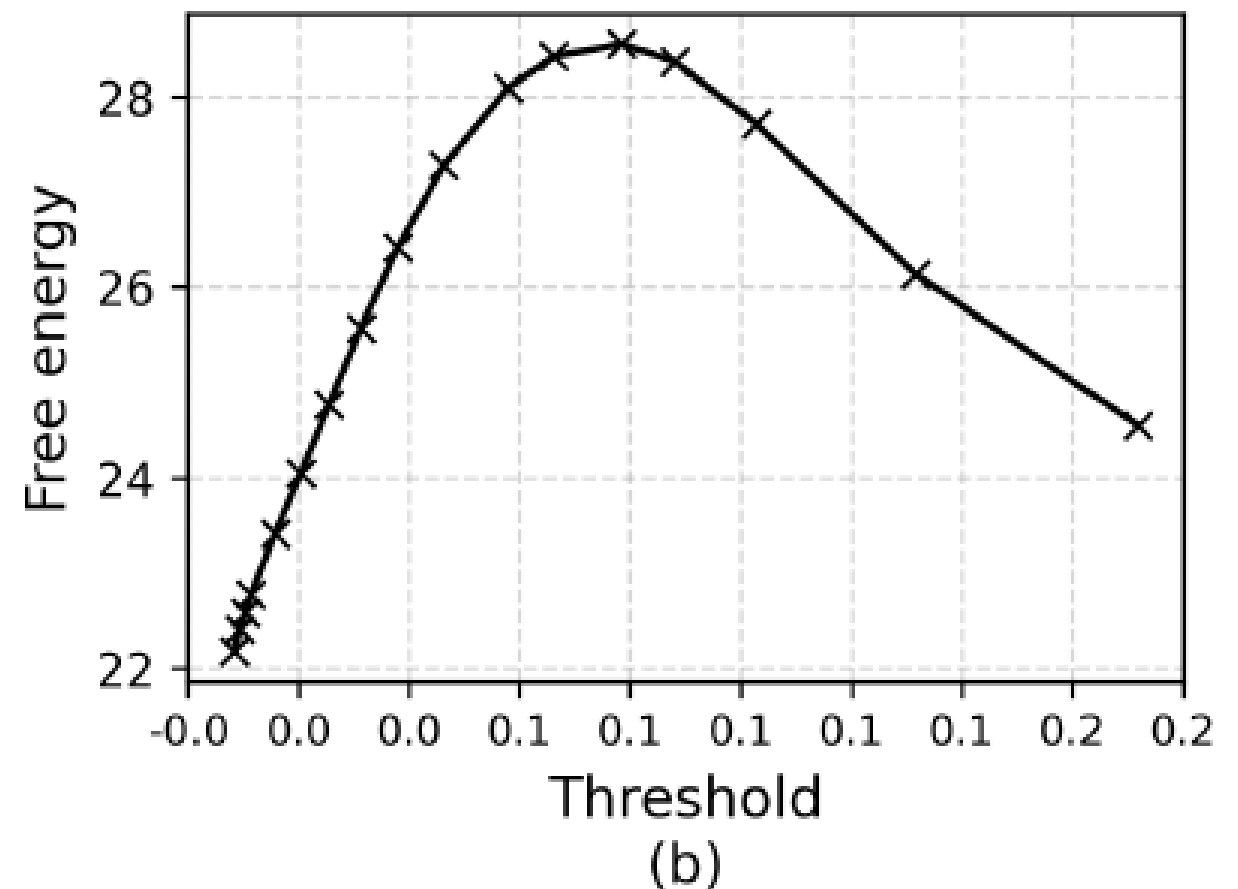
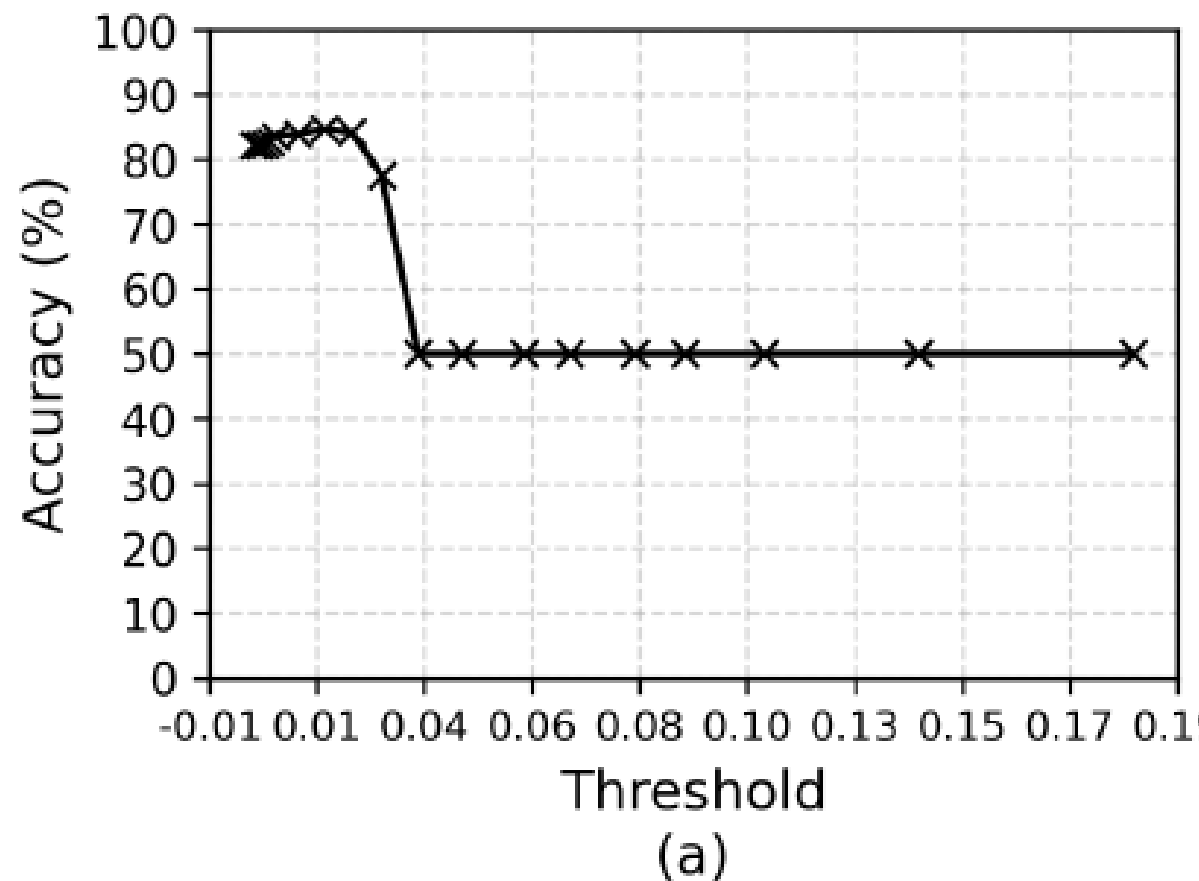
Encoder model for Reviews dataset.

