Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Кольцов Сергей Николаевич

ЭНТРОПИЙНЫЕ ТЕМАТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ ИХ АГРЕГИРОВАНИЯ

РЕЗЮМЕ
диссертации на соискание ученой степени
доктора компьютерных наук

Санкт-Петербург – 2022

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования ‘Национальный исследовательский университет «Высшая Школа Экономики»’

Научный консультант: Миркин Борис Григорьевич, доктор технических наук, ведущий научный сотрудник Международного центра анализа и выбора решений, профессор факультета компьютерных наук, департамент анализа данных и искусственного интеллекта НИУ ВШЭ.

Введение

Анализ больших текстовых данных стал одним из самых востребованных научных направлений в современном мире в связи с развитием электронных средств хранения и передачи таких данных. По своему объему эти данные становятся сравнимыми с физическими мезоскопическими системами, в связи с чем, для их анализа можно использовать модели машинного обучения, на основе математического формализма, заимствованного из статистической физики. Примером такой модели является тематическое моделирование (topic modeling), на основе процедуры сэмплирования, где используется формализм модели Потса для вычисления распределения слов по темам [15]. Задачей тематического моделирования является выделение набора распределений наблюдаемых величин – текстов или изображений и их элементов – по скрытым переменным, называемым темами.

На данный момент разработано множество тематических моделей [1], с различными способами определения скрытых распределений и разными мерами анализа качества, однако при достаточно широком применении этих моделей в различных областях, нерешенным остается ряд проблем, которые ограничивают использование ТМ.

Одной из основных является проблема определения количества компонент в смеси распределений, поскольку параметр, определяющий размер смеси в модели, должен задаваться явным образом в тематических моделях. Следует отметить, что в тематическом моделировании разработан подход, в котором авторы постулируют автоматический подбор числа тем, однако такие модели обладают множеством скрытых параметров, которые существенно влияют на результаты моделирования [2]; более того эта модель не способна корректно определить число тем в датасете.

Второй проблемой является нестабильность тематического моделирования, которая выражается в том, что результаты тематического моделирования не являются идентичными друг другу, при разных запусках модели на одном и том же датасете и с одними и теми же параметрами. Данная проблема, с одной стороны, связана с неоднозначностью матричных разложений (для тематических моделей на основе E-M алгоритма), а с другой стороны, с наличием множества локальных минимумов и максимумов в подинтегральной функции (для тематических моделей на основе процедуры сэмплирования Гиббса).

Третья нерешенная проблема, вытекающая из второй, связана с разработкой процедур регуляризации, которые могут использоваться как для улучшения стабильности ТМ, так и для других целей [1]. Под регуляризацией понимается добавление априорной информации в тематические модели в виде различных связей и ограничений, что приводит к уменьшению возможного числа решений. На данный момент, в литературе предложено большое число генеративных моделей с регуляризаторами, однако нет ясного критерия выбора комбинации регуляризаторов и подбора коэффициентов регуляризации.

Вышеуказанные проблемы естественным образом влияют на качество тематического моделирования. На данный момент, основными мерами определения качества тематических моделей являются энтропия Шеннона, дивергенция Кульбака-Лейблера, логарифм правдоподобия и перплексия (perplexity). Кроме того, известно, что распределения слов, по крайней мере, в европейских языках, имеют степенной характер, свойственный именно сложным статистическим системам. Также известно, что поведение сложных систем эффективнее исследовать с помощью методов, развиваемых в рамках математического формализма, заимствованного из теории сложных систем.

Цели и задачи исследования

Цель диссертации – разработка и исследование нового класса вычислительных тематических моделей – энтропийных тематических моделей, ориентированных на продвижение в решении проблем определения оптимальных гиперпараметров тематических моделей, включая определение наличия плоских или иерархических структур в датасетах, и разработке стабильных моделей кластеризации текстовых коллекций.

Полученные результаты:

1. Энтропийная тематическая модель на основе однопараметрической энтропии (энтропии Реньи и Тсаллиса). Данная модель реализована для следующих генеративных алгоритмов: 1) LDA (Gibbs Sampling algorithm), 2) pLSA (E-M algorithm), 3) VLDA (E-M algorithm), 4) GLDA (Gibbs Sampling algorithm).
2. Энтропийная тематическая модель на основе двухпараметрической энтропии Шарма−Миттала (Sharma–Mittal Entropy). Данная модель реализована для следующих генеративных моделей: 1) pLSA (E-M algorithm). 2) LDA (Gibbs Sampling algorithm). 3) ARTM с регуляризаторами разреживания матриц Φ и Θ (E-M algorithm).
3. Иерархическая энтропийная тематическая модель. Данная модель реализована для генеративных иерархических моделей: 1) hLDA. 2) hPAM. 3) hARTM. 4) алгоритм кластерного анализа HCA ('complete' method).
4. Фрактальная модель оценки работы генеративных тематических моделей. Данная модель реализована для следующих алгоритмов: 1) pLSA (E-M algorithm), 2) ARTM (E-M algorithm), 3) LDA Gibbs sampling algorithm.
5. Метод агрегации тематических моделей на основе процедуры ренормализации. Метод реализован для следующих алгоритмов: 1) VLDA (E-M algorithm). 2) LDA (Gibbs sampling algorithm). 3) pLSA (E-M algorithm).
6. Метод агрегации реализован для трех вариантов объединения тем: 1) Объединение на основе минимума энтропии Реньи. 2) Случайное объединения тем. 3) Объединение на основе минимума дивергенции Кульбака-Лейблера.
7. Гранулированная тематическая модель на основе процедуры сэмплирования Гиббса. Данная модель реализована для трех вариантов функции локального распределения тем: 1) GLDA. 2) ELDA. 3) TLDA.

Личный вклад автора включает в себя:

- общая математическая формулировка энтропийной модели на основе однопараметрической энтропии Реньи, опубликованная в двух статьях одним автором.
- организация и участие в проведении широкомасштабных компьютерных экспериментов по анализу применимости энтропийной модели для оценки качества работы тематических моделей разного типа.
- лидирующее участие в математической формулировке энтропийной модели на основе двухпараметрической энтропии Шарма−Миттала и проверке этой модели в серии компьютерных экспериментов.
- формулировка фрактальной модели оценки работы генеративных тематических моделей и проведение компьютерных экспериментов по тестированию данной модели.
- математическая формулировка метода агрегации тематических моделей с помощью процедуры ренормализации и проведение компьютерных экспериментов по проверке эффективности процедуры ренормализации.
- общая математическая формулировка гранулированного метода сэмплирования.

По теме данной диссертации опубликовано 8 статей в журналах уровня Q1-Q2 WoS: а также 11 статей, индексируемых в Scopus.

Научная новизна:

1. Впервые предложено использование однопараметрической энтропии Реньи и двухпараметрической энтропии Шарма−Миттала для целей оптимизации работы тематических моделей.
2. Впервые показано, что мера качества на основе параметризованных энтропий превосходит традиционные меры, такие как логарифм правдоподобия или перплексия, так они позволяют одновременно настраивать величины гиперпараметров тематических моделей и количество распределений в смеси.
3. Впервые предложена фрактальная модель оценки работы генеративных тематических моделей, демонстрирующая самоподобное поведение тематических моделей, что позволяет применять к ним процедуру ренормализации.
4. Впервые предложена процедура ренормализации тематических моделей и продемонстрирована ее эффективность для скоростного определения оптимального количества распределений в смеси.
5. Предложен гранулированный вариант тематической модели, которая превосходит по своей стабильности другие тематические модели.

Публикации повышенного уровня (Q1-Q2 по WOS и Scopus)

1. Koltcov, S. Analysis and tuning of hierarchical topic models based on Renyi entropy approach / Koltcov, S., Ignatenko, V., Terpilovskii, M., Rosso, P. // PeerJ Computer Science, Том 7, 2021. Open access**:** https://peerj.com/articles/cs-608/
2. Koltsov S., Ignatenko V., Boukhers Z., Staab S. Analyzing the Influence of Hyper-parameters and Regularizers of Topic Modeling in Terms of Renyi entropy // *Entropy*. 2020. Vol. 22. No. 4.pp. 1-13. doi
3. Koltcov S, Ignatenko V. Renormalization Analysis of Topic Models // *Entropy*. 2020. Vol. 22. No. 5. pp. 1-23. doi
4. Koltsov S., Ignatenko V., Koltsova O. Estimating Topic Modeling Performance with Sharma–Mittal Entropy // *Entropy*. 2019. Vol. 21. No. 7. pp. 1-29. doi
5. Koltsov S. Application of Rényi and Tsallis entropies to topic modeling optimization // *Physica A: Statistical Mechanics and its Applications*. 2018. Vol. 512. pp. 1192-1204. doi
6. Koltcov, S.N. A thermodynamic approach to selecting a number of clusters based on topic modeling / Koltcov, S.N. // Technical Physics Letters. 2017. Т. 43. № 12. С. 90-95. doi
7. S. N. Koltsov, S. I. Nikolenko, and E. Yu. Koltsova Gibbs Sampler Optimization for Analysis of a Granulated Medium // Pis'ma v Zhurnal Tekhnicheskoi Fiziki, 2016, Vol. 42, No. 16, pp. 21–25.
8. Sergey Nikolenko, Sergei Koltcov, Olessia Koltsova. Topic modelling for qualitative studies // *Journal of Information Science*. 2017. Vol. 43. No. 1. pp. 88-102. doi

Публикации стандартного уровня по теме исследования (Scopus)

1. Koltsov S., Ignatenko V., Pashakhin S. How many clusters? An Entropic Approach to Hierarchical Cluster Analysis, in: *Intelligent Computing: SAI 2020: Volume 3* Vol. 1230. Book 3. Cham : Springer, 2020. doi pp. 560-569. doi
2. Koltsov S., Ignatenko V. Renormalization approach to the task of determining the number of topics in topic modeling, in: *Intelligent Computing: SAI 2020: Volume 1* Vol. 1228. Part 1. Switzerland : Springer, 2020. pp. 234-247. doi
3. Ignatenko V., Sergei Koltcov, Staab S., Boukhers Z. Fractal approach for determining the optimal number of topics in the field of topic modeling // *Journal of Physics: Conference Series*. 2019. Vol. 1163. No. 1. pp. 1-6. doi
4. Koltsov S., Pashakhin S., Dokuka S. A Full-Cycle Methodology for News Topic Modeling and User Feedback Research, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics). 10th International Conference on Social Informatics, SocInfo 2018; St.Petersburg. Cham: Springer, 2018. pp. 308-321. doi
5. Mavrin A., Filchenkov A., Koltsov S. Four Keys to Topic Interpretability in Topic Modeling, in: *Artificial Intelligence and Natural Language, 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings* Issue 930. Switzerland : Springer, 2018. doi pp. 117-129. doi
6. Koltsov S., Nikolenko S. I., Koltsova O., Filippov V., Bodrunova S. Stable Topic Modeling with Local Density Regularization, in: *Internet Science, Proc. of 3d conf INSCI 2016, Lecture Notes in Computer Science series* Vol. 9934. Switzerland : Springer, 2016. doi pp. 176-188. doi
7. Koltsov S., Nikolenko S. I., Koltsova O., Bodrunova S. Stable topic modeling for web science: Granulated LDA, in: *WebSci 2016 - Proceedings of the 2016 ACM Web Science Conference*. Elsevier, 2016. pp. 342-343. doi
8. Koltsov S., Koltsova O., Nikolenko S. I. Latent Dirichlet Allocation: Stability and Applications to Studies of User-Generated content, in: *Proceedings of WebSci '14 ACM Web Science Conference, Bloomington, IN, USA — June 23 - 26, 2014*. NY : ACM, 2014. pp. 161-165.
9. Nikolenko S. I., Koltsov S., Koltsova O. Measuring Topic Quality in Latent Dirichlet Allocation, in: *Proceedings of the Philosophy, Mathematics, Linguistics: Aspects of Interaction 2014 Conference*. St. Petersburg : The Euler International Mathematical Institute, 2014. pp. 149-157.
10. Koltsov S., Ignatenko V., Pashakhin S. Fast Tuning of Topic Models: An Application of Rényi Entropy and Renormalization Theory, in: *Proceedings of the 5th International Electronic Conference on Entropy and Its Applications* Vol. 46. Issue 1. MDPI AG, 2020. Ch. 5. pp. 1-8.
11. Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A. Interval Semi-supervised LDA: Classifying Needles in a Haystack, in: *Proceedings of the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013)* Part I: Advances in Artificial Intelligence and Its Applications. Berlin : Springer, 2013. pp. 265-274.

1. Аналитический обзор научной литературы

1.1. Подходы к проблеме выбора числа кластеров

В обзоре рассматриваются исследования, наиболее интересные и полезные для данной диссертационной работы. Главная проблема в поиске оптимального числа кластеров в кластерном анализе и тематическом моделировании состоит в выборе функции, на основе которой осуществляется такой поиск. Обсуждение множества мер качества кластеризации, включая функции для выбора числа кластеров, производится в работах [10, 11]. Эти и другие работы показывают, что для данных целей в кластерном анализе часто используется минимальное внутрикластерное расстояние. Однако проблема этой и сходных мер качества заключается в том, что функция зависимости такой меры от числа кластеров является монотонно возрастающей (или убывающей). Соответственно, требуется разработка процедур трансформации для выделения особенностей из таких функций. В работе [3] был сформулирован алгоритм определения оптимального числа тем на основе ‘rate distortion theory’. Модернизация данного подхода в рамках неэкстенсивной статистической физики для кластеризации изображений была реализована в работе [4].

В кластерном анализе существуют и другие подходы к решению данной проблемы [5, 6, 7]. В работе Тибширане [6] предложен подход, называемый ‘gap statistic’. Его ключевой идеей является измерение разницы между базовым распределением (null reference distribution) и распределением, которое получается в ходе кластеризации из этого же распределения. Данная разница измеряется при различном числе кластеров, после чего строится соответствующая кривая. В рамках данного подхода авторы предполагают, что оптимальное число кластеров соответствует ситуации, когда логарифм от среднего внутри-

кластерного расстояния падает ниже, чем аналогичный логарифм, рассчитанный по null reference distribution. По сути дела, это аналог процедуры измерения зависимости энтропии от числа кластеров по отношению к исходной энтропии. В работе [8] предлагается процедура кластеризации, основанная на принципе поиска максимума энтропии (maximum entropy principle), при этом авторы также опираются на классический вариант энтропии (энтропия Шеннона). Однако уже в работе [9], метод кластеризации реализован с применением принципа максимизации энтропии Тсаллиса, за счет вариации параметра q.

Из всех существующих подходов в кластерном анализе наиболее интересным и содержательным является подход на основе минимизации свободной энергии [12]. Его основная идея заключается в следующем: каждый элемент статистической системы характеризуется вероятностями принадлежности к различным кластерам. Соответственно, для каждого элемента можно сформулировать понятие внутренней энергии (выраженной через вероятность принадлежности элемента кластеру) и рассчитать свободную энергию всей системы. Температура в такой системе становится свободным параметром, который варьируется с целью минимизации свободной энергии. Недостатком данной работы является тестирование модели только на кластерах с гауссовыми распределениями. Кроме того, как показывают вычисления авторов, функция свободной энергии выглядит монотонной функцией без явного минимума.

Данная диссертационная работа базируется на идеях работы [12], но в отличие от нее, во-первых, температура рассматривается как число кластеров, во-вторых, вместо свободной энергии рассматриваются параметризованные энтропии, которые обладают явным минимумом. Теоретические положения энтропийных тематических моделей изложены в главе 2.

1.2. Обзор типов моделей в тематическом моделировании

На данный момент в научной литературе по тематическому моделированию предложено более сорока различных тематических моделей, а число статей, в которых используется тематическое моделирование, превышает несколько сотен. В целом можно выделить три основных вида моделей: 1. Плоские модели (plain topic models) с различными видами регуляризации [13, 14, 15, 16]. 2. Иерархические тематические модели [17, 18, 19, 20]. 3. Тематические модели с элементами нейронных сетей, в которых используются либо эмбединги разного типа, либо слои нейронных сетей [21, 22]. Наиболее полные обзоры разновидностей моделей и мер качества представлены в работах [1, 23]. В целом, в научной литературе доминируют два основных алгоритма определения распределения слов по темам и тем по документам: 1. Алгоритм Expectation-Maximization (E-M алгоритм). В рамках него матрица слов в документах F представлена в виде произведения двух матриц $F = \Phi\Theta$, где Φ – матрица распределения слов по темам, Θ – матрица распределения тем в документах. 2. Алгоритм определения вероятности принадлежности слова к теме в виде многомерного интеграла. В данном алгоритме вычисление вероятности реализовано с помощью процедуры сэмплирования Гиббса. Несмотря на разный математический формализм указанных алгоритмов, оба они дают сходные результаты [24]. Поэтому ниже рассматриваемые проблемы, справедливые для разных алгоритмов.

Для тематического моделирования проблема поиска оптимального числа тем/кластеров также актуальна и даже более сложна для решения. Это обусловлено следующими причинами. Во-первых, этот поиск связан с лингвистическим определением понятия темы, что вызывает значительные трудности, так как сложно сформулировать лингвистический критерий разделения двух тем на множестве документов. Кроме того, тематические модели часто генерируют трудно интерпретируемые темы, которые сложно рассматривать как темы. Во-вторых, в тематическом моделировании, так же, как и в кластерном анализе, сложно сформулировать адекватную функциональную зависимость, которая, с одной стороны, характеризовала бы тематическую модель, а, с другой стороны, являлось бы функцией от числа тем и гиперпараметров. Тем не менее, существует

несколько работ, в которых авторы попытались решить проблему выбора числа тем именно в тематическом моделировании. Авторы работы [25], основываясь на идеях кластерного анализа, рассматривали тему как семантический кластер (набор слов), в рамках которого можно рассчитать внутрикластерное расстояние. В качестве функции для минимизации авторы использовали косинусную меру. Таким образом, с точки зрения авторов, число тем, при котором находится минимум средней величины косинусной меры, рассчитанной по всем темам, соответствует оптимуму. Другой подход к поиску оптимального числа тем предложен Аруном и соавторами в работе [26] как поиск минимума дивергенции Кульбака-Лейблера при изменении числа тем. Для этого авторы предлагают делать SVD разложение матриц Φ и Θ, после чего рассчитывать дивергенцию Кульбака-Лейблера на основании двух векторов, содержащих сингулярные величины. В этом случае оптимальное число тем соответствует ситуации, когда обе матрицы описываются одинаковым числом сингулярных величин. Недостатками этих двух подходов являются следующее. Во-первых, не ясно каким образом минимумы выбранных функций соотносятся с энтропийным принципом, широко используемым в теории информации. Во-вторых, добавление еще одного этапа расчета, а именно SVD разложения и расчета дивергенции Кульбака-Лейблера, существенно ограничивает применение подхода Аруна к обработке больших данных. Арун и коллеги искали минимум дивергенции Кульбака-Лейблера на текстовых коллекциях, не превышающих 2500 текстов. В-третьих, в обоих выше изложенных подходах отсутствует учет влияния исходного распределения на результаты тематического моделирования, хотя известно, что такое влияние есть [27]. В-четвертых, в выше указанных подходах не принимается во внимание эффект семантической нестабильности, который присущ тематическим моделям [28].

Отдельно следует выделить тематическую модель на основе алгоритма аддитивной регуляризации (ARTM), предложенной в работе [16]. Данная модель основана на поиске максимума линейной комбинации логарифма правдоподобия и ряда регуляризаторов. Уровень влияния регуляризаторов на тематическую модель определяется величиной коэффициентов. Несмотря на широкое применение данной модели в русскоязычной научной литературе, она обладает одним существенным недостатком: в теории аддитивной регуляризации не сформулирован принцип выбора величин коэффициентов регуляризации. Они должны задаваться явным образом перед проведением тематического моделирования. В рамках данной работы предлагается решение этой проблемы.

Одной из основных мер качества тематических моделей является максимум логарифма правдоподобия (‘log-likelihood’) [1] и связанная с ним мера ‘perplexity’. В целом, логарифм правдоподобия позволяет настраивать гиперпараметры ‘плоских’ тематических моделей, однако не дает возможности определять оптимальное количество тематических кластеров. При этом ‘log-likelihood’ мало пригоден для настройки иерархических тематических моделей [2], в которых существуют дополнительная нерешенная проблема выбора числа тем на каждом из уровней иерархии, а также традиционная проблема задания величин гиперпараметров.

Кроме этого в тематических моделях активно используется мера когерентности (‘coherence’), которая позволяет оценить связность тем в тематическом решении [67]. Суть данной меры заключается в подсчете того, насколько часто слова с высокими вероятностями совместно встречаются в высоко вероятностных документах. Высокая когерентность тем соответствует наилучшему решению. Данная мера также не позволяет определить оптимальное количество компонент в смеси распределений, в виду своего монотонного поведения.

Таким образом, тематическим моделям разного типа присущи следующие нерешенные проблемы: 1. Не ясно, каким образом определять оптимальное число кластеров в тематическом решении. 2. Существующие меры качества не являются универсальными, то есть работают не на всех моделях. 3. Не существует меры качества, которые позволяли

бы настраивать несколько параметров модели одновременно (включая гиперпараметры, число кластеров и семантическую связность).

В данной работе предлагается решение указанных проблем за счет использования параметризованных энтропий в тематических моделях. Теоретическая и экспериментальная оценка применения параметризованных энтропий в тематических моделях приведена в главе 2.

1.3. Применение энтропийных принципов в области тематического моделирования

Следующая часть обзора работ посвящена применению модели отжига для выделения скрытых распределений в тематическом моделировании. В работах [29, 30] используется классический вариант алгоритма отжига на основе марковского процесса. В работе Тсаллиса [74] предложен модифицированный вариант отжига, однако в области машинного обучения данный алгоритм не применялся.

В работе Zhu [31] предложена модель ‘maximum entropy discrimination latent Dirichlet allocation (MedLDA)’, суть которой заключается во введении в логарифм правдоподобия дивергенции Кульбака-Лейблера, который является энтропийным регуляризатором. Следует также отметить работу [32], где предложена тематическая модель, в которой для определения оптимальной величины коэффициента регуляризации используется минимум энтропии Шеннона, рассчитанный по словам. Существенным недостатком данной работы является тестирование предложенной модели на датасетах, размеченных всего на две темы.

1.4. Стабильность тематических моделей

Несмотря на множество работ, посвященных тематическим моделям, число работ, связанных с оценкой их стабильности невелико. Проблема стабильности тематических моделей связана с особенностями работы тематических моделей.

Решение задачи тематического моделирования эквивалентно стохастическому матричному разложению, в котором большая матрица F, содержащая документы d и слова w, аппроксимируется произведением двух матриц Θ и Φ меньшей размерности. Однако стохастическое матричное разложение определяется не единственным образом, а с точностью до невырожденного преобразования [16]. Если $F = \Phi\Theta$ – решение, то $F = (\Phi S)(S^{-1}\Theta)$ также является решением для всех невырожденных S, при которых матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ являются стохастическими. В терминах алгоритма ТМ неоднозначность восстановления многомерной плотности смеси распределений связана с тем, что алгоритм, стартуя из различных начальных приближений, будет сходиться к различным точкам из множества решений. Это выражается в том, что при разных запусках алгоритма на одних и тех же исходных данных, содержимое матриц Θ и Φ будет различным. Задачи, решение которых не единственно или неустойчиво, называются некорректно поставленными. Общий подход к их решению даёт регуляризация по Тихонову [38]. Суть регуляризации заключается в доопределении априорной информации, что позволяет сузить множество решений. Регуляризация осуществляется либо за счет введения ограничений на матрицы Θ и Φ [16], либо за счет модификации процедуры сэмплирования [72, 73].

В исследованиях оценки стабильности тематических решений можно выделить несколько работ. В работе Гриффитса и Стайверса [15] предлагается использовать симметричную дивергенцию Кульбака-Лейблера (KLB) для оценки сходства между двумя темами из разных тематических решений. Однако в данной работе не проведено детального исследования применимости данной меры в практических экспериментах. Модифицированный вариант симметричной меры KLB предложен в работе Кольцова и соавторов [28], где также представлен полноценный практический алгоритм оценки стабильности тематических моделей по трем запускам. В работе Белфорда предложена мера ‘Average Descriptor Set Difference (ADSD)’ [39], которая характеризует среднее значение количества совпадающих слов в двух тематических решениях. Кроме того, в данной работе

также рассматривается мера ‘Average Term Stability’ на основе усредненного расстояния Жаккара. Авторы данной работы предлагают способ выделения стабильных тем с помощью ‘K-Fold ensemble approach’, который тестируются на моделях LDA с сэмплированием Гиббса и NMF (Non-negative Matrix Factorization approach) – модели, близкой к ТМ. В работе Грина и соавторов [40] также была использована мера ‘Average Jaccard (AJ) measure’ для определения оптимального числа тем в англоязычных размеченных датасетах. В работе [41] де Ваал показал, что мера перплексии мало пригодна для оценки стабильности тематической модели, так как она, во-первых, зависит от размера датасета, что затрудняет сравнение различных датасетов между собой, а во-вторых, ведет себя монотонно-убывающим образом.

Указанные работы направленны только на разработку и тестирование меры стабильности. Тем не менее, есть несколько работ, в которых предлагается модификация самой тематической модели с целью увеличения ее стабильности. В работе Кольцова и соавторов [42] показано, что выбор коэффициентов регуляризации в моделях LDA с сэмплированием Гиббса и ARTM существенно влияет на стабильность тематической модели. Кроме того, в данной работе предложен гранулированный вариант процедуры сэмплирования (GLDA), которые дает чрезвычайно высокий уровень стабильности тематической модели. Подробное описание GLDA приведено в главе 5 данной работы. В данной работе было показано, что добавление регуляризаторов влияет стабильность тематической модели.

Наиболее подробный обзор статей, связанных с проблемой стабильности/нестабильности тематических моделей, приведен в работе Агарвала [44]. В целом, проблема нестабильности тематических моделей полностью не решена.

На основании обзора литературы можно сделать следующее заключение. В англоязычной литературе наиболее широкое распространение получили следующие модели: 1. pLSA (E-M алгоритм). 2. LDA с сэмплированием Гиббса. 3. Variation LDA (E-M алгоритм). Данные модели чаще всего используются в качестве базовых (baseline) при сравнении с другими тематическими моделями. В русскоязычной научной литературе большое распространение получила модель ARTM, в которой реализован подход, альтернативный как вариационному принципу вывода тематических моделей, так и выводу на основе физической модели Потса (расширенный вариант модели Изинга), реализованной с помощью сэмплирования Гиббса.

Среди мер качества тематических моделей наибольшее распространение получили следующие: 1. Максимум логарифма правдоподобия (настройка тематических моделей). 2. Дивергенция Кульбака-Лейблера (определение стабильности тематических моделей). 3. Когерентность (определение связности тем в тематических моделях).

В области тематического моделирования выявлены следующие проблемы: 1. Проблема определения оптимального количества тем. В существующих моделях оно задается явным образом, при неясных критериях задания. 2. Проблема оценки величин гиперпараметров, включая коэффициенты регуляризации, от которых тематические модели сильно зависят. Выбор таких параметров может быть частично решен за счет поиска логарифма правдоподобия, однако данный подход работает лишь для нескольких ‘плоских’ тематических моделей. 3. Проблема разработки стабильных тематических моделей, которая усугубляется тем, что стабильность сильно зависит от числа тем и от величин гиперпараметров, выбор которых не ясен. 4. Проблема одновременной оценки тематической модели, как с точки зрения настройки гиперпараметров, так и с точки зрения определения семантической связности тем. Выше указанные проблемы не решены, так как развитие области тематического моделирования пошло, прежде всего, по пути разработки большого числа моделей. Исследования, связанные с анализом настроек моделей или с решением проблемы стабильности, фрагментарны и малочисленны. Исходя из этого, данная диссертационная работа направлена на частичное решение выше указанных проблем.

В рамках данной работы рассматривается следующий список тематических моделей: 1. LDA (с сэмплированием по Гиббсу), 2. pLSA (E-M алгоритм), 3. VLDA (E-M алгоритм), 4. GLDA (с сэмплированием по Гиббсу). 5. ARTM с регуляризаторами разреживания матриц Φ и Θ (E-M алгоритм). 6. hLDA. 7. hPAM. 8. hARTM. Данный выбор обусловлен, во-первых, тем, что эти модели чаще всего используются в научной литературе (особенно как базовые при разработке новых моделей). Во-вторых, они основаны на двух разных принципах (E-M алгоритм и процедура сэмплирования Гибсса). В-третьих, эти модели предназначены для работы с датасетами, имеющими разную тематическую структуру. В данной диссертационной работе использовались разноязычные датасеты, как размеченные, так и нет, с тематической структурой разной глубины иерархии и без нее. Это дает возможно оценить эффективность разработанных моделей для определения различных тематических структур.

2. Энтропийная тематическая модель на основе параметризованных энтропий Реньи и Шарма−Миттала

В данной главе рассматривается теоретическая формулировка энтропийнной тематической модели для одно- и двухпараметрических энтропий, а также приводятся серия компьютерных экспериментов на размеченных и не размеченных разноязычных датасетах, которые показывают полезность параметризованных энтропий для целей настройки тематических моделей и для определения наличия ‘плоской’ или иерархической тематической структуры.

Предлагаемая энтропийнная тематическая модель базируется на идеях работы Rose [12], где было показано, что процедуру кластеризации можно рассматривать в терминах вероятности принадлежности к кластеру, при этом такая вероятность выражается через свободную энергию всей статистической системы (то есть через стат. сумму системы). В такой модели кластеризации температура является параметром настройки кластерной модели, который находится посредством процедуры отжига. В отличии от модели Rose, энтропийная тематическая модель рассматривает температуру в виде количества кластеров, а в качестве целевой функции качества используется параметризованные энтропии. Данное различие позволяет сформулировать энтропийную модель настройки гиперпараметров, включая количество тем на основе поиска минимума параметризованной энтропии. В целом, энтропийная тематическая модель базируется на следующих положениях [45, 46]. 1) Коллекция документов является мезоскопической информационной системой, которая состоит из множества элементов (слов и документов), поэтому поведение такой 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. Соответственно, такая информационная система может не достигать равновесного состояния в смысле максимума энтропии Шеннона, но может находиться в промежуточном равновесном состоянии, которое определяется локальным минимумом параметризованной энтропии Реньи или Тсаллиса. 2) Под темой понимается состояние (аналог направления спина), которое может принимать каждое слово и документ в коллекции. При этом как слово, так и документ принадлежат ко всем темам с различной вероятностью (матрицы этих вероятностей обычно обозначаются как Φ и Θ, соответственно). Множество слов и документов с высокой вероятностью по теме формирует то, что можно назвать тематическим кластером. 3) Информационная система обменивается с внешней средой только энергией за счет изменения температуры. В данном подходе, под температурой информационной системы понимается число тем, задаваемое извне и являющееся параметром, который нужно определить за счет поиска минимума меры несимметричной дивергенции Кульбака-Лейблера (физический аналог – свободная энергия). Поскольку последняя эквивалентна разнице свободных энергий [47], где одна часть свободной энергии отвечает за начальное (равновесное) состояние, а вторая часть

характеризует неравновесное состояние системы [47], то в качестве меры неравновесности такой информационной системы можно использовать следующее выражение Λ_F=F(T) - F_0, где F_0 – свободная энергия начального состояния (хаос) тематической модели, F(T) – свободная энергия при заданном числе тем T, полученная после проведения процедуры кластеризации (тематического моделирования). 4) Минимум Λ_F зависит от различных параметров тематической модели. 5) Оптимальное число тем и набор оптимальных параметров тематической модели соответствует ситуации, когда достигается информационный максимум $S = -I$ [48], то есть минимум величины Λ_F, а также минимум энтропии Реньи, которая может быть выражена через разницу свободных энергий.

В тематических моделях сумма всех вероятностей слов равна числу тем: $T = \sum_{t=1}^{T}\sum_{n=1}^{W} p_{tn}$. В рамках статистической физики принято исследовать распределение статистической системы по уровням энергии, где энергия уровня выражается через вероятность. В соответствии с этим подходом в данной работе диапазон вероятностей делится на фиксированное число интервалов, определяется энергия этих уровней, а также число слов, лежащих на каждом из уровней. Следует отметить, что число слов, лежащих в каждом интервале, зависит от числа тем и величин параметров тематической модели. Разделение на интервалы является условным и удобным с вычислительной точки зрения. При стремлении такого интервала к нулю, распределение числа слов по интервалам будет стремиться к функции плотности распределения ρ. Однако, для простоты изложения, в рамках данной работы мы будем рассматривать двухуровневую систему, в которой на одном уровне будут находиться слова с высокой вероятностью, а на другом уровне будут находиться слова с низкой вероятностью, то есть, с вероятностью, близкой к нулю.

2.1. Энтропийная тематическая модель на основе энтропии Реньи

Введем функцию плотности распределения слов для уровня слов с высокой вероятностью при заданном количестве тем и фиксированном наборе параметров следующим образом [46]:

$$\check{\rho} = \frac{\sum_{t=1}^{T}\sum_{n=1}^{W} N_{tn}}{WT} \qquad (1),$$

N_{tn} – число слов с высокой вероятностью, T – число тем, n – суммирование по списку уникальных слов, t – суммирование по всем темам. Под высокой вероятностью мы понимаем вероятности, чьи величины $p_{tn} > 1/W$, где W – число уникальных слов в датасете. Выбор данного уровня обусловлен тем, что величина $1/W$ является исходной величиной при инициализации матрицы Φ. Величина WT определяет общее число всех микросостояний в тематической модели (под микросостоянием понимается вероятность одного слова в одной теме), то есть, размер матрицы Φ является нормировкой функции плотности распределения. В ходе тематического моделирования вероятности слов перераспределяются относительно данного порога. Небольшая часть слов попадает на уровень с высокой вероятностью $p_{tn} > 1/W$, а большая часть слов попадает на другой уровень, где $p_{tn} > 1/W$. Уровень слов с высокой вероятностью в тематической модели можно характеризовать величиной энергии, которая выражается через сумму вероятностей слов, находящихся на данном уровне, нормированных на общее число тем:

$$E = -T \cdot \ln \check{P} \qquad (2),$$

где $\check{P} = \sum_{t=1}^{T}\sum_{n=1}^{W} p_{tn}\,/T$, суммирование производится по всем словам с высокой вероятностью, находящимся на данном уровне, T – число тем. Таким образом, уровень определяется двумя экспериментально измеряемыми величинами: 1. Суммой вероятностей слов на данном уровне $\check{P}$. 2. Числом слов, лежащим на данном уровне (плотностью распределения слов $\check{\rho}$).

Для двухуровневой системы основной вклад в энтропию и энергию всей системы дают именно слова с высокой вероятностью, поэтому свободная энергия всей системы приблизительно определяется через энтропию и энергию одного уровня. Свободная энергия такой системы выражается через энтропию Гиббса (энтропию Шеннона) и энергию

следующим образом [47]: $F = E - T \cdot \mathrm{S} = \mathrm{E} - \mathrm{S}/$, где $q = 1/T$. Энтропия информационной системы (энтропия Шеннона) выражается через число слов на одном уровне по следующей формуле: $S = \ln(\check{\rho}(T))$ [45]. Разница свободных энергий системы выражается через величины $\check{P}$ и $\check{\rho}$ следующим образом:

$$\Lambda_F = F(T) - F_0 = (E(T) - E_0) - (S(T) - S_0) \cdot T = -\ln\left(\check{P}\right) - T \cdot \ln(\check{\rho}) \quad (3),$$

где E_0, S_0 – энергия и энтропия системы при начальном распределении, которые соответствуют максимуму энтропии, то есть $S_0 = \ln(T)$ и $E_0 = -\ln(W \cdot T)$. Таким образом, уровень неравновесности тематической модели определяется как разность свободных энергий и выражается через экспериментально определяемые величины $\check{\rho}$ и $\check{P}$. При этом нормировка данных величин по своей сути является энтропией начального состояния, то есть хаоса. Величины $\check{\rho}$ и $\check{P}$ подсчитываются для каждой тематической модели при вариации свободного параметра T и параметров тематической модели; таким образом, величина Λ_F является функцией от числа тем T, размера словаря W, то есть датасета, а также зависит от величин параметров генеративной тематической модели.

2.2. Связь свободной энергии с энтропией Реньи в тематических моделях

За счет использования статистической суммы следующего вида: $\mathbf{Z_q} = \sum \check{\boldsymbol{\rho}} \cdot \check{\mathbf{P}} = \sum \check{\boldsymbol{\rho}} \cdot \mathbf{e}^{-\mathbf{q}\cdot\mathbf{E}} = \sum \mathbf{e}^{-\mathbf{q}\cdot\boldsymbol{\Lambda}_\mathbf{F}}$, q=1/T [49] можно выразить свободную энергию тематической модели через энтропию Реньи и через экспериментально определяемые величины $\check{\boldsymbol{P}}$ и $\check{\boldsymbol{\rho}}$:

$$S_q^R = \frac{\ln(Z_q)}{q-1} = \frac{\ln(e^{-q\cdot F})}{q-1} = \frac{-\mathrm{q}\cdot\Lambda_F}{q-1} = \frac{\Lambda_F}{T-1} \qquad (4).$$

Необходимо отметить, что связь между свободной энергией и энтропией Реньи можно также осуществить при помощи эскорт-распределения [50, 51], так как задание выше указанной статистической суммы эквивалентна эскорт преобразованию.

Таким образом, энтропия Реньи в тематических моделях выражается через свободную энергию, параметр q, равный обратному числу тем $q = 1/T$, и через экспериментально определяемые величины $\check{P}$ и $\check{\rho}$. При таком подходе энтропия Реньи, во-первых, характеризует меру неравновесности тематической модели, так как ее вычисление основано на разнице свободных энергий. Во-вторых, оптимизация моделей машинного обучения может быть реализована на основе поиска минимума энтропии Реньи. В-третьих, энтропия Реньи в своей формулировке, в отличии от энтропии Шеннона, включает в себя два разнонаправленных процесса, а именно увеличение числа тем, которое, с одной стороны, приводит к уменьшению энтропии Шеннона, а с другой стороны, к увеличению общей энергии, и, следовательно, к увеличению общей суммы вероятностей в модели. Таким образом, разность между двумя разнонаправленными процессами имеет область баланса, где два процесса уравновешивают друг друга. В этой области энтропия Реньи является минимальной. При этом, минимум энтропии соответствует максимуму информации в тематической модели. Следовательно, настройка параметров тематической модели может осуществляться на основе поиска минимума однопараметрической энтропии Реньи.

2.3. Энтропийная модель на основе энтропии Шарма−Миттала

Тематическая модель на основе энтропии Реньи не включает в себя семантическую составляющую, которая играет важную роль при практическом применении моделей кластеризации на текстовых датасетах. Однако, энтропийную тематическую модель можно расширить за счет использования двухпараметрической энтропии Шарма−Миттала [52, 53]. Она имеет следующий вид: $S_{S,M} = \frac{1}{1-r}\left[\left(\sum_i p_i^q\right)^{\frac{1-r}{1-q}} - 1\right]$, где r, q – параметры, определяющие характер параметризации энтропии.

Энтропия Шарма−Миттала включает в себя энтропию Реньи и Тсаллиса как частные случаи задания параметров r, q. Например, при $\mathrm{r} \to 1$ энтропия $\mathrm{S_{S,M}}$ совпадает с энтропией

Реньи , а при $r \to q$ $S_{S,M}$ совпадает с энтропией Тсаллиса. Следует отметить, что предел энтропии Шарма–Миттала при $r \to 0$ равен экспоненте от энтропии Реньи без единицы, что можно рассмотреть как параметризованную перплексию. В этом случае $\lim_{r\to 0} S_{S,M} = e^{S_q^R} - 1$. Покажем, что $e^{S_q^R} - 1 > S_q^R$ при условии, что $S_q^R \neq 0$. Рассмотрим $f(x) = e^x - 1 - x$ при $x \neq 0$. Получаем, что $f'(x) = e^x - 1$. Таким образом, f возрастает при $x > 0$ и убывает при $x < 0$, и, соответственно, $\min f(x) = f(0) = 0$. Например, при $S_q^R = 6$, $e^{S_q^R} - 1 \cong 402$; при $S_q^R = 1$, $e^{S_q^R} - 1 \cong 1.7$; при $S_q^R = 0.1$, $e^{S_q^R} - 1 \cong 1.005$, то есть, флуктуация параметра r приводит к очень большим значениям энтропии.

Исходя из того, что в энтропийной тематической модели параметр $q = 1/T$ связан с числом тем, использование энтропии Шарма–Миттала для анализа тематических моделей требует доопределения смысла параметра r. Он, прежде всего, меняется в диапазоне [0-1]. Более того, если $r = 1$, то энтропия Шарма–Миттала преобразовывается в энтропию Реньи, и, следовательно, качество тематической модели определяется исключительно энтропией Реньи и параметром q. На основании этого, можно заключить, что параметр q для энтропии Шарма–Миттала представляет собой обратную величину числа тем. В случае, когда $r = 0$, энтропия Шарма–Миттала выглядит следующим образом: $S_{S,M} = e^{S_q^R} - 1$, то есть становится очень большой. Исходя из принципа, что максимум энтропии соответствует минимуму информации, мы можем заключить, что минимальные значения параметра r, которые ведут к максимальным значениям $S_{S,M}$, соответствуют минимальным значениям информации.

В научной литературе используется понятие расстояние Жаккара, определяемого следующим образом [54]: $J(X,Y) = 1 - \frac{X \cap U}{X \cup Y}$. Коэффициент Жаккара измеряет сходство между двумя множествами (в данном случае между двумя наборами слов), и определяется как размер пересечения множества слов, деленный на размер объединения множества слов. В случае идентичности двух множеств данное расстояние равно нулю. Расстояние Жаккара играет важную роль, в особенности в области компьютерных наук, при исследовании регулярных языков [55], и связано с ‘entropy distance’ следующим образом:

$$D_H(X,Y) = 1 - \frac{I(X,Y)}{H(X,Y)} = J(X,Y) = 1 - J,$$

где $I(X,Y)$ - взаимная информация X и Y, а $H(X,Y)$ - совместная энтропия X и Y. В теории информации взаимная информация соответствует пересечению множеств X и Y, а совместная энтропия – объединению X и Y, и, следовательно, энтропийное расстояние соответствует расстоянию Жаккара. Если $J(X,Y) = 0$, то и $D_H(X,Y) = 0$. Таким образом, на основании выше сказанного, мы можем доопределить суть параметра r следующим образом. Параметр r в $S_{S,M}$ энтропии будет отвечать за семантический состав тематической модели, то есть будет измеряться с помощью расстояния Жаккара. Данный параметр характеризует величину изменения семантического состава при изменении числа тем (а также при вариации величин гипрепарметров тематической модели). Это связано с тем, что вариация гиперпараметров модели и числа тем влияет на состав высоко вероятностных слов в тематической модели.

Таким образом, настройка энтропийной тематической модели осуществляется за счет подбора числа тем (параметр $q = 1/T$) и гиперпараметров модели, при условии достижения минимума двухпараметрической энтропии Шарма–Миттала. То есть среди множества параметров необходимо выбирать те параметры, которые соответствуют максимуму информации тематической модели для выбранного датасета.

2.4. Энтропийная тематическая модель на основе энтропии Шарма−Миттала

На основании формулы 4 и статистической суммы $Z_q = \sum \check{\rho} \cdot e^{-q \cdot E}$, энтропия Шарма−Миттала тематической модели в терминах экспериментально определяемых величин $\check{P}$ и $\check{\rho}$ выглядит следующим образом [56]:

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

$$= \frac{1}{1-r}\left[\left(\left(\frac{P(T)}{T}\right)^{q} \cdot \left(\frac{N_{tn}}{WT}\right)\right)^{\frac{1-r}{q-1}} - 1\right], \quad (5),$$

где, W – число слов в словаре, T – число тем, $P(T)$ – сумма вероятностей слов на втором уровне, N_{tn} – число слов с высокой вероятностью, то есть число слов на втором уровне, n – суммирование по списку уникальных слов, t – суммирование по всем темам. Следовательно, формула (5) позволяет вычислять двухпараметрическую энтропию тематической модели на основании экспериментально наблюдаемых величин: ‘нормированной суммы вероятностей слов на данном уровне – $\check{P}$’ и ‘нормированной плотности распределения слов – $\check{\rho}$’. Таким образом, $S_{S,M}$, с одной стороны, позволяет оценить параметры тематических моделей, например, такие как параметры регуляризации в моделях LDA Gibbs sampling и ARTM, а также число тем на основе поиска минимума $S_{S,M}$, которая, в свою очередь, характеризуется разницей энтропий между начальным распределением и распределением, полученным в ходе моделирования. С другой стороны, она позволяет оценить, какой вклад в энтропию вносит расстояние Жаккара между двумя разными тематическими решениями, которые характеризуются разными значениями параметров и числом тем. Соответственно, наилучшие значения параметров тематических моделей соответствуют ситуации, когда энтропия имеет минимальное значение, а наихудшие значения – максимальной величине энтропии.

2.5. Иерархическая энтропийная тематическая модель

Текстовые коллекции могут содержать как обычную тематическую структуру (plane topic structure), так и иерархическую структуру. На данный момент не существует методов хорошего определения структуры, за исключением энтропийной модели, предлагаемой в данной работе [2]. Общая идея определения структуры заключается в следующем. Как было показано ранее [46], в датасете может присутствовать несколько локальным минимумов параметризованной энтропии, которые соответствуют разному количеству тем. Соответственно, эти минимумы можно сопоставить с разными уровнями иерархии. Таким образом, маркером той или иной тематической структуры может служить количество минимумов. Если в датасете присутствует только один минимум, то в данном датасете присутствует только один уровень тем; если в датасете есть два минимума, то можно говорить о двух уровнях тематической иерархии. Исходя из этого, рассмотренную энтропийную тематическую модель также следует расширить на иерархические модели следующим образом [2]. Так как иерархическая структура в ТМ может быть изображена в виде графа, где каждый узел представляет собой одну тему, процедура иерархического ТМ приводит к построению иерархического дерева, где на каждом уровне присутствует фиксированное число тем. У каждого узла-темы есть список слов и документов с вероятностями, которые определяют величину их принадлежности к данной теме. Общее число слов на каждом уровне является константой, которая равна общему числу элементов W в статистической системе. Совокупность узлов-тем на одном уровне представляет собой матрицу Ф (распределение слов по темам).

Процедура иерархического тематического моделирования заключается в построении последовательности матриц Φ, в которых число слов постоянно, а число тем

последовательно увеличивается (от уровня к уровню иерархии). Соответственно, при иерархическом тематическом моделировании, по мере перехода от уровня к уровню, происходит изменение доли слов с вероятностями выше величины $1/W$. Таким образом, каждый уровень иерархии характеризуется следующими параметрами: 1. Количество тем T_i на -ом уровне. 2. Количество слов на i уровне $N_i = \sum_t N_{it} \left(\phi_{it} > \frac{1}{W}\right)$, вероятность которых выше порога $1/W$, где W – размер словаря, t – индекс суммирования по темам. 3. Сумма вероятностей слов $\tilde{P} = \sum_{t=1}^{T_i} \phi_{it} \left(\phi_{ti} > \frac{1}{W}\right)$. На основании этих величин можно определить внутреннею энергию и энтропию Шеннона (S) текущего уровня по отношению к равновесному состоянию этого же уровня: $E_i = -ln(\tilde{P}/T_i)$, $S_i = ln(\frac{N_i}{WT_i})$, где i – номер уровня. Далее, при помощи S_i и E_i можно определить свободную энергию и энтропию Реньи i-го уровня иерархии. Свободная энергия иерархического уровня выражается следующим образом: $\Lambda_{F_i} = E_i - T_i \cdot S_i$. Энтропия Реньи i-го уровня выражается через свободную энергию i -го уровня следующим образом: $S_i^R = \frac{\Lambda_{F_i}}{T-1}$, где $q = 1/T_i$ – параметр, характеризующий каждый уровень иерархии.

Таким образом, измеряя величину энтропии на каждом уровне иерархии, при вариации параметров модели (включая число уровней) для заданного 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 параметров. При этом процесс кластеризации слов по темам начинается с максимума энтропии, когда все элементы (слова) статистической системы относятся к одной или двум темам, и закачивается также максимумом энтропии, где все элементы приблизительно с одинаковой вероятностью принадлежат всем темам (для большого числа тем). Расположение глобального минимума и ряда локальных минимумов энтропии Реньи в терминах числа тем определяется особенностями датасета. Энтропия Реньи S_i^R выступает как мера неравновесности системы, где минимум энтропии соответствует максимуму информации, а количество минимумов энтропии Реньи служит маркером тематической структуры.

Необходимо отметить, что данный принцип был использован для настройки процедуры иерархической кластеризации (на основе ‘complete’ method) [57] при кластеризации пользователей социальной сети VK.

2.6. Экспериментальная проверка применения энтропии Реньи и Тсаллиса в тематических моделях

В данной работе проводилось исследование четырех тематических моделей с точки зрения поведения энтропий Реньи и Тсаллиса как функций от числа тем: 1. LDA GB. 2. Granulated LDA (GLDA GB). 3. PLSA (E-M algorithm). 4. Variational LDA (E-M algorithm). Выбор этих моделей обусловлен тем, что, во-первых, эти модели являются ‘base line’ для большого числа работ в области тематических моделей. Во-вторых, эти модели представляют собой основные типы ‘inference’ тематических моделей. В каждом эксперименте для каждой модели измерялось количество микросостояний, чьи вероятности больше заданной величины $p_{tn} > \frac{1}{W}$. Далее рассчитывалась функция плотности состояний от числа тем, внутренняя энергия, энтропия и свободная энергия для каждой модели. На основании свободной энергии рассчитывались энтропии Реньи и Тсаллиса для каждого тематического решения.

Датасеты: 1. Датасет ‘Live Journal’. Набор русскоязычных постов из социальной сети ‘Live Journal’, размера: 101481 постов; размер словаря: 172939 уникальных слов. Число тем варьировалось в диапазоне: T = [2; 330] с шагом в 2 темы. 2. Англоязычный датасет ‘20

newsgroup' [58]. Размер – 15404 поста и N=50948 уникальных слов, размеченных на 20 тем. Число тем для второго датасета варьировалось в диапазоне: T = [2; 120] с шагом в 2 темы. Выбор этих датасетов обусловлен следующими причинами. Во-первых, это датасеты на разных языках, что позволяет показать кросс-языковую применимость энтропийных тематических моделей и установить общие черты моделей, присущие разным языкам. Во-вторых, разный размер коллекций показывает, что изменение размера может приводить к появлению дополнительных локальных минимумов. Кроме того, на англоязычной коллекции тестировались различные модели кластеризации [59], что дает возможность сравнить результаты тематического моделирования с результатами кластерного анализа.

На рисунках (1), (2) приведены энтропии Шеннона, Реньи для четырех моделей (датасет '20 newsgroup'). Каждая модель запускалась по три раза, затем результаты расчета усреднялись. Энтропии рассчитывались на основе усредненных значений. Усреднение результатов моделирования связано с учетом нестабильности тематических моделей.

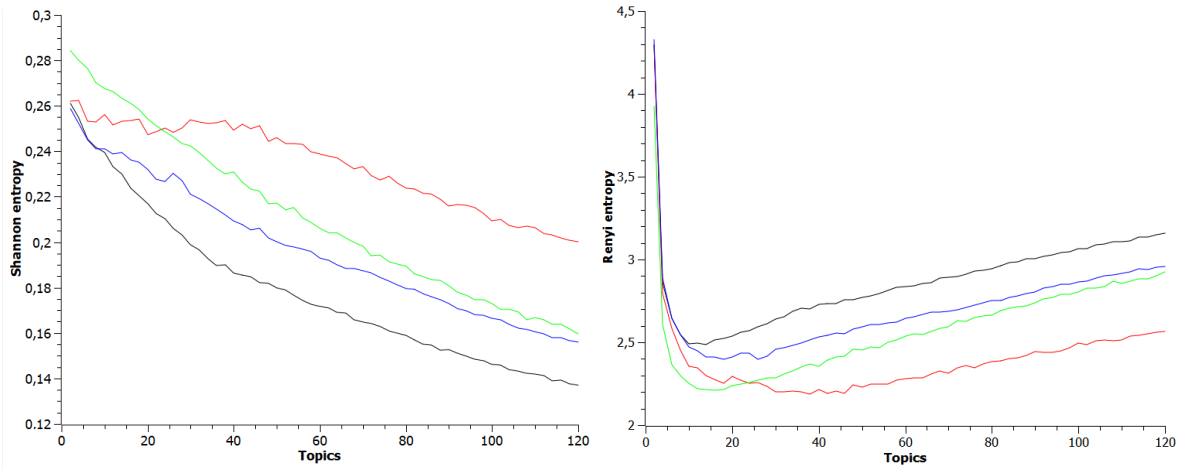

Рис. 1. Распределения энтропии Шеннона. Рис. 2. Распределения энтропии Реньи.

Распределения энтропии Шеннона и Реньи как функция от числа тем для '20 newsgroup dataset'. LDA (Gibbs sampling) – черный цвет, GLDA (Gibbs sampling) – красный цвет, PLSA (E-M algorithm) – синий цвет, LDA (E-M algorithm) – зеленый цвет.

Энтропия Реньи, в отличие от энтропии Шеннона, имеет глобальный минимум и показывает правильные результаты на граничных значениях числа тем. При $T \to 1$ энтропия Реньи дает максимум, так как тематическое моделирование, равно как любой другой кластерный алгоритм, не дает распределение кластеров, то есть информация близка нулю. В то же время, увеличение числа кластеров/тем (то есть $T \to \infty$) приводит к равномерному распределению каждого слова по темам, что соответствует увеличению энтропии. Однако разные модели дают немного разное положение минимума энтропии Реньи и разную глубину этого минимума. Чтобы определить, какая из указанных моделей дает более точный результат, необходимо сравнить результаты ТМ с результатами кластерного анализа на этой же коллекции. Авторы работы [59] тестировали ряд алгоритмов кластеризации на '20 newsgroup dataset' и показали, что оптимальное количество кластеров варьируется в диапазоне от 15 до 20 кластеров для разных алгоритмов за счет корреляции некоторых тем.

Модели LDA (Gibbs sampling, LDA GB) и LDA (E-M algorithm) показывают, что оптимальное число тем – порядка 15, модель PLSA (E-M algorithm) дает 20 тем. Однако, самое существенное отличие дает модель GLDA, которая почти вдвое завышает число тем

по сравнению с моделями LDA (Gibbs sampling) и LDA (E-M algorithm). Это связано с тем, что в модели GLDA в процедуре сэмплирования присутствует сильное усреднение, что приводит к высокой величине стабильности, но при этом сдвигает глобальный минимум энтропии Реньи.

Результаты расчета энтропии Реньи и Тсаллиса для 4 тематических моделей на русскоязычном датасете приведены на рисунках (3), (4).

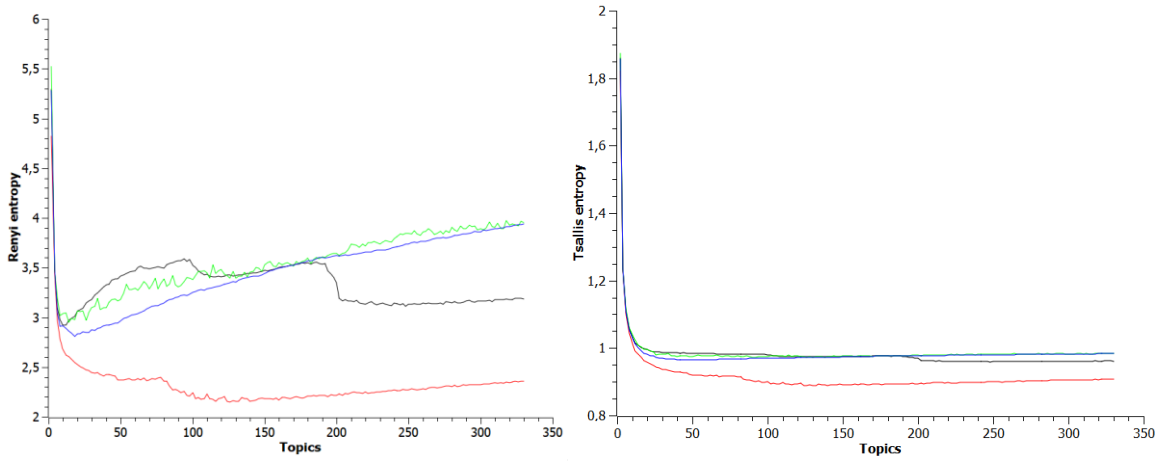

Рис. 3. Распределение энтропии Реньи. Рис. 4. Распределение энтропии Тсаллиса.

Распределение энтропии Реньи и Тсаллиса как функции от числа тем для датасета из ‘Live Journal’: LDA (Gibbs sampling) – черный цвет, GLDA (Gibbs sampling) – красный цвет, PLSA (E-M algorithm) – синий цвет, LDA (E-M algorithm) – зеленый цвет

Вычисления показывают, что модели на основе E-M алгоритма показывают сильное отличие от моделей на основе сэмплирования Гиббса для русскоязычного датасета, причем именно при большом числе тем (свыше 100). Модель LDA GB демонстрирует наличие сильных скачков энтропии Реньи, связанных со значимыми флуктуациями функции плотности распределения. Однако модели LDA (E-M algorithm) и PLSA (E-M algorithm) не видят этих скачков. Флуктуации плотности распределения в моделях сэмплирования Гиббса не могут быть объяснены особенностями процедуры сэмплирования, так как в работе [45] проводились исследования на этом же датасете, в котором модель LDA (Gibbs sampling) запускалась по три раза для каждой темы, а темы варьировались с шагом 1, в диапазоне [105 - 120], и с шагом 10 в диапазоне [120-600]. Скачок в области [110 – 120] тем наблюдался во всех запусках модели. Таким образом, модели на основе сэмплирования Гиббса обладают большей чувствительностью по сравнению с другими моделями. Энтропия Тсаллиса, рассчитанная на модели LDA(Gibbs sampling), также показывает скачок в области [110-120] тем и в области [190-200] тем, однако амплитуда скачка существенно ниже. Это обусловленно тем, что энтропия Тсаллиса является более стабильной с точки зрения Lesche [60].

На основании проведенных вычислений можно сделать следующее заключение. Во-первых, параметризованная энтропия Реньи пригодна для определения количества тем в текстовых датасетах, так как ее минимум совпадает с результатами человеческой разметки. Параметрами энтропии является число тем. Во-вторых, разные тематические модели демонстрируют разное количество минимумов параметризованной энтропии, но при этом положение глобального минимума для разных моделей практически одинаково. Положение минимумов и их количество характеризуются особенностями датасетов.

2.7. Численные эксперименты по семантической стабильности в тематических моделях

При описание различных физических статистических систем важным фактором является неразличимость частиц. Это позволяет использовать комбинаторный подход для подсчета числа состояний и оценки функции плотности распределения. В этом случае не важно, какие именно частицы заселяют состояния с высокой вероятностью. Однако в случае информационных систем, состоящих из множества документов, тема формируется из множества различных слов, семантические различия между которыми важны. Поэтому при исследовании поведения текстовых систем необходимо проверять, насколько воспроизводимо распределение слов с семантической точки зрения, при изменении параметра ‘число тем’ и гипер-параметров. В данной работе, семантическая воспроизводимость в ТМ двух облаков слов T_1 и T_2 (соответствующие двум разными темам). измерялось при помощи расстояния Жаккара.

Расстояние Жаккара вычислялись попарным сравнением каждого тематического решения со всеми остальными решениями и сохранялись в матрицу, где ячейка содержит величина расстояния Жаккара $J_{t1,t2}$, где $t1, t2$ – номера тем.

На рисунках (5) и (6) приведены кривые диагональных расстояний Жаккара по моделям LDA (Gibbs sampling) и LDA (E-M algorithm) для русскоязычного датасета. Мы не приводим величины расстояния Жаккара для англоязычного датасета, так как все модели показали приблизительно одинаковые величины порядка 0.99.

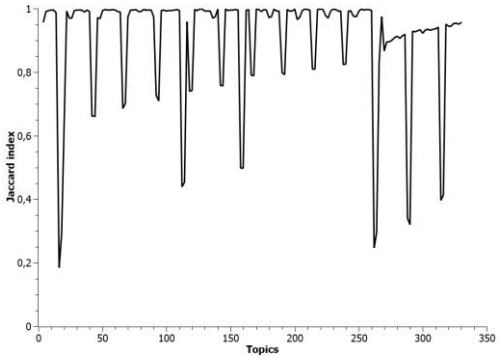

Рис. 5. Поведение расстояния Жаккара модели LDA Gibbs sampling.

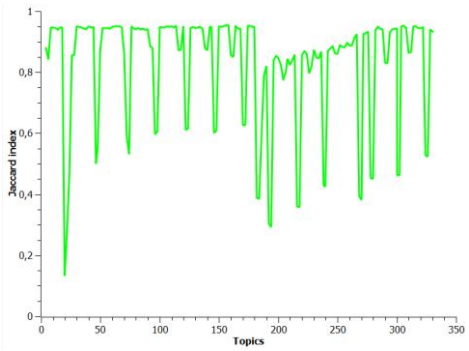

Рис. 6. Поведение расстояния Жаккара модели LDA (E-M алгоритм)

Распределение расстояний Жаккара показывает, что модели обоих типов демонстрируют существование зон семантической стабильности. При этом, существуют зоны с высоким уровнем коэффициента $J_{t1,t2} \cong 0.9$ и зона более низким уровнем $J_{t1,t2} \cong 0.5$. Однако, если использовать достаточно большое количество топовых слов, например, 1000 в каждом тематическом решении, то такая периодическая структура практически исчезает.

2.8. Экспериментальная проверка энтропии Шарма−Миттала и энтропии Реньи как меры качества для оценки числа тем и семантической связности тематических моделей

2.8.1. Эксперименты по применению энтропии Реньи

В данной части работы исследовалась возможность определения оптимальных величин параметров в тематических моделях. Исследование проводилось для моделей: 1. LDA Gibbs sampling (LDA GB) [20], 2. pLSA (E-M) [21], 3. ARTM с регуляризаторами ‘sparse Φ’ и ‘sparse Θ’ [16]. В LDA GB гиперпараметрами является величины α, β, характеризующие распределения Дерихле, и T – число тем. В модели ARTM параметрами являются коэффициенты регуляризаторов разреживания матриц Φ, Θ, число тем. Модель

PLSA имеет только один параметр – число тем, поэтому даная модель сравнивалась двумя другими моделями.

Датасеты: 1. ‘20 newsgroup dataset’ (человеческая разметка, диапазон [15 - 20] тем). 2. Датасет на русском языке (‘lenta_ru’) (пользовательская разметка на 10 тем). Анализ корреляции тем показывает, что «реальное» число тем находится в диапазоне [7-10]. Размер датасета 82852 документов, размер словаря 172939. Все датасеты просчитывались для каждой модели, при вариации параметров. Далее для каждого из полученных тематических решений проводился подсчет функции плотности распределения и распределения величин расстояний Жаккара, энтропии Реньи и логарифмп правдоподобия. Кроме того, рассчитывалась двухпараметрическая энтропия Шарма−Миттала ($S_{S,M}$) с целью оценки семантической стабильности тематических решений при вариации гиперпараметров моделей, включая количество тем.

2.8.1.1. Модели pLSA и LDA GB – энтропия Реньи

Кривые энтропии Реньи для модели pLSA, LDA GB для двух датасетов приведены на рисунках (7), (8).

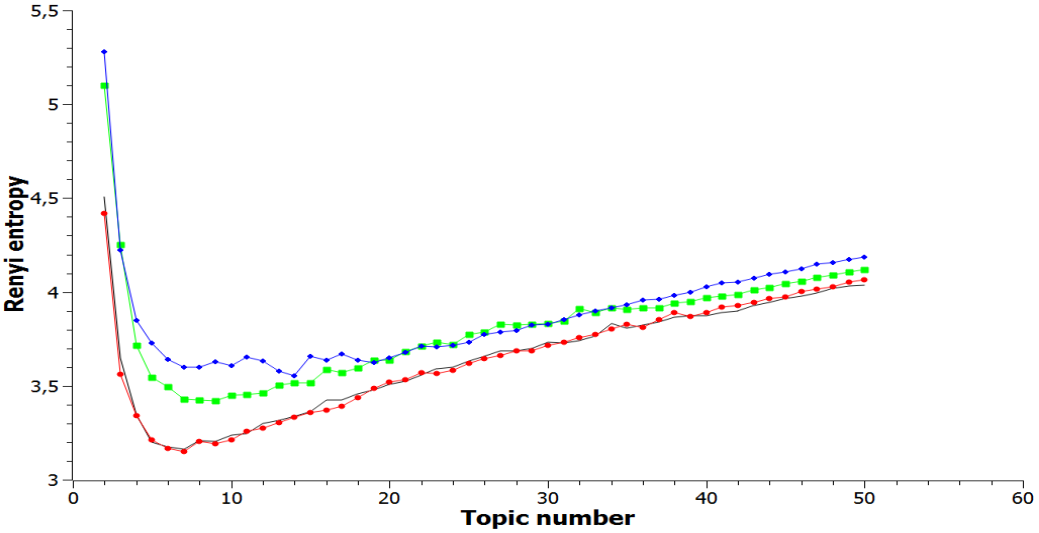

Рис. 7. Энтропия Реньи (‘lenta_ru’). pLSA,– черная линия, LDA GB (α=0.1, β=0.1) – красная линия, LDA GB (α=0.5, β=0.1) – зеленая линия, LDA GB (α=1, β=1) – синяя линия.

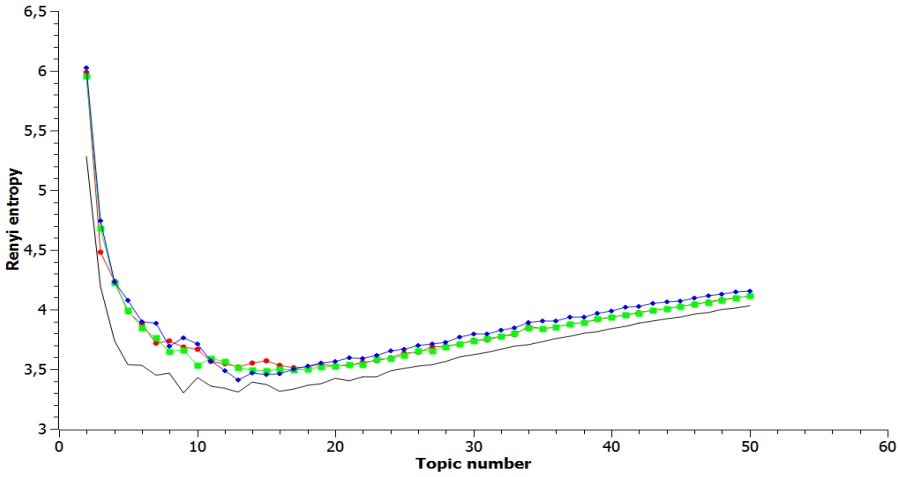

Рис. 8. Энтропия Реньи (’20 topics news’). pLSA, – черная линия, LDA GB (α=0.1, β=0.1) – красная линия, LDA GB (α=0.5, β=0.1) – зеленая линия, LDA GB (α=1, β=1) – синяя линия.

Рисунки (7), (8) показывают, что энтропии Реньи модели pLSA и модели LDA Gibbs sampling с параметрами α=0.1, β=0.1 очень близки к друг к другу. Увеличение параметров регуляризации α, β приводит к увеличению энтропии Реньи, при этом также происходит сдвиг минимума параметризованной энтропии. На рисунке (9) приведены кривые логарифма правдоподобия как функция от числа тем. Видно, что увеличение значений α, β приводит к ухудшению логарифма правдоподобия, что эквивалентно увеличению энтропии. Таким образом, сравнивая поведение кривых логарифма правдоподобия с кривыми энтропии Реньи, можно заключить следующее: 1. Энтропия Реньи пригодна для настройки параметров тематической модели, а минимальное значение энтропии Реньи соответствуют оптимальным значениям параметров рассмотренных тематических моделей. 2. Энтропия Реньи позволяет определить оптимальное число тем, в отличие от логарифма правдоподобия, на основе ее локального минимума.

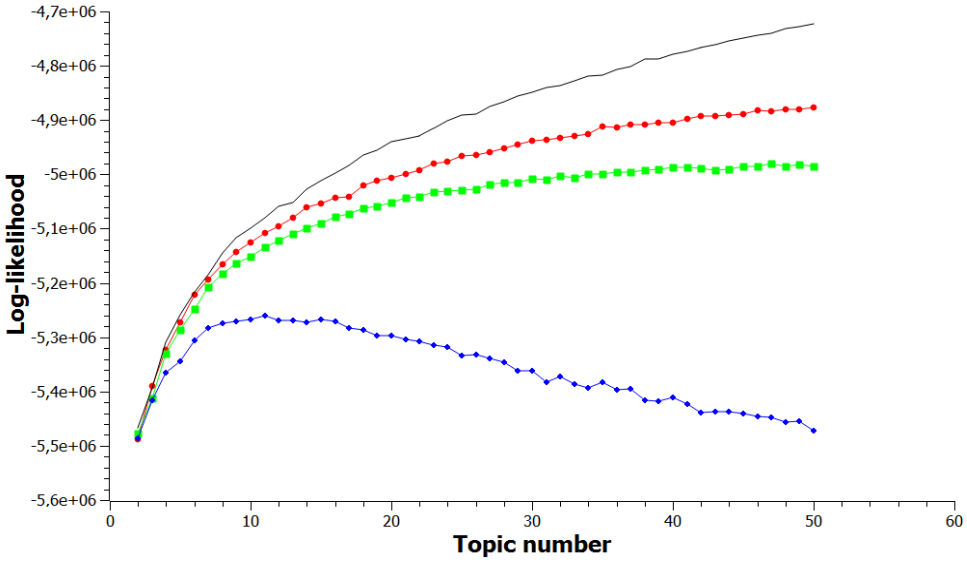

Рис. 8. Логарифм правдоподобия (‘lenta_ru’). pLSA,– черная линия, LDA GB (α=0.1, β=0.1) – красная линия, LDA GB (α=0.5, β=0.1) – зеленая линия, LDA GB (α=1, β=1) – синяя линия.

2.8.1.2. Модель ARTM с разреживанием матрицы Ф – энтропия Реньи

Результат тематического моделирования на основе модели ARTM существенно зависит от величин коэффициентов регуляризации [16]. Их увеличение может приводить к резкому изменению уровня стабильности тематической модели [43]. Исходя из этого, в данной части работы анализируется эффект влияния параметра регуляризации Ф (τ_Φ) и числа тем на поведение энтропии Реньи, при вариации гиперпараметров модели ARTM. При исследовании данной модели производилось вариация числа тем в диапазоне [2-50] и величины τ_Φ в диапазоне [-10, 10]. Набор кривых энтропии Реньи в виде функции от числа тем приведен на рисунке (9). Как видно, увеличение параметра τ_Φ приводит к тому, что минимум энтропии Реньи смещается в область малого числа тем (порядка 2), что намного меньше «реального» (7-10). Таким образом, сильная регуляризация типа ‘разреживание’ матрицы Ф приводит к неправильному количеству тем. Отметим, что изменение знака у коэффициента регуляризции не влияет на результаты моделирования. Кривая энтропии Реньи данной модели для англоязычного датасета ’20 topics news’ приведена на рисунке (10). Увеличение параметра регуляризации приводит к существенному смещению минимума энтропии Реньи. Тогда как в датасете 14-17 тем, при τ_Φ = 1 минимум смещается к 10 темам, а дальнейшее увеличение τ_Φ приводит к поломке модели. Таким образом, наилучший результат тематической модели соответствует минимальной величине

коэффициента регуляризации, а кривая энтропия Реньи практически совпадает с аналогичной кривой для модели pLSA.

2.8.1.3. Модель ARTM с разреживанием матрицы Θ – энтропия Реньи

В данной модели параметрами являются коэффициент регуляризации матрицы Θ (τ_Θ) и число тем. В отличие от предыдущей модели, в данном случае производится разреживание матрицы распределения тем в документах. В расчетах производилось вариация числа тем в диапазоне [2-50] и коэффициента τ_Θ в диапазоне [-10, 10]. Набор кривых энтропии Реньи для данного регуляризатора в виде функции от числа тем приведен на рисунке (11) (датасет ‘lenta’). Кривые энтропии Реньи для коэффициентов регуляризации τ_Θ = [0.01, 0.1, 1] практически не различаются между собой. Однако коэффициент τ_Θ =10 не позволяет рассчитать свободную энергию и энтропию Реньи, так как модель ломается, аналогично предыдущей. Аналогичный результат получается для датасета ’20 topics news’. Таким образом, лучший результат получается при малой величине коэффициента регуляризации, так как его увеличение приводит как к существенному уменьшению логарифма правдоподобия, так и к увеличению энтропии Реньи.

2.8.2. Эксперименты по применению меры качества на основе энтропии Шарма−Миттала к тематическим моделям

В рамках данного набора экспериментов проводилось исследование поведения двухпараметрической энтропии Шарма−Миттала, при вариации гиперпараметров в моделях pLSA, ARTM и LDA Gibbs sampling. Использование данного типа параметризованной энтропии дает возможность оценить изменение семантической составляющей тематических моделей на уровень энтропии при вариации гиперпараметров. Выбранные тематические модели являются одними из наиболее часто используемыми моделями в англоязычной и русскоязычной научной литературе.

2.8.2.1. Модель plsa – энтропия Шарма−Миттала

Для расчета двухпараметрической энтропии $S_{S,M}$, прежде всего, были рассчитаны попарные величины расстояний Жаккара при вариации числа тем. Примеры таких расчетов визуализированы в виде тепловых карт на рисунках (12), (13). Поведение кривых энтропии $S_{S,M}$ для модели plsa (для двух датасетов) приведены на рисунке (14). Большие скачки энтропии Шарма−Миттала обусловлены малыми величинами величин расстояний Жаккара.

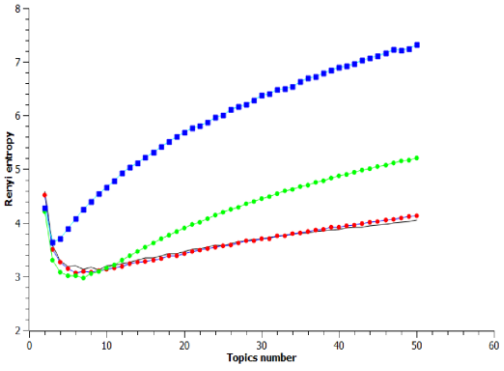

Рис. 9 (датасет ‘lenta’).

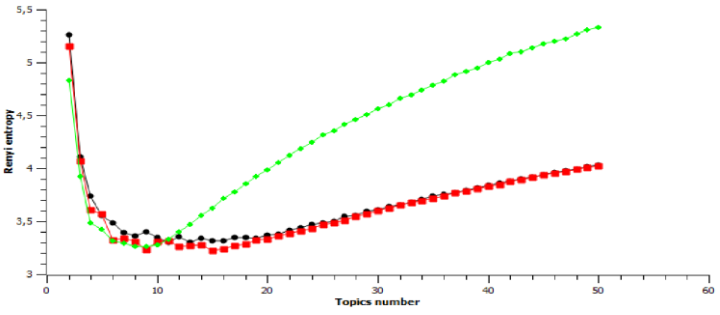

Рис. 10 (датасет ’20 topics news’).

Кривые энтропии Реньи (датасеты ‘lenta’, датасет, ’20 topics news’) при вариации регуляризатора τ_Φ ‘sparse Φ’ (ARTM). Черный цвет: τ_Φ =0.01, красный цвет: τ_Φ =0.1, зеленый цвет: τ_Φ =1, синий цвет: τ_Φ =10

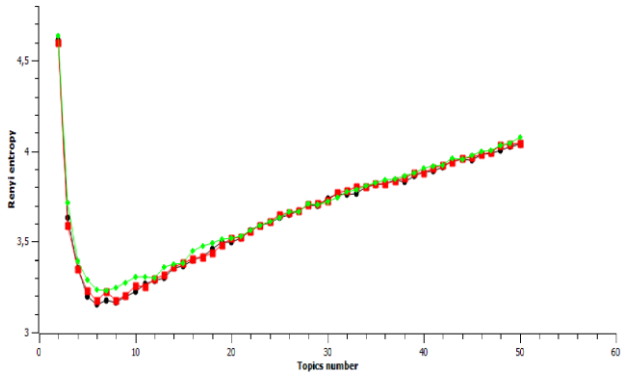

Рис. 11. Кривые энтропии Реньи (датасет ‘lenta’) при вариации регуляризатора τ ‘sparse Θ’ (ARTM). Черный цвет: τ_θ =0.01, красный цвет: τ_θ =0.1, зеленый цвет: τ_θ =1.

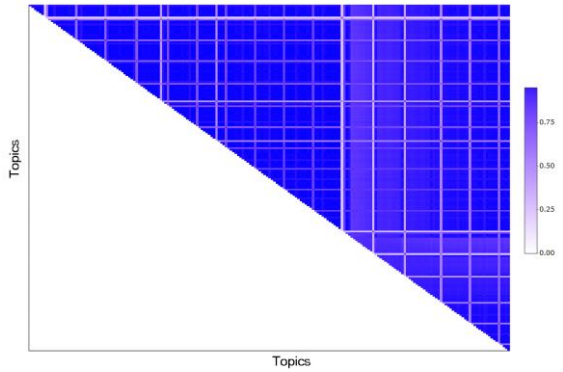

Рис. 12 (LDA Gibbs sampling).

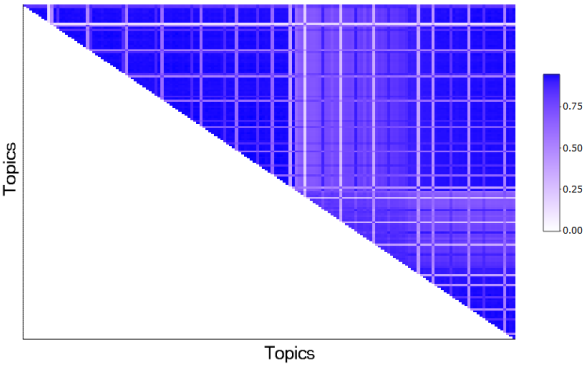

Рис. 13 (VLDA (E-M алгоритм)).

Расстояния Жаккара для модели LDA Gibbs sampling и VLDA (E-M алгоритм).

Тем не менее, двухпараметрическая энтропия также обладает минимумом, который позволяет найти оптимальное число тем. На рисунках (15), (16) приведены кривые энтропии $S_{S,M}$ для модели plsa, с обрезанными пиками (с целью визуализации минимумов, так как $S_{S,M}$ дает большой скачок при маленькой величине расстояния Жаккара). Данные рисунки показывают, что для русскоязычного датасета минимум двухпараметрической энтропии лежит в области [7-10], а для англоязычного датасета минимум находится в районе [18-20] тем, что полностью соответствует человеческой разметке.

2.8.2.2. Модели LDA GB – энтропия Шарма−Миттала

Результаты вычисления энтропии $\mathbf{S_{S,M}}$ для модели LDA GB в сравнении с моделью plsa приведены на рисунках (17), (18). Они показывают, что двухпараметрическая энтропия также позволяет правильно оценить «реальное» число тем для двух разноязычных датасетов. При этом, увеличение величины коэффициентов регуляризации α, β приводит к увеличению энтропии и смещению минимума, что нарушает возможность корректного определения числа тем в датасете. Таким образом, можно прийти к следующим выводам. Во-первых, энтропия $\boldsymbol{S_{S,M}}$ позволяет корректно определять оптимальное число тем в разноязычных датасетах. Во-вторых, энтропия $\boldsymbol{S_{S,M}}$ дает возможность корректно подбирать гиперпараметры модели LDA GB.

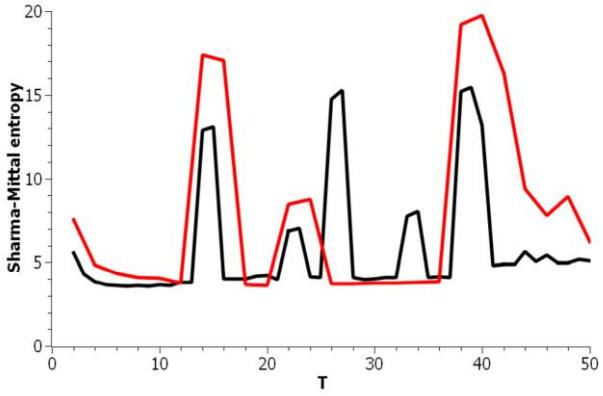

Рис. 14. Кривая энтропии $S_{S,M}$для датасета ‘lenta’ и ’20 topics news’ (модель plsa) для диагональных элементов матрицы расстояний Жаккара. Русскоязычный датасет - черная линия; Англоязычный датасет – красная линия.

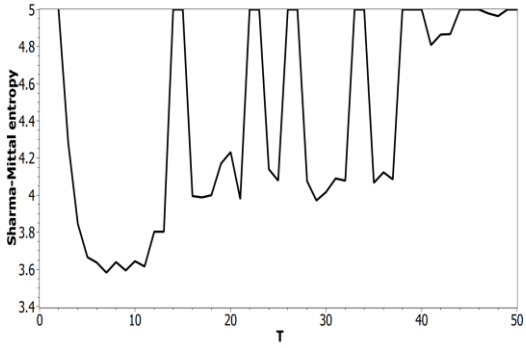

Рис. 15 $S_{S,M}$ (датасета ‘lenta’).

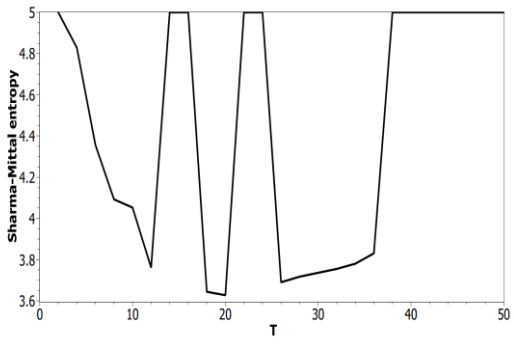

Рис. 16 $S_{S,M}$ (датасета ’20 topics news’).

Кривые энтропии $S_{S,M}$для датасета ‘lenta’ и ’20 topics news’, модель plsa с усеченными пиками.

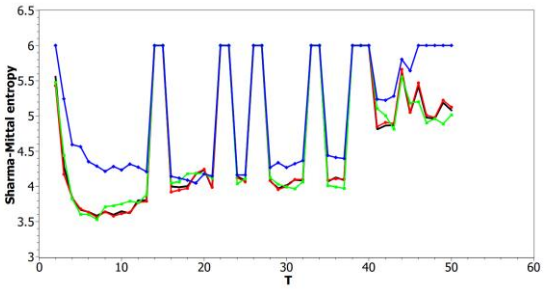

Рис. 17 $S_{S,M}$ (датасета ‘lenta’).

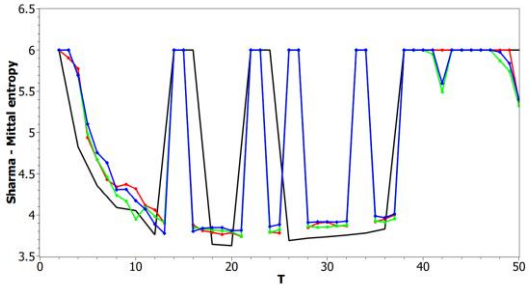

Рис. 18 $S_{S,M}$ (датасета ’20 topics news’). Кривые энтропии $S_{S,M}$(модель LDA GB vs pLSA) в зависимости от числа тем. (‘lenta’, ’20 topics news’). pLSA – черная линия; LDA (α = 0.1, β = 0.1): красная линия; LDA (α = 0.5, β = 0.1): зеленая линия; LDA (α = 1, β = 1), blue. Пики обрезаны.

2.8.2.3. Модель ARTM с разреживанием матрицы Φ и матрицы Θ – энтропия Шарма−Миттала

Модель ARTM реализована на основе принципа аддитивной регуляризации, где коэффициент регуляризации, задаваемый пользователем, определяет уровень вклада, задаваемого регуляризатора, в результат тематического моделирования. На данный момент хорошего способа определения оптимального значения коэффициента не существует. Поэтому целью данной является экспериментальная демонстрация возможности применения параметризованной энтропии для настройки коэффициентов резуляризации в модели ARTM.

В данной модели параметрами являются величины коэффициентов регуляризации и число тем. Соответственно, при исследовании данной модели производилось вариация числа тем в диапазоне [2-50] и коэффициентов τ_Φ, τ_Θв диапазоне [-10, 10]. Набор кривых энтропий как функций от числа тем приведен на рисунках (19), (20). Увеличение параметра τ_Φ, τ_Θ в энтропии $S_{S,M}$, так же, как и для энтропии Реньи, приводит к увеличению общей величины энтропии, то есть к ухудшению работы тематической модели.

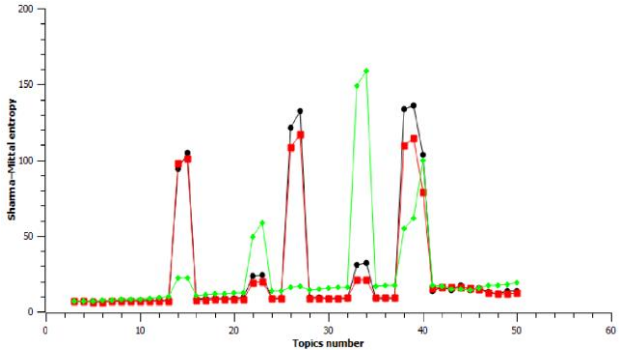

Рис. 19 энтропия $S_{S,M}$(sparse Φ).

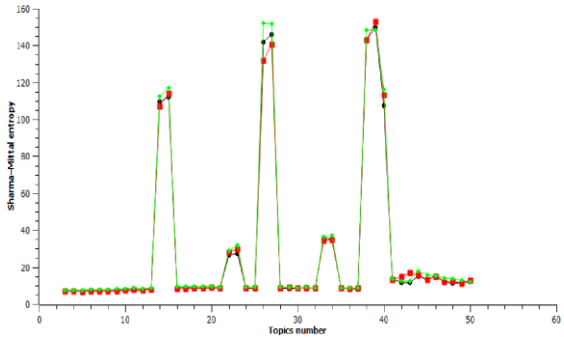

Рис. 20 энтропия $S_{S,M}$(sparse Θ).

Кривые энтропий $S_{S,M}$ модели ARTM в случае регуляризаторов sparse Φ и sparse Θ. Черный цвет: τ_Φ, τ_Θ=0.01, красный цвет: τ_Φ, τ_Θ=0.1, зеленый цвет: τ_Φ, τ_Θ=1.

Таким образом, на основании анализа проведенных компьютерных экспериментов на размеченных датасетах можно сказать следующее: 1. При вариации параметра $q = 1/T$, энтропия $S_{S,M}$ и энтропия Реньи позволяет определить оптимальное число тем, а также подобрать оптимальную величину коэффициента регуляризации; 2. Вариация параметра r (расстояние Жаккара) в энтропии $S_{S,M}$ приводит к выделению зон семантической стабильности, разделенных пиками с большой величиной энтропии. Однако, величина скачка зависит от объема слов, используемых при расчете расстояний Жаккара. 3. Минимум параметризованных энтропий, при маленьких величинах коэффициентов параметризации, соответствует результатам человеческой разметки текстовых коллекций.

2.9. Эксперименты по применению энтропии Реньи к анализу иерархических тематических моделей

Как уже было отмечено, в области тематического моделирования существует проблема не только определения оптимального числа тем, но и определения наличия ‘плоской’ или иерархической тематической структуры. В данной главе приводятся результаты экспериментального анализа поведения трех иерархических моделей на разноязычных размеченных датасетах. В экспериментах демонстрируется возможность использования параметризованной энтропии Реньи как в качестве маркера тематической стркутуры, так и для определения оптимальной числа тем на разных уровнях иерархии.

С целью проверки теоретической концепции, изложенной в параграфе 2.5, были проведены следующие эксперименты. Во-первых, в компьютерных экспериментах по применению энтропии Реньи для анализа иерархических тематических моделей были использованы следующие тематические модели: 1. HLDA (модель иерархического латентного размещения Дирихле) [61]. 2. HPAM (модель иерархического размещения Пачинко) [62]. 3. hARTM (Иерархическая аддитивная регуляризация тематических моделей) [63]. Данные модели тестировались с помощью шести размеченных датасетов, два из которых имеют плоскую разметку, а четыре - двухуровневую разметку.

Описание датасетов: 1. Датасет на русском языке (‘lenta_ru’) (пользовательская разметка на 10 тем). 2. Англоязычный датасет ‘20 newsgroup’ [58] (пользовательская разметка на 20

тем). 3. “WoS”– имеет иерархическую разметку на два уровня. Содержит 46.985 аннотаций опубликованных статей (*Web of Science)* и 80.337 уникальных слов. Первый уровень разметки содержит 7 тем (компьютерные науки, электротехника, психология, машиностроение, гражданское строительство, медицина, биохимия), а второй - 134 темы. Стоит отметить, что данный датасет сильно не сбалансирован по распределению числа документов по темам второго уровня, поэтому в данной работе также рассматривается его сбалансированное подмножество. Балансировка датасета заключалась в удалении из датасета тем, которые содержат менее 260 документов. Сбалансированный датасет “WoS” сожержит11.967 аннотаций статей и 36.488 уникальных слов, 7 тем на первом уровне и 33 на втором. 4. “Amazon” (https://data.mendeley.com/datasets/9rw3vkcfy4/1)) – датасет, имеющий иерархическую разметку на три уровня с 6, 64 и 510 темами соответственно. Содержит 40.000 отзывов о товарах из интернет-магазина *Amazon* и 31.486 уникальных слова. Третий уровень содержит пустые метки, поэтому в данной работе рассматриваются только первые два уровня иерархической разметки, а также его балансированная версия, содержащая 6 тем на первом уровне и 27 на втором. Общее число документов– 32.774, число уникальных слов – 28.422.

2.9.1. Модель HPAM

Иерархическая модель HPAM зависит от следующих параметров: 1. Число тем на втором уровне. 2. Число тем на третьем уровне. 3. Параметр ‘eta’ (η – параметр, характеризующий функцию Дирихле). 4. Параметр ‘alpha’ (α). Нужно отметить, что в модели HPAM на первом уровне число тем всегда равно 1. Также, параметр α задается в виде начального значения, а далее алгоритм его подстраивает. Исследование показало, что вариация исходной величины параметра α не влияет на результаты моделирования, поэтому параметр α не был использован в работе. Настройка параметров модели HPAM для датасетов с плоской тематической структурой осуществлялось в два этапа. На первом этапе на третьем уровне число тем фиксировалось, а число тем и параметра η на втором уровне варьировалось. На втором этапе выбирались и фиксировались число тем и величина η для второго уровня, для которых на первом этапе было получено минимальное значение энтропии Реньи, и варьировалось число тем на третьем уровне. Первый уровень по определению разработчиков модели имеет одну тему.

2.9.1.1. Датасет ‘Lenta’

Для данного датасета проводились следующие эксперименты. В первой части на первом и третьем уровнях было установлено число тем, равное единице, а число тем на втором уровне варьировалось в диапазоне [2-200]. Величина η варьировалась в диапазоне [0,001 - 1]. Так как тематическое ТМ обладает нестабильностью, все вычисления проводились по 6 раз (для заданной комбинации параметров), а энтропия Реньи усреднялась. Далее для анализа поведения тематической модели на третьем уровне иерархии отбирались наилучшие комбинации параметров, соответствующие минимальным значениями энтропии Реньи на втором уровне. Для них рассчитывалась величина энтропии Реньи на третьем уровне иерархии.

Результаты расчета энтропии Реньи как функции от числа тем и параметра η на втором уровне иерархии модели HPAM приведены на рисунке (21).

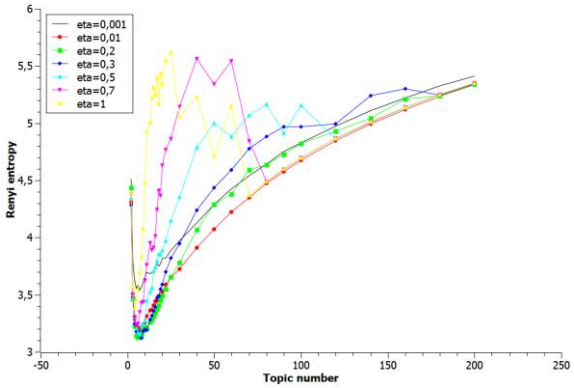

Рис. 21. Энтропия Реньи.

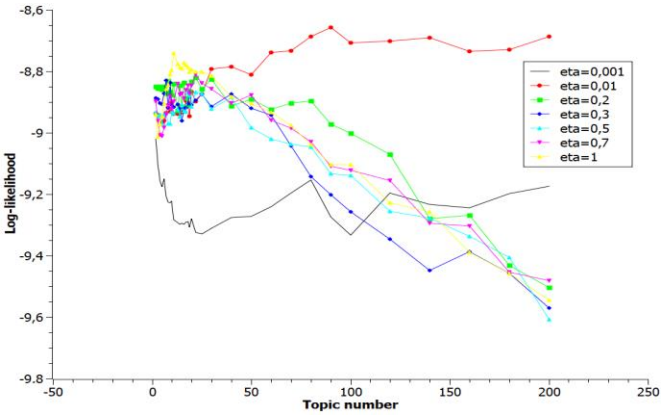

Рис. 22 Логарифма правдоподобия.

Зависимость минимума энтропии Реньи и логарифма правдоподобия от параметра η и числа тем (Lenta) на втором уровне иерархии модели HPAM.

На рисунке (22) приводится кривые логарифма правдоподобия модели HPAM. Видно, что они мало применимы для анализа, так как данная мера дает очень большие флуктуации, что не позволяет определить число тем в датасете и найти оптимальное значение параметра η. Кроме того, так как перплексия является обратной величиной от логарифма правдоподобия, она также не пригодна для реальной настройки модели HPAM.

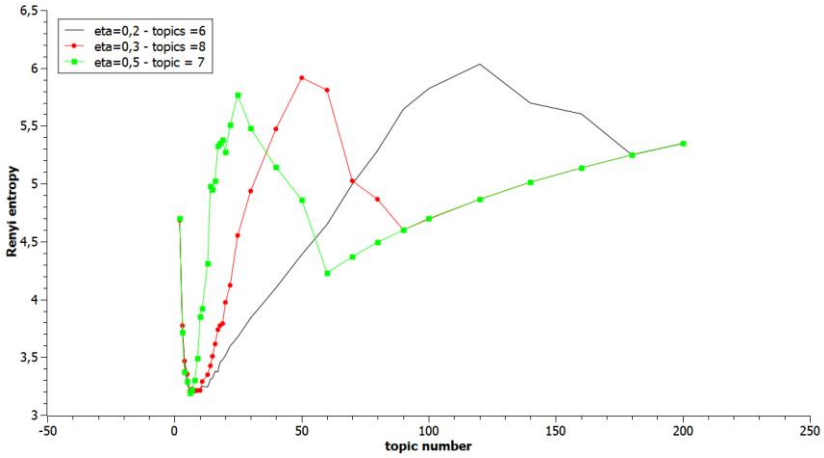

Рис. 23. Зависимость энтропии Реньи от числа тем на третьем уровне, при фиксированном числе тем на втором уровне иерархии и заданных параметров η. Модель HPAM (Lenta).

Вычисление энтропии на третьем уровне показывает, что вариация числа тем приводит к тому, что присутствует один глобальный минимум в районе 6 тем, а также резкие флуктуации энтропии при увеличении числа тем свыше 50 тем. Резкие изменения энтропии заменяются на практически прямые линии. Это связано с тем, что в области сильных флуктуаций модель ломается: число слов с высокой вероятностью и сумма вероятностей становится константой, а рост энтропии объясняется только тем, что в формуле расчета энтропии присутствует число тем, то есть число тем растет, а статистические особенности модели не меняются. Следовательно, модель HPAM может видеть один глобальный минимум при небольшом числе тем.

2.9.1.2. Датасет ’20Newsgroups’

Модель HPAM для ‘20Newsgroups’ исследовалась так же, как и для русскоязычного датасета. Кривые энтропии Реньи как функции от числа тем на втором уровне иерархии при разных параметрах η приведены на рисунке (24). В целом, их поведение при изменении числа тем и параметра η аналогично кривым энтропии Реньи для датасета ‘Lenta’.

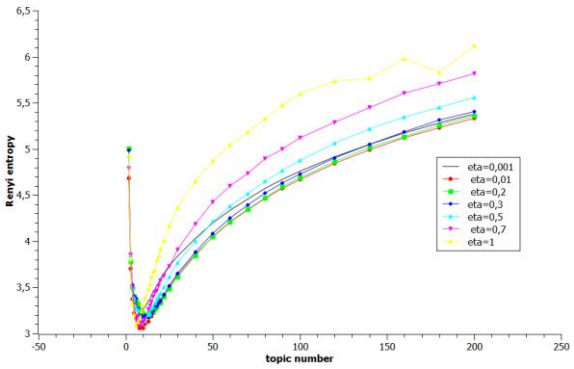

Рис. 24. Энтропия Реньи.

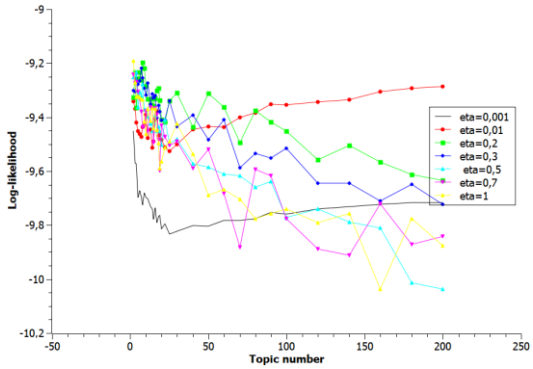

Рис. 25 Логарифма правдоподобия

Зависимость минимума энтропии Реньи и логарифма правдоподобия от параметра η и числа тем (‘20Newsgroups’) на втором уровне иерархии модели HPAM.

Рисунок (25) показывает, что логарифм правдоподобия также не пригоден для настройки модели HPAM для англоязычного датасета ‘20Newsgroups’.

2.9.1.3. Сбалансированный и несбалансированный датасеты WoS

Для этих датасетов на первом этапе проводились вычисления модели HPAM для следующего диапазона параметров: 1. Число тем варьировалось в диапазоне [2-60] с шагом в две темы, 2. Параметр η варьировался следующем образом: [0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1]. Число тем на первом и третьем уровнях иерархии фиксировалось равным единице. Модель HPAM с каждой комбинацией параметров запускалась по 6 раз, после чего рассчитывалось среднее значение величины энтропии Реньи. На рисунке (26), (27) приведены усредненные кривые энтропии Реньи для разных величин параметра η при вариации числа тем.

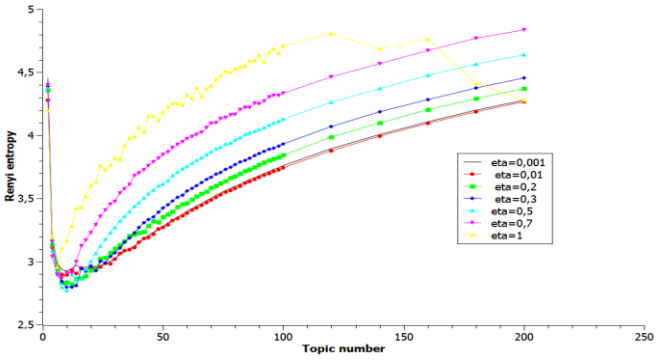

Рис. 26. Энтропия Реньи.

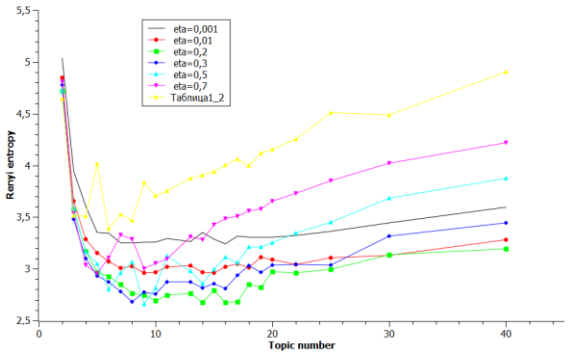

Рис. 27 Логарифма правдоподобия

Энтропия Реньи на втором уровне иерархии (сбалансированный и несбалансированный ‘WoS’).

Последние два рисунка показывают, что балансировка датасета приводит к появлению резко очерченного минимума энтропии Реньи, то есть, балансировка датасетов улучшает тематическое моделирование. Кроме того, в этом случае точность определение числа тем существенно выше.

Вычисления на третьем уровне проводились при фиксированном числе тем и соответствующая величина η на втором уровне иерархии. Далее, на третьем уровне варьировалось число тем для нескольких величин η. Результаты расчета энтропии Реньи показаны на рисунках (28), (29).

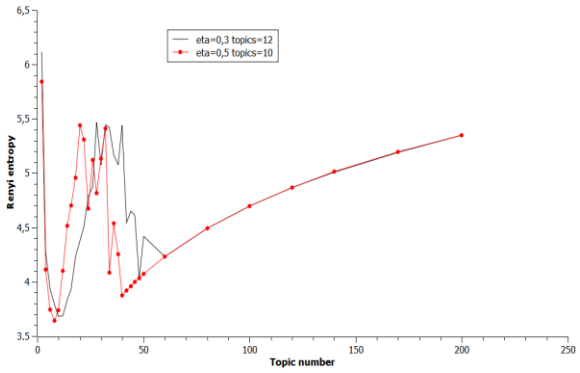

Рис. 28. Кривые энтропии Реньи (сбалансированный датасет ‘WoS’).

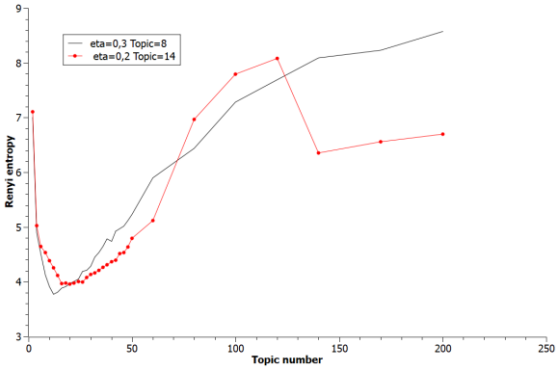

Рис. 29. Кривые энтропии Реньи (несбалансированный датасет ‘WoS’)

Кривые энтропии Реньи при вариации числа тем и параметра η на третьем уровне иерархической модели (сбалансированный и не сбалансированный датасет ‘WoS’)

2.9.1.4. Сбалансированный и несбалансированный датасеты “Amazon”

Для этих двух датасетов были проведены вычисления, аналогичные вычислениям на датасетах ‘WoS’. Результаты приведены на рисунках (30), (31) (вариация числа тем и параметра η на втором уровне) и результаты вычисления энтропии Реньи на третьем уровне показаны на рисунках (32), (33).

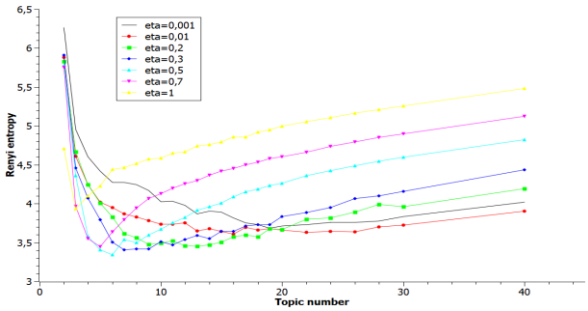

Рис. 30. Кривые энтропии Реньи (сбалансированный датасет ‘Amazon’).

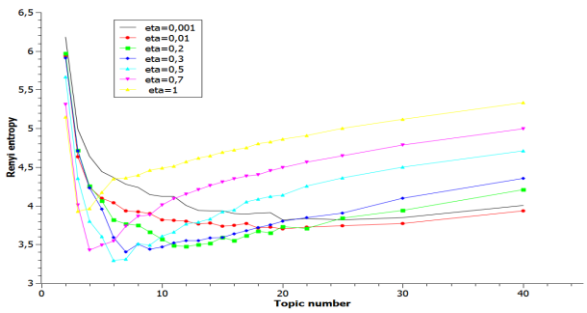

Рис. 31. Кривые энтропии Реньи (несбалансированный датасет ‘Amazon’)

Кривые энтропии Реньи на втором уровне иерархии (сбалансированный и несбалансированный датасет ‘Amazon’) при вариации числа тем и параметра η.

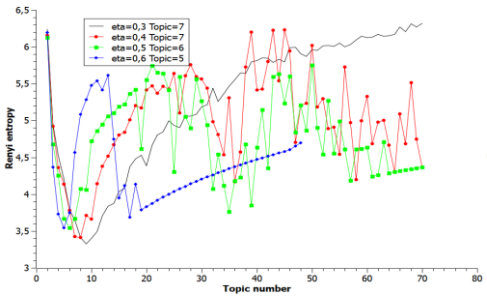

Рис. 32. Кривые энтропии Реньи (сбалансированный датасет ‘Amazon’).

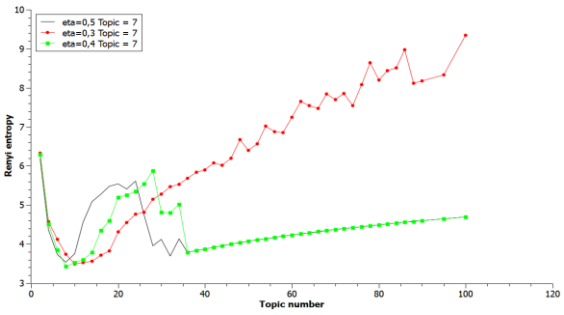

Рис. 33. Кривые энтропии Реньи (несбалансированный датасет ‘Amazon’).

Энтропия Реньи на третьем уровне модели HPAM (сбалансированный и несбалансированный датасеты ‘AMAZON’) на третьем уровне при вариации числа тем и и параметра η.

Видно, что модель HPAM дает резкие скачки энтропии как для плоских датасетов, так и для иерархически размеченных датасетов при большом числе тем. Поэтому можно заключить, что модель HPAM не способна различать плоскую и иерархическую структуру датасетов, и ее можно использовать лишь для определения одного уровня иерархии.

2.9.2. Модель HLDA

Авторы данной модели утверждают, что эта модель автоматически находит число тем в датасете, на основе иерархического процесса ‘китайский ресторан’ [61]. Однако, как показало исследование, данная модель существенно зависит от параметра концентрации и приводит к большому разбросу числа тем при вариации данного параметра [2]. Поскольку возможности корректно определять параметр концентрации не обнаружено, данная модель, не рассматривается в данной работе. Полное исследование данной модели приведено в работе [2].

2.9.3. Модели hARTM

Модель hARTM, предложенная авторами [63], имеет следующие параметры: 1. Число тем на каждом уровне. 2. Seed – параметр, характеризующий процедуру инициализации (задание генератора случайных чисел). Данная модель исследовалась для четырех рассмотренных датасетов. Результаты вычисления энтропии Реньи для плоских датасетов приведены на рисунках (34), (35).

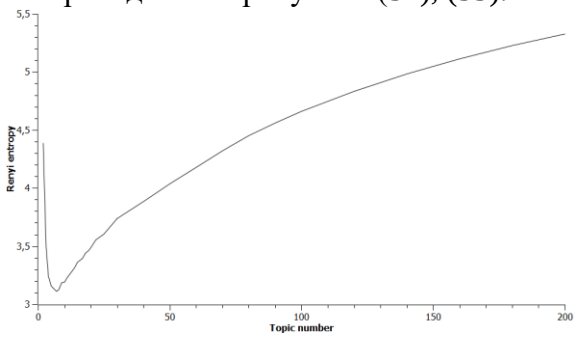

Рис. 34. Кривые энтропии Реньи (датасет ‘Lenta’).

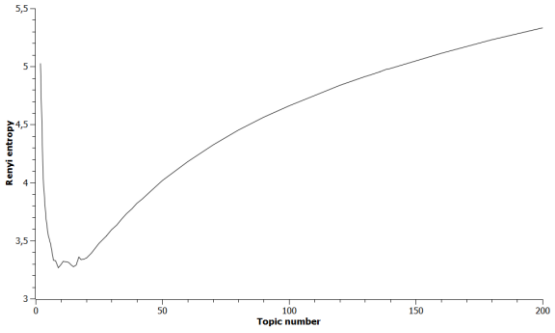

Рис. 35. Кривые энтропии Рен (датасет ’20Newsgroups’).

Зависимость энтропии Реньи от числа тем на первом уровне иерархии в модели hARTM (‘Lenta’, ’20Newsgroups’).

Вычисления показывает, что hARTM хорошо определяет плоскую структуру датестов и не дает флуктуации при большом числе тем. Результаты вычислений для датасета ‘WoS’ и приведены на рисунках (36), (37). Эти кривые показывают, что, во-первых, балансировка для датасета приводит к снижению энтропии тематической модели; во-вторых, балансировка приводит к изменению положения второго энтропийного минимума. При этом первый минимум практически не меняется. Это значит, что удаление документов, которые составляют маленькие темы, не влияет на совокупность слов с высокими вероятностями на первом уровне. Существование второго уровня иерархии демонстрируется наличием второго локального минимума. При этом, процедура балансировки влияет на положение минимума для второго уровня иерархии.

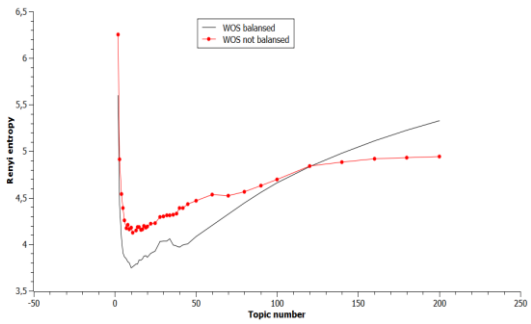

Рис. 36, 37. Кривые энтропии Реньи (датасет ‘WOS’).

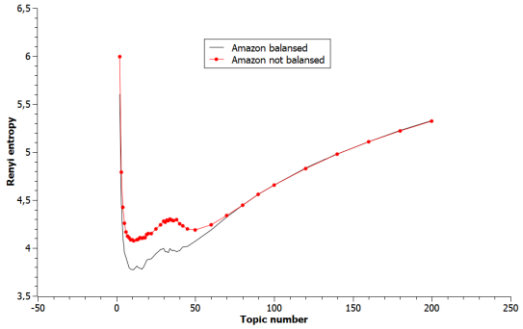

Рис. 37. Кривые энтропии Реньи (датасет ‘Amazon’).

Кривые энтропии Реньия для сбалансированного и несбалансированного датасета (‘WOS’ и ‘Amazon’) на первом уровне. Черный цвет – сбалансированный датасет, красный цвет – несбалансированный датасет.

Тем не менее, следует отметить, что второй уровень иерархии определяется с меньшей точностью, чем первый уровень. Это связано с тем, что слова на втором уровне обладают меньшими величинами вероятности, поэтому определение разницы между словами на втором уровне и словами ниже уровня $\frac{1}{W}$ затруднено в связи с нестабильностью тематических моделей.

Таким образом, на основании проведенных исследований энтропийных тематических моделей можно сделать следующий вывод. Во-первых, модели на основе энтропий Реньи и Шарма−Миттала, а именно модели LDA (Gibbs Sampling algorithm), pLSA (E-M algorithm), VLDA (E-M algorithm), GLDA (Gibbs Sampling algorithm) и ARTM с регуляризаторами разреживания матриц Φ и Θ (E-M алгоритм) позволяют определять оптимальные гиперпараметры тематических моделей. Оптимальное число тем в тематических моделях определяется при помощи поиска минимума параметризованных энтропий. Вариация параметров регуляризации приводит к смещению всей кривой энтропии, при этом лучшая величина параметра регуляризации соответствует низшей кривой энтропии (среди всех кривых, полученных при вариации параметра). Во-вторых, использование двухпараметрической энтропии позволяет оценить семантическую стабильность тематических моделей, при вариации гиперпараметров моделей, включая количество тем. В-третьих, иерархическая энтропийная тематическая модель на основе hARTM дает возможность определять наличие иерархической или ‘плоской’ структуры в разноязычных датасетах и корректно устанавливать оптимальное число тем на двух уровнях иерархии.

3. Фрактальная модель оценки результатов тематической модели

Поведение информационной статистической системы можно исследовать при помощи фрактальной модели. Это обусловлено тем, что энтропия Реньи хорошо описывает фрактальные статистические системы [50]. В основе данного математического формализма лежит процедура скейлинга, то есть изменение масштаба. Фрактальная модель выглядит следующим образом [64]. Тематическое решение при фиксированном числе тем представляет собой матрицу Φ, в которой общее число ячеек равно $T * W$, где T – число тем (колонок в матрице), W – число уникальных слов (число строк). Каждая ячейка матрицы содержит вероятность p_{ij} принадлежности слова w_i к теме T_j, а размер ячейки равен величине $\varepsilon \sim 1/(WT)$. При фиксированном размере словаря $W = const$, размер ячейки определяется только количеством тем в модели, и при $T \to \infty$ размер ячеек стремится к нулю. Функция плотности распределения слов имеет вид: $\check{\rho} = \frac{N_i}{W\mathrm{T}}$, где N_i – число ячеек в

тематическом решении, вероятность которых (p_{ij}) выше величины $\frac{1}{W}$, то есть данная функция оценивает облако высоко вероятностных слов, и она является функцией от числа тем, и в ходе тематического моделирования меняется с 0 до некоторого значения $\check{\rho}_i(E)$<1, которое зависит от количества тем. Следовательно, плотность $\check{\rho}(E)$ зависит от размера ячеек и степени $D(\varepsilon)$ [64]: $\check{\rho}(E) \cong \varepsilon^{-D(\varepsilon)}$. Распределение фрактальных размерностей D(ε) определялось при помощи алгоритма 'box counting' [65]. Его применение к подсчету фрактальных размерностей в ТМ состоит из следующих этапов: 1. Пространство слов покрывается сеткой фиксированного размера, которая является матрицей $\Phi = \phi_{wt}$. 2. Подсчитывается количество ячеек, где вероятность слов больше величины $p_{wt} > 1/W$. 3. Рассчитывается величина ρ_{wt} для заданной величины числа тем T_t. 4. Повторяются шаги 1, 2, 3 при изменении размера ячейки, то есть при изменении числа тем. 5. Строится график зависимости $\check{\rho}(E)$ в билогарифмических координатах. 6. Методом наименьших квадратов оценивается наклон на графике, и он представляет собой фрактальную размерность, взятую с обратным знаком: $D(E) = -\frac{\ln(p(E))}{\ln(\varepsilon)}$. Линейные части графика в билогарифмических координатах характеризуют процесс самовоспроизводства функции плотности распределения слов в тематических моделях.

3.1. Эксперименты по определению фрактальной размерности в тематических моделях

В исследовании фрактальных свойств тематических моделей проводился набор компьютерных экспериментов. В расчетах использовались датасеты: 1. 'Lenta'. 2. '20 newsgroups dataset'. Для обеих коллекций была проведена серия расчетов, в которых число тем менялось в диапазоне [2-50] с шагом 1. Все модели запускались по три раза, а результаты моделирования усреднялись. Для каждого усредненного решения рассчитывалось значение $\check{\rho}(E)$. Полученные кривые анализировались в билогарифмических координатах. В экспериментах были использованы тематические модели: 1. pLSA (E-M алгоритм); 2. ARTM (E-M алгоритм); 3. LDA Gibbs sampling. Примеры моделирования и расчеты фрактальной степени приведены на рисунках (38), (39), (40), (41).

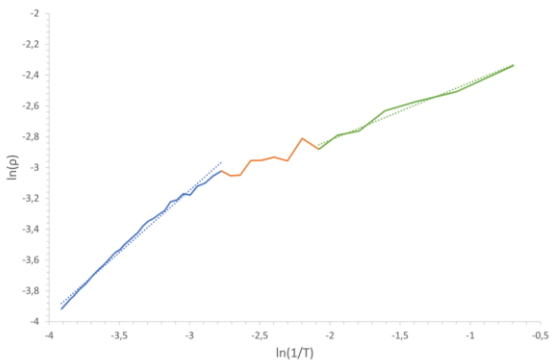

Рис. 38. Распределение фрактальных размерностей (модель pLSA)

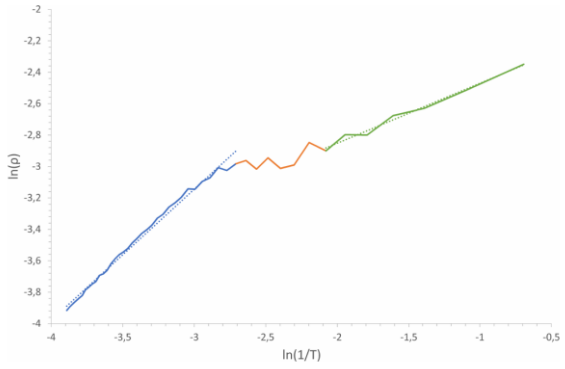

Рис. 39. Распределение фрактальных размерностей (модель LDA GB (α=0.4, β=0.5))

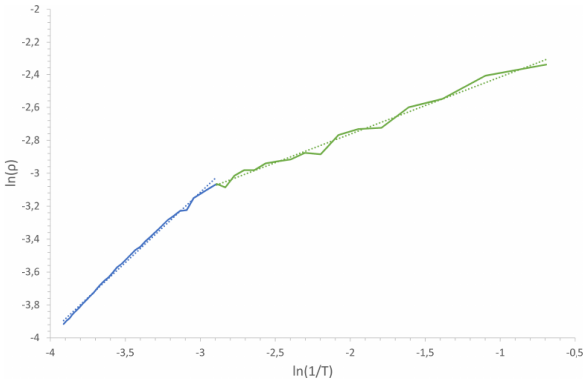

Рис. 40. Распределение фрактальных размерностей (ARTM (sparse Ф = 0.01)).

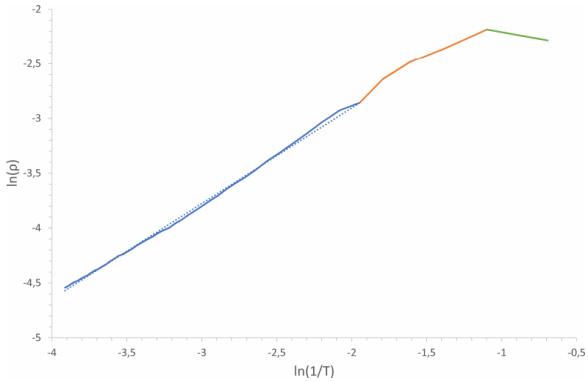

Рис. 41. Распределение фрактальных размерностей (ARTM (sparse Ф = -10)

Фрактальный анализ поведения тематических моделей показывает, что в текстовых коллекциях существуют самоподобные области, а между ними существует область перехода, которая соответствует минимуму энтропии Реньи. При этом промежуточная область между линейными областями соответствует минимуму энтропии Реньи [64]. Таким образом, проблему анализа эволюции тематических моделей при вариации числа тем можно свести к проблеме поиска областей перестройки тематических моделей. Последняя рассматривается в следующей главе и решается за счет применения к тематическому моделированию теории ренормализации.

4. Метод агрегации тематических моделей на основе процедуры ренормализации.

4.1. Введение в теорию ренормализации

Ренормализация – математический формализм, который активно используется в различных областях физики, таких как анализ перколяций, и анализ фазовых переходов. Ренормализация заключается в построении процедуры изменения масштаба системы, при котором поведение системы остается тем же самым. Теоретические основы процедуры ренормализации были заложены в работах Каданоффа [33], Вильсона [34]. Широкое развитие процедуры перенормировки (ренормализации) получили в теории фракталов, так как фрактальное поведение обладает свойством самоподобия [35, 36].

Суть процедуры ренормализации заключается в следующем. Рассмотрим решетку, состоящую из совокупности узлов. Мы не рассматриваем физические особенности этих узлов и приводим лишь формулировку процедуры ренормализации. Каждый из узлов характеризуется направлением спина. В свою очередь спин может занимать определение направление, количество которых зависит от задачи. Так, например, в модели Изинга рассматривается всего два положения спина, в модели Потса число положений может быть порядка 3-5 состояний [37].

Узлы с одинаковыми спинами составляют кластеры. Процедура скейлинга или ренормализаци происходит по принципу блочного объединения, в котором несколько ближайших узлов заменяются на один узел. В качестве направления нового спина берется направление спинов, составляющих большинство в выбранном блоке. Процедура блочного объединения проводится по всей поверхности. Соответственно, в результате появляется новая конфигурация спинов. Процедуру огрубления можно проводить несколько раз. Исходя из принципа, что новая конфигурация спинов должна быть эквивалентна старой конфигурации, возникает возможность построить процедуру вычисления полевых параметров и значений показателей критических индексов. Следует отметить, что последовательное применение процедур ренормализации или процедуры огрубления исходной системы дает приближенные результаты, однако, несмотря на это, данная методика широко используется, так как позволяет получить оценки критических индексов при фазовых переходах, где стандартные математические модели не применимы.

Процедура ренормализации успешно применяется там, где наблюдается масштабная инвариантность. Масштабная инвариантность характеризуется степенными распределениями. Математическое выражение самоподобия выражается следующим образом. Пусть $f(x)=cx^{\alpha}$, где c, α - константы. Если сделать масштабное преобразование вида $x \rightarrow \lambda x$, то получим тот же функциональный вид, но с другим коэффициентом, то есть, $f(\lambda x)=\beta x^{\alpha}$. Таким образом, степенные законы характеризуются масштабной инвариантностью. Степенной параметр можно определять при помощи разных алгоритмов, например, таких как 'box counting'.

4.2. Общая постановка задачи агрегирования в виде процедуры ренормализации в тематическом моделировании

Общая задача агрегирования тематических моделей при вариации размера смеси распределения заключается в применении технологии ренормализации. Понятие ренормализации (или перенормировки) заимствуется из в квантовой теории поля и представляет собой итеративный метод перенормировки, в котором переход от областей с меньшей энергией к областям с большей связан с изменением масштаба системы. Ренормализация тесно связана с масштабной инвариантностью и конформной инвариантностью, и с симметриями, в которых система кажется одинаковой на всех масштабах (так называемое самоподобие). Ренормализация тематических моделей реализована следующим образом [65]. Результатом ТМ является матрица $\Phi = \phi_{wt}$, которая состоит из ряда одномерных распределений слов по темам. Размеры матрицы определяются числом слов W и числом тем T. В данной работе рассматривается фиксированный словарь уникальных слов, поэтому изменение масштаба тематической модели зависит только от параметра $q = 1/T$. Процедура ренормализации — это процедура слияния двух тем в одну. После объединения двух тем, нормируется новая тема, так как сумма вероятностей всех слов в одной теме всегда равна единице. В силу того, что расчет элементов матрицы ϕ_{wt} зависит от типа модели, математическая формулировка процедуры ренормализации специфична для каждой модели. Кроме того, результат слияния зависит от того, какие темы попарно объединяются. В данной работе рассматриваются три принципа объединения тем:

1) Принцип попарного объединения тем на основе минимума дивергенции Кульбака-Лейблера. Он предполагает, что объединять нужно темы, которые имеют сходные распределения вероятностей. Для этого производится попарный расчет по следующей формуле:

$$D_{KL}(p \mid q) = \sum_{i=1} p(x_i) \cdot ln\left(\frac{p(x_i)}{q(x_i)}\right) = \\ = -\sum_{i=1} p(x_i) \cdot ln\big(q(x_i)\big) + \sum_{i=1} p(x_i) \cdot ln\big(p(x_i)\big) \qquad (6).$$

Объединяются темы, между которыми получается наименьшая величина дивергенции Кульбака-Лейблера.

2) Принцип объединения тем на основе минимума энтропии Реньи, рассчитываемой для каждой темы. Расчет производится по формуле (4), только при суммировании используются вероятности слов одной темы. Далее, две темы объединяются в одну, если у этих тем наименьшие значения энтропии.
3) Объединение случайно выбранных тем.

Ниже рассмотрены три процедуры ренормализации тематических моделей, основанные на разных алгоритмах восстановления скрытых распределений. Первая и третья модель основана на E-M алгоритме (VLDA, pLSA), а вторая модель – на процедуре сэмплирования Гиббса (LDA GB).

4.3. Процедура ренормализации для модели VLDA на основе процедуры E-M алгоритма

В модели VLDA (variational Latent Dirichlet Allocation) [66] параметрами является число тем, а результатами расчета: T-мерный вектор α, где каждая величина α_i характеризует распределение Дерихле для каждой темы, и матрица распределения слов w по темам t: $\Phi = (\phi_{wt})_{w \in W, t \in T}$. Для оценки значений матрицы Φ используется вариационный E-M алгоритм, для оценки значений вектора α используется метод Ньютона-Рапсона. Счетчик в данном алгоритме определяется следующим выражением [65]:

$$\mu_{nt} = \phi_{w_n t} \exp\left(\psi\left(\alpha_t + \frac{L}{T}\right)\right), \qquad (7)$$

где L – длина документа, n – номер текущего слова документа, w_n – слово из списка словаря уникальных слов, соответствующие данному текущему слову, ψ – дигамма-функция, μ_{nt} – вспомогательная переменная, играющая роль счетчика, так как через неё выражается ϕ_{wt} с учетом нормировки в ходе вариационного E-M алгоритма.

Для задачи ренормализации был использована сумма счетчиков. На выходе данного алгоритма мы имеем матрицу $(\phi_{wt})_{w \in W, t \in T}$ и вектор α. Алгоритм ренормализации состоит из следующих шагов:

1) Выбор пары тем для склейки одним из перечисленных в 4.1 способом. Обозначим две выбранные темы за t_1 и t_2.
2) Склеивание выбранных тем. Значения распределения “новой” темы $\phi_{\cdot t_1}$, полученной при склеивании t_1 и t_2, получается следующим образом [29]:
$$\phi_{wt_1} := \phi_{wt_1} \cdot \exp\left(\psi\left(\alpha_{t_1}\right)\right) + \phi_{wt_2} \cdot \exp(\psi(\alpha_{t_2})). \qquad (8)$$
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 Φ и элемент α_{t_2} из вектора α. На этом шаге происходит уменьшение числа тем на одну, то есть получаем $T - 1$ тем. Далее, новые значения вектора α нормируются так, чтобы сумма компонент вектора была равна 1.
3) Расчет итогового значения энтропии Реньи для уменьшенного числа тем. После того, как сформировано новое тематическое решение, по итогам нового решения рассчитывается энтропия Реньи по формуле (4), то есть по всему решению.

Шаги 1, 2, 3 повторяются до тех пор, пока не останется только 2 темы. По итогам ренормализации строится кривая энтропии Реньи как функция от параметра ренормализации, то есть от числа тем. Далее кривая Реньи, полученная при помощи ренормализации, сравнивается с кривой Реньи, полученной при последовательном расчете тематических моделей, при вариации числа тем. Сравнивая две кривые, можно оценить эффект ренормализации для данной модели. Область минимума энтропии Реньи соответствует области оптимального числа тем.

4.4. Процедура ренормализации для модели LDA на основе процедуры сэмплирования Гиббса

Модель LDA (Латентное размещение Дирихле) с сэмплированием Гиббса [15] использует симметричные распределения Дирихле, где распределение слов в теме характеризуется параметром β, а распределение тем в документах характеризуется параметром α.

Матрица $\Phi = (\phi_{wt})$ рассчитывается с помощью процедуры сэмплирования Гиббса. Значения β, α и число тем задаются пользователем. Вычисление матрицы Ф состоит из двух этапов. На первом этапе происходит процедура сэмплирования, в ходе которого

формируется счетчик c_{wt}. На втором этапе происходит вычисление элементов матрицы ϕ_{wt} по следующей формуле:

$$\phi_{wt} = \frac{c_{wt} + \beta}{(\sum_{w \in W} c_{wt}) + \beta W}, \qquad (9)$$

где c_{wt} –счетчик, равный тому, сколько раз слово w было отнесено к теме t. Для задачи ренормализации результатов данной модели мы используем содержимое счетчика c_{wt} и формулу (9). На выходе алгоритма мы имеем матрицы $(\phi_{wt})_{w \in W, t \in T}$ и c_{wt}. На вход процедуры ренормализации подается матрица c_{wt}, которая подвергается процедуре ренормализации, и на основании нее рассчитывается итоговая ренормализационная матрица ϕ_{wt}.

Алгоритм ренормализации для процедуры сэмплирования Гиббса состоит из следующих шагов:

1) Выбор пары тем для склейки одним из трех способов. Обозначим 2 выбранные темы за t_1 и t_2.
2) Склеивание выбранных тем. Новая тема получается за счет суммировании частот слов c_{wt} двух выбранных тем. Затем на основании новых значений счетчика производится расчет элементов матрицы ϕ_{wt}. Формула ренормализации выглядит следующим образом [68]:

$$\phi_{wt_1} := \frac{c_{wt_1} + c_{wt_2} + \beta}{(\sum_{w \in W} c_{wt\_1} + c_{wt_2}) + \beta W}. \qquad (10)$$

Новая тема $\phi_{\cdot t_1}$ уже удовлетворяет свойству: $\sum_{w \in W} \phi_{wt_1} = 1$. Затем удаляем столбец $\phi_{\cdot t_2}$ из матрицы Ф, то есть уменьшаем размеры тематического решения.

Шаги 1 и 2 повторяются до тех пор, пока не останется только 2 темы. По итогам ренормализации строится кривая энтропии Реньи как функция от параметра ренормализации, то есть от числа тем.

4.5. Процедура ренормализации для модели pLSA

Модель pLSA является наиболее простой, так как в ней отсутствуют регуляризаторы, параметры, и единственным параметром является число тем [13]. Алгоритм ренормализации в данном случае состоит из следующих шагов:

1) Выбор пары тем для склейки одним из трех способов. Обозначим пару выбранных тем за t_1 и t_2.
2) Склеивание выбранных тем. В этой модели новая тема выражается через простое суммирование вероятностей:

$$\phi_{wt_1} := \phi_{wt_1} + \phi_{wt_2} \qquad (11).$$

3) Нормировка новой темы. После суммирования производится нормировка новой темы, так, чтобы сумма вероятностей в новой теме была равна 1. Далее мы удаляем столбец $\phi_{\cdot t_2}$ из матрицы Ф.

Шаги 1, 2, 3 повторяются до тех пор, пока не останется только 2 темы. По итогам ренормализации строится кривая энтропии Реньи как функция от параметра ренормализации, то есть от числа тем.

4.6. Эксперименты ренормализации

В рамках исследования ренормализации были использованы три датасета: 1. Датасет на русском языке (Lenta.ru). 2. Датасет ‘20 newsgroups dataset’. 3. Франкоязычный датасет из 25000 документов на французском языке и 18749 уникальных слов. Франкоязычный датасет не содержит разметки по темам. На этих датасетах производилось тематическое моделирование в диапазоне 2-100 тем с шагом в 1 тему. Далее были использованы следующие параметры для модели LDA с сэмплированием Гиббса: α=0.1, β=0.1. Исследование оптимальных гиперпараметров для этих датасетов было проведено в работе [15], поэтому в данной работе параметры не варьировались. Далее тематические решения

на 100 тем во всех датасетах были подвергнуты процедуре ренормализации с шагом в одну тему. На основании процедуры ренормализации строились кривые энтропии Реньи как функции от числа тем. Наконец, кривые энтропии Реньи, полученные в ходе ренормализации, были сравнены с кривыми энтропии Реньи, которые были получены обычным способом (без ренормализации).

4.6.1. Ренормализация модели LDA GB (датасет ‘Lenta’)

Результаты расчета для данной модели приведены на рисунках (43)-(45).

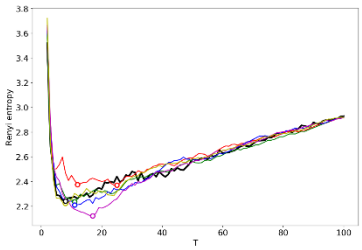

Рис. 43.

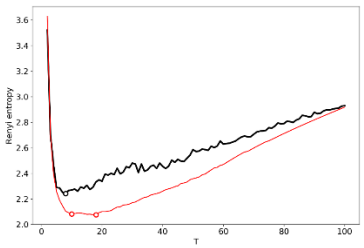

Рис. 44.

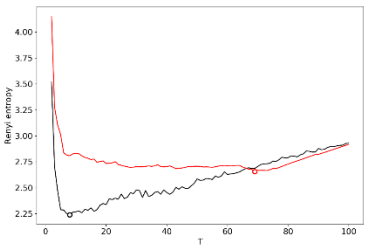

Рис. 45.

Кривые энтропии Реньи как функции от числа тем для русскоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. (43) - случайное объединение тем. Рис. (44) – объединение на основе минимума энтропии Реньи. Рис. (45) – объединение на основе дивергенции Кульбака – Лейблера.

Последний график демонстрирует, что ренормализация тематической модели на основе минимума дивергенции Кульбака-Лейблера производит наихудший результат среди трех типов ренормализации. Наилучший результат, с точки зрения определения оптимального числа тем, демонстрирует процедура объединения на основе минимума энтропии Реньи.

4.6.2. Ренормализация модели LDA GB (датасет ’20Newsgroups’)

Результаты расчета для данной модели приведены на рисунках (46)-(48).

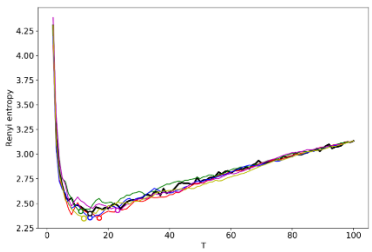

Рис. 46.

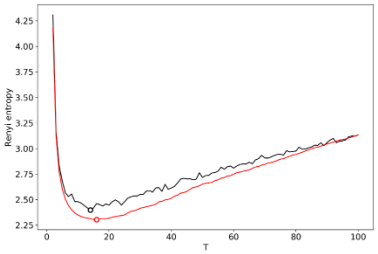

Рис. 47.

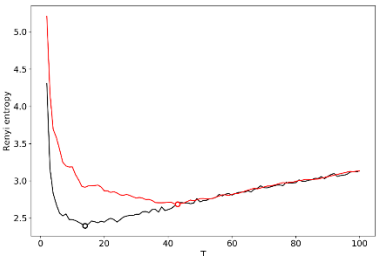

Рис. 48.

Кривые энтропии Реньи как функции от числа тем для англоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 46 – случайное объединение тем. Рис. 47 – объединение на основе минимума энтропии Реньи. Рис. 48 – объединение на основе дивергенции Кульбака – Лейблера.

Эти вычисления также показывают, что наилучший результат демонстрирует объединение при помощи минимума энтропии Реньи.

4.6.3. Ренормализация модели LDA GB (франкоязычный датасет)

Результаты ренормализации для франкоязычного датасета приведены на следующих 3 рисунках.

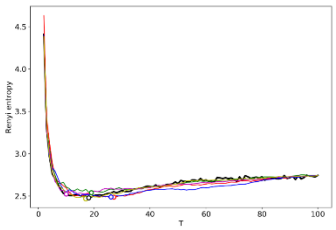

Рис. 49.

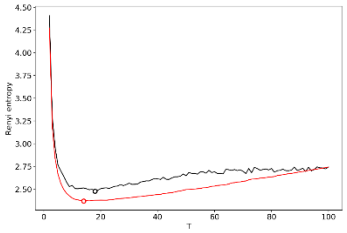

Рис. 50.

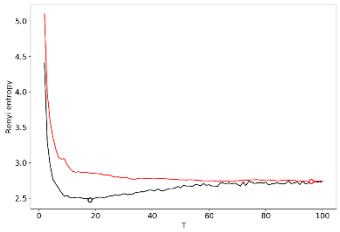

Рис. 51.

Кривые энтропии Реньи как функции от числа тем для франкоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 49 - случайное объединение тем. Рис. 50 – объединение на основе минимума энтропии Реньи. Рис. 51 – объединение на основе дивергенции Кульбака – Лейблера.

4.6.4. Ренормализация модели VLDA (датасет ‘Lenta’)

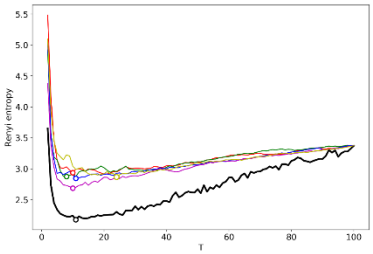

Рис. 52.

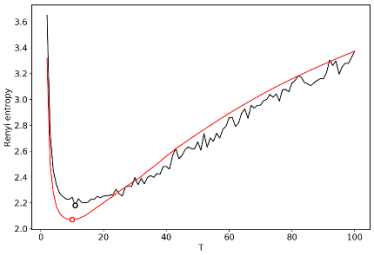

Рис. 53.

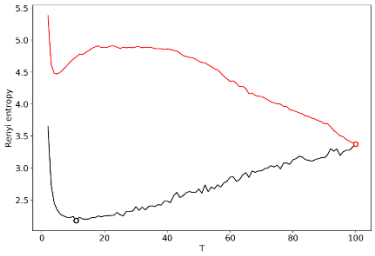

Рис. 54.

Кривые энтропии Реньи как функции от числа тем для русскоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. (52) – случайное объединение тем. Рис. (53) – объединение на основе минимума энтропии Реньи. Рис. (54) – объединение на основе дивергенции Кульбака – Лейблера.

4.6.5. Ренормализация модели VLDA (датасет ’20Newsgroups’)

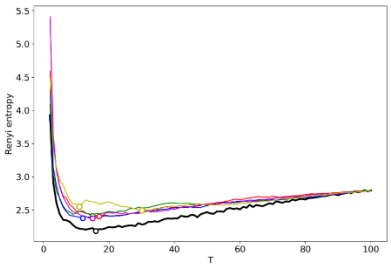

Рис. 55.

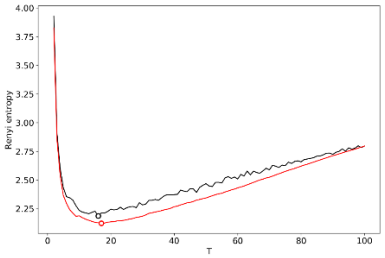

Рис. 56.

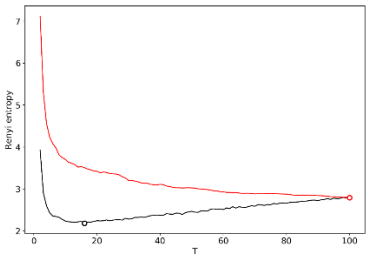

Рис. 57.

Кривые энтропии Реньи как функции от числа тем для англоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 55 – случайное объединение тем. Рис. 56 – объединение на основе минимума энтропии Реньи. Рис. 57 – объединение на основе дивергенции Кульбака – Лейблера.

4.6.6. Ренормализация модели VLDA (франкоязычный датасет)

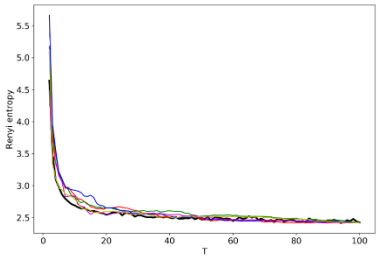

Рис. 58.

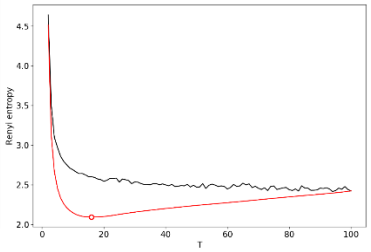

Рис. 59.

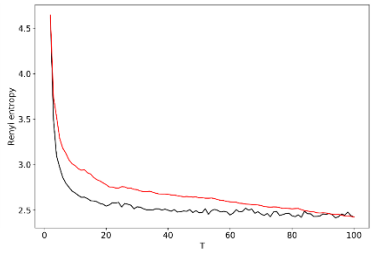

Рис. 60.

Кривые энтропии Реньи как функции от числа тем для франкоязычный датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 58 – случайное объединение тем. Рис. 59 – объединение на основе минимума энтропии Реньи. Рис. 60 – объединение на основе дивергенции Кульбака – Лейблера.

Ренормализация модели VLDA также показывает, что лучший результат, с точки зрения определения оптимального числа тем, показывает процедура объединения на основе минимума энтропии Реньи, а наихудший результат – с помощью минимума дивергенции Кульбака-Лейблера. Однако ренормализационная модель VLDA работает хуже, чем модель LDA GB.

4.6.7. Ренормализация модели pLSA (датасет ‘Lenta’)

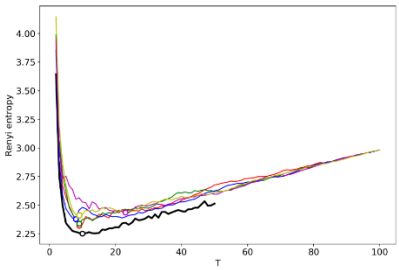

Рис. 61.

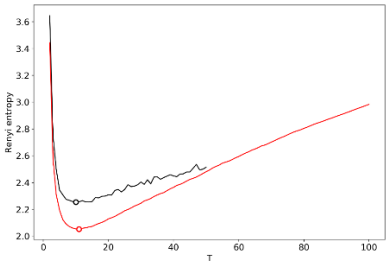

Рис. 62.

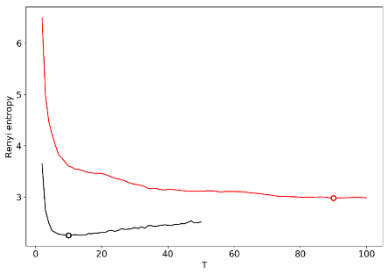

Рис. 63.

Кривые энтропии Реньи как функции от числа тем для русскоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 61 – случайное объединение тем. Рис. 62 – объединение на основе минимума энтропии Реньи. Рис. 63 – объединение на основе дивергенции Кульбака – Лейблера.

4.6.8. Ренормализация модели pLSA (датасет ’20Newsgroups’)

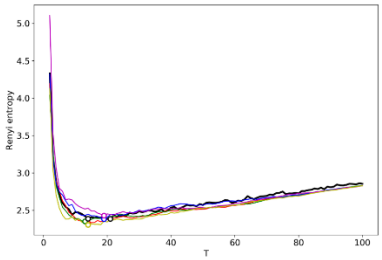

Рис. 64.

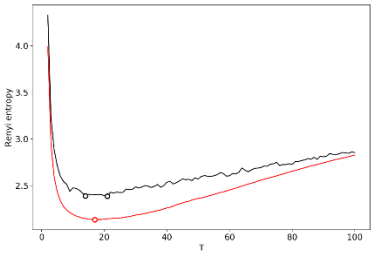

Рис. 65.

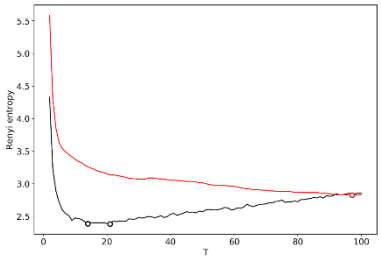

Рис. 66.

Кривые энтропии Реньи как функции от числа тем для англязычный датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 64 – случайное объединение

тем. Рис. 65 – объединение на основе минимума энтропии Реньи. Рис. 66 – объединение на основе дивергенции Кульбака – Лейблера.

4.6.9. Ренормализация модели pLSA (франкоязычный датасет)

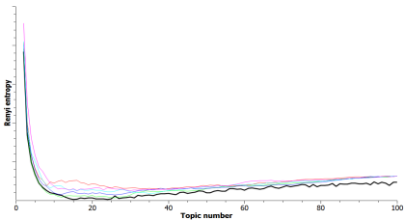

Рис. 67.

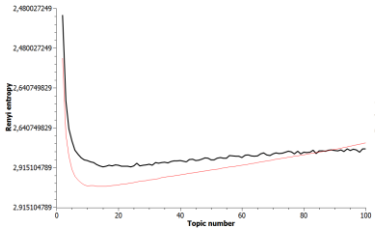

Рис. 68.

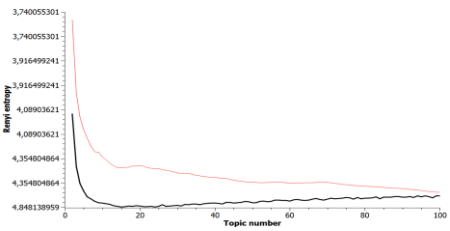

Рис. 69.

Кривые энтропии Реньи как функции от числа тем для франкоязычного датасета: черная линия – расчет энтропии по каждой тематической модели. Рис. 67 – случайное объединение тем. Рис. 68 – объединение на основе минимума энтропии Реньи. Рис. 69 – объединение на основе минимума дивергенции Кульбака – Лейблера.

Таким образом, на основании проведенных исследований можно заключить следующее: 1. Лучше всего работает процедура ренормализации на основе минимума энтропии Реньи. 2. Ренормализация на основе минимума дивергенции Кульбака-Лейблера не пригодна для определения оптимального числа тем в текстовых коллекциях. 3. Ренормализация работает для различных европейских языков.

4.7. Сравнение методов агрегации по быстродействию моделей на основе процедуры ренормализации

Время вычисления последовательного расчета тематических моделей и время проведения процедуры ренормализации для размеченных датасетов приведено в таблице (4).

Таблица 4 – Время расчета 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

Таблица демонстрирует, что процедура ренормализации выполняется в 10-800 раз быстрее, чем последовательный расчет тематических моделей при вариации количества тем. Расчет по трем процедурам ренормализации на разноязычных датасетах показывает, что, во-первых, наиболее быстрыми процедурами являются ренормализации на основе случайного объединения тем и объединения тем с минимальными значениями локальной энтропии Реньи. Самой медленной является ренормализация на основе дивергенции Кульбака-Лейблера, она также показывает наихудший результат с точки зрения сходства между кривой энтропии Реньи, полученной на основе последовательного расчета

тематических моделей, и ренормализационной кривой энтропии Реньи. Наилучший результат обеспечивает ренормализация на основе слияния тем с помощью локальной энтропии Реньи. Случайное объединение тем приводит к существенной флуктуации минимума энтропии Реньи, однако если произвести усреднение ренормализационных кривых по разным запускам, то усредненная кривая энтропии Реньи также позволяет определить оптимальное значение числа тем. Таким образом, наиболее удобной и с точки зрения скорости расчета, и с точки зрения корректного определения оптимального количества распределений в смеси тем является процедура ренормализации на основе минимума локальной энтропии Реньи.

5. Гранулированный вариант тематической модели

Как уже было отмечено, в области тематического моделирования одной из основных проблем является проблема стабильности. При этом основная часть научных работ направлена на измерения стабильности при помощи разных мер качества. В данной главе, в отличие от других работ предлагается новая модель, которая позволяет существенно улучшить стабильность тематических моделей на основе процедуры сэмплирования Гиббса. В данной части работы рассматривается модель, которая является модификацией модели LDA на основе сэмплирования Гибсса за счет задания явного вида функции локальной плотности распределения слов по темам внутри окна заданного размера. Задается окно Парзена-Розенблатта [69]:

$$p(r) = \frac{1}{mh}\sum_{i=1}^{m} K(\frac{r-r_i}{h}), \quad (12),$$

где K(w) — произвольная чётная функция, называемая ядром. Ядро K(w) должно удовлетворять условию нормировки $\int K(r)dr = 1$. На практике часто используются следующие ядра: 1. Прямоугольное ядро. $K_r = const$ при заданной ширине окна h. 2. Ядро Епанечникова (Epanechnikov kernel) [70] $K_r = const \cdot (1 - r^2)$. 3. Треугольное ядро $K_r = const \cdot (1 - |r|)$. Эти же ядра использовались в данной работе для регуляризации тематического моделирования.

5.1. Регуляризация тематической модели при помощи задания локальной плотности распределения слов по темам

В настоящей работе регуляризация тематической модели основана на идее о существовании тематической зависимости между парой уникальных слов, то есть предположение о существовании локальной функции плотности распределения тем, которую можно задать с помощью ядерной функции. Мы предполагаем, что тема состоит из слов, которые не просто описываются распределением Дирихле, но и часто встречаются рядом с другом в тексте. Задавая вид функции распределения слов по темам, внутри окна (локальная плотность) и размер окна, можно влиять на характер регуляризации модели.
В общем виде алгоритм сэмплирования Гиббса с учетом локальной плотности распределения слов по темам выгляди следующим образом:

- Начальная инициализации матриц $\Phi = \phi_{wt}, \Theta = \theta_{td}$.
- Внешний цикл – определяет количество итераций.
    - Цикл по документам
        - Цикл по словам внутри текущего документа.

Во внутреннем цикле реализовано случайное сэмплирование в соответствии с распределениями Дирихле [21]. Оно заключается в том, что для случайно выбранного якорного слова (центрального слова в окне) рассчитывается принадлежность теме, а темы остальных слов внутри окна определятся посредством локальной функции плотности: $T(w_i) = T_0 \cdot K(w_0)$, где T_0 – тема якорного слова, полученная из распределений Дирихле, $K(w_0)$ – функция локальной плотности, w_i – слова из окна.

        - Конец цикла по словам

- Конец цикла по документам
- Конец внешнего цикла (по числу итераций)

На заключительном этапе тематического моделирования, после того, как закончится сэмплирование, на основании счетчиков производится окончательный расчет матриц ϕ_{wt}, θ_{td} распределений слов и документов по темам. Таким образом, задавая вид локальной функции плотности распределения слов по темам и размер окна, мы производим регуляризацию тематической модели [72].

5.1.1. Прямоугольное ядро регуляризации (гранулированный метод сэмплирования, GLDA)

В качестве первого ядра в данной работе рассматривалась функция типа ‘ступенька’. Ее смысл в том, что все слова внутри заданного окна имеют одну и ту же тему $K(T) = T(anchor\ word)$, то есть тему якорного слова. Вторым регуляризационным параметром выступает ширина окна. Таким образом, каждый документ рассматривается как гранулированная поверхность, состоящая из гранул (тем). Пример гранулированного текста приведен на рисунке (70). Так как изначально комбинация слов, часто встречающихся внутри одной гранулы не известна, то гранулированный вариант сэмплирования формирует статистическую связь между близко расположенными словами. Результаты вычисления стабильности тематической модели с прямоугольным ядром приведены в таблице 5 (модель GLDA).

5.1.2. Ядро Епанечникова (Epanechnikov kermel, ELDA)

Ядро Епанечникова является симметричной функцией, которая показывает, что темы слов внутри заданного окна распределены следующим образом:

$$K(w) = T(anchor\ word) \cdot (1 - r^2), \quad (13),$$

где $r = 1$ соответствует крайнему правому слову в окне, $r = -1$ – крайнему левому слову в окне. Это значит, что чем дальше слово от якорного слова, тем сильнее тема слова отличается от темы центрального слова (отличие идет в сторону уменьшения номера темы). Результаты расчета стабильности тематической модели с ядром Епанечникова приведены в таблице 5 (модель ELDA).

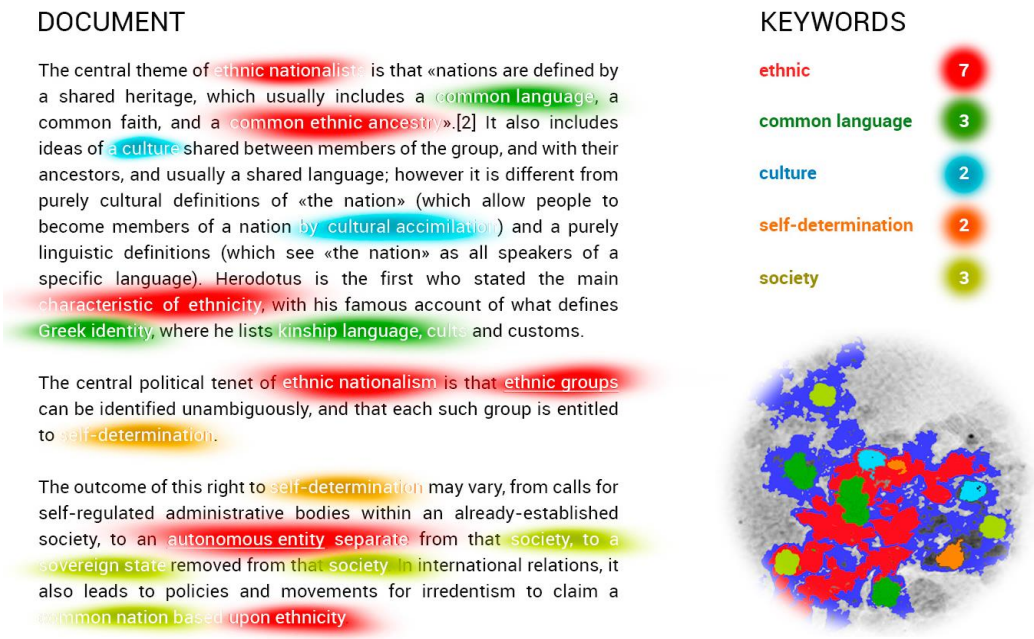

Рис. 70. Пример гранулированной поверхности в физике и текста, представленного в виде гранулированной поверхности.

5.1.3. Треугольное ядро Епанечникова (triangle kernel, TLDA)

В данном случае функция локальной плотности задается в виде треугольника:

$$K(w) = T(anchor\ word) \cdot (1 - |r|) \quad 14.$$

Данная модель почти аналогична модели ELDA, с той разницей, что падение номера темы для слов по краям окна происходит быстрее, чем в модели ELDA. Результаты стабильности тематической модели с прямоугольным ядром приведены в таблице 5 (модель TLDA).

5.2. Исследование стабильности моделей ARTM, GLDA, pLSA, SLDA

В целом, в рамках тематического моделирования, доминирующее положение занимают модели, основанные именно на LDA. Однако добавление регуляризаторов в модели LDA сопряжено со сложностью байесовского вывода, что, в свою очередь, затрудняет построение многоцелевых тематических моделей, удовлетворяющих одновременно большому числу регуляризаторов. В работе [16] авторы предлагают альтернативу байесовскому подходу – аддитивную регуляризацию тематических моделей, ARTM. Она имеет несколько принципиальных отличий от байесовского подхода.

В этом случае построение многоцелевых тематических моделей существенно упрощается благодаря аддитивности регуляризаторов. Добавление регуляризатора требует небольшой модификации М-шага в готовом Е-М-подобном алгоритме. В рамках данной работы проводилось исследование стабильности модели ARTM с двумя регуляризаторами: 1. Регуляризатор разреживания матрицы Φ. 2. Регуляризатор разреживания матрицы Θ.

5.2.1. Эксперименты по определению стабильности тематических моделей

Проводилось исследование 9 тематических моделей: 1. pLSA. 2. ARTM sparse Ф. 3. ARTM sparse Θ. 4. VLDA, 5. LDA GB. 6. Semi-supervised LDA GB. 7. GLDA. 8. ELDA, 9. TLDA). В тестировании моделей были использованы документы из социальной сети «Живой Журнал», общим количеством 101481 документ. В каждом моделировании было использовано 200 тем. Результаты вычислений моделей приведены в таблице (7). Каждая модель запускалась по три раза при вариации числа тем. Стабильность тематических моделей рассчитывалась по трем запускам при помощи дивергенции Кульбака – Лейблера.

Таблица 7.

Тематическая модель	Число стабильных тем	Средняя величина расстояния Жаккара
PLSA	54	0.47
PLSA+регуляризатор разреживания матрицы φ(w,t), коэффициент регуляризации α=0.5	9	0.44
PLSA+регуляризатор разреживания матрицы Θ(t,d), коэффициент регуляризации β=0.2	87	0.47
Variational Latent Dirichlet Allocation (VLDA)	111	0.53

LDA (Gibbs sampling)	77	0.56
SLDA (Gibbs sampling)	84	0.62
GLDA (размер гранулы ±1)	195	0.64
GLDA (размер гранулы ±2)	195	0.71
GLDA (размер гранулы ±3)	197	0.73
ELDA (размер гранулы ±1)	184	0.23
ELDA (размер гранулы ±2)	192	0.33
ELDA (размер гранулы ±3)	199	0.20
TLDA (размер гранулы ±1)	162	0.63
TLDA (размер гранулы ±2)	200	0.3
TLDA (размер гранулы ±3)	200	0.68

Вычисления показывает, что задание регуляризаторов может как увеличивать стабильность, так и уменьшать ее. При этом добавление информации о локальной связи между словами может существенно улучшать стабильность тематической модели. Предложенный вариант процедуры сэмплирования работает существенно стабильнее, чем такие модели как pLSA, VLDA, ARTN.

Заключение.

Как уже было отмечено, в тематическом моделировании есть три основные проблемы: 1. Определение количества компонент в смеси распределений, включая определение наличие плоских и иерархических структур в датасетах. 2. Проблема настройки гиперпараметров и коэффициентов регуляризации. 3. Проблема стабильности (воспроизводимости тематического решения). Соответственно, в рамках данного диссертационного исследования были предложены способы решения этих проблем.

Во-первых, реализована энтропийная тематическая модель (на основе энтропии Реньи) определения оптимального количества распределений в смеси для генеративных тематических моделей. Данная модель позволяет оценить количество тем в датасетах на европейских языках, а также определить оптимальные гиперпараметры тематических моделей. Во-вторых, реализована иерархическая энтропийная тематическая модель, которая позволяет оценить количество уровней иерархии, а также определить тип иерархии в датасете. При этом, количество минимумов энтропии Реньи соответствует количеству уровней иерархии в датасетах. В-третьих, реализована энтропийная тематическая модель на основе двухпараметрической энтропии Шарма−Миттала (Sharma–Mittal Entropy). В данной модели один параметр энтропии выражен в терминах обратной величины числа тем, второй параметр – через расстояние Жаккара, что позволяет учесть семантическое сходство между тематическими решениями при вариации размера смеси распределений. В-

четвертых, реализована фрактальная модель оценки работы генеративных тематических моделей. Данная