Дополнительные результаты экспериментов



Pretrained encoder model for Reviews dataset

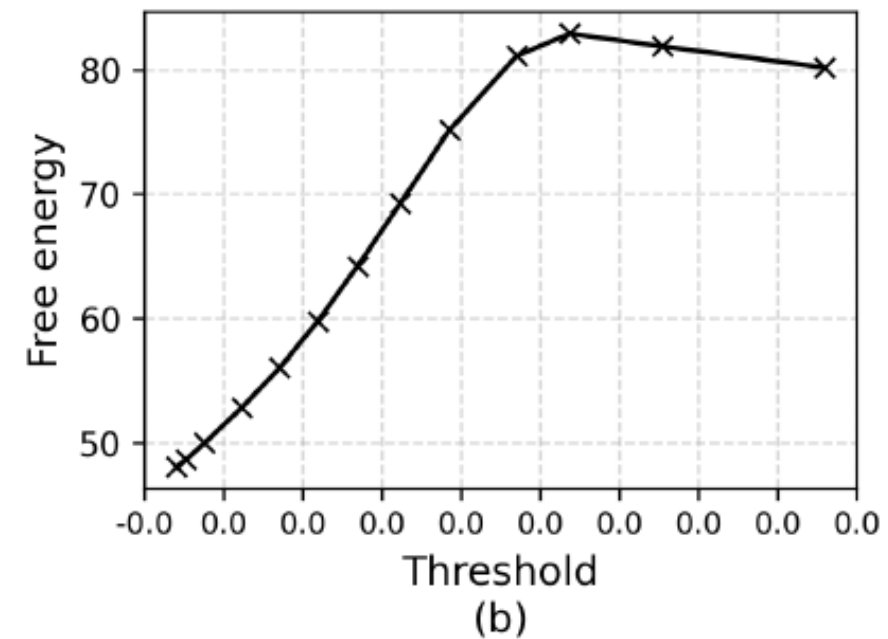
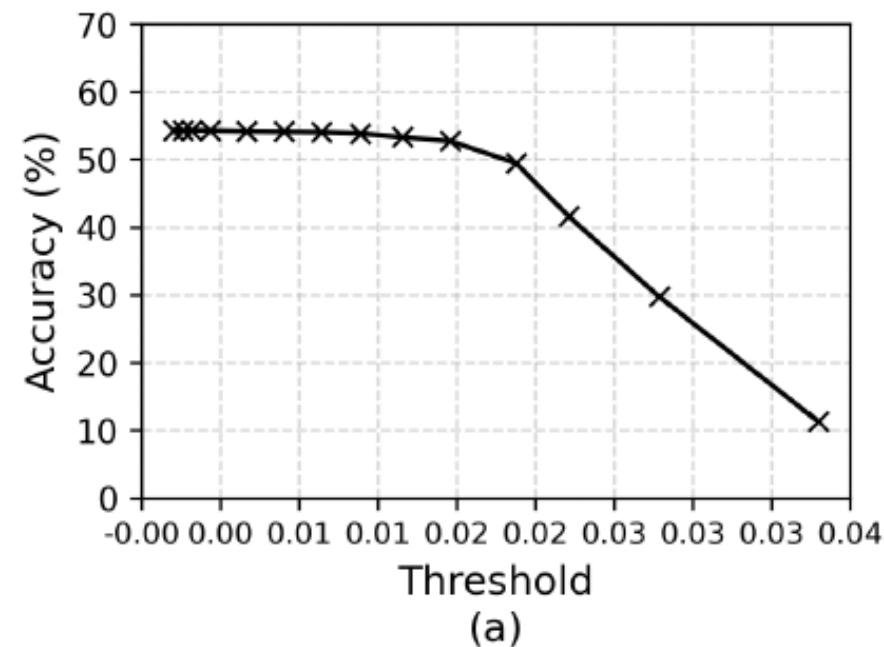
Дополнительные результаты экспериментов



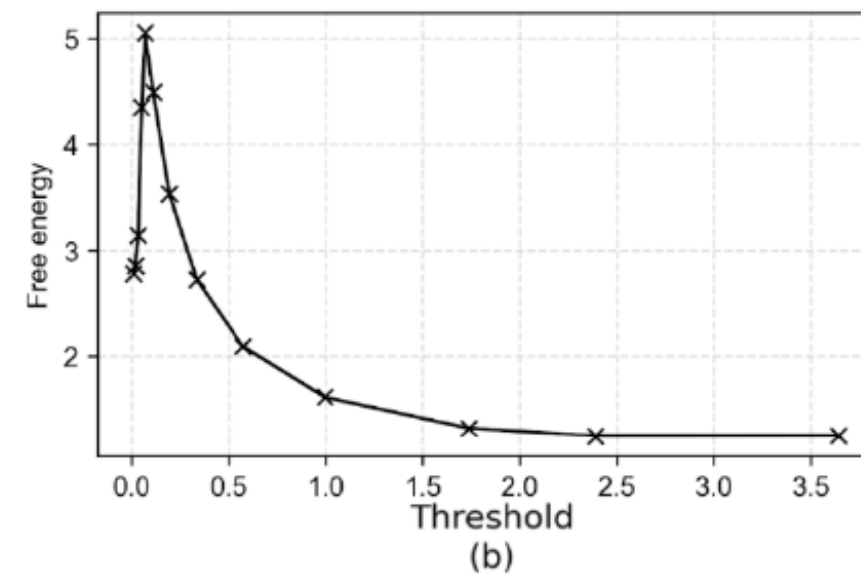
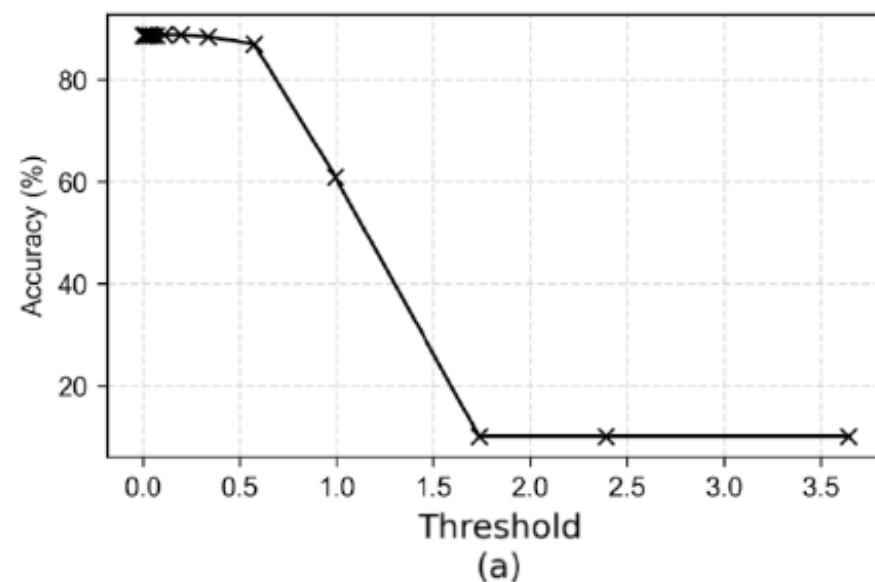
Pretrained encoder model with embedding pruning for Reviews dataset.

Дополнительные результаты экспериментов

MLP model for Cifar-10 dataset.



VGG-11 model for Cifar-10 dataset





Ренормализация VS pruning

Model	Dataset	Accuracy	Free energy
MLP	Reviews	514s	1s
MLP	IMDB	112s	1s
Encoder-decoder transformer	Reviews	494s	0.7s
Encoder-decoder transformer	IMDB	135s	0.7s
Encoder-only transformer	Reviews	189s	0.7s
Encoder-only transformer	IMDB	50s	0.7s
Pretrained transformer	Reviews	1532s	1.1s
Pretrained transformer	IMDB	481s	1.1s

Calculation time in seconds (s).



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

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

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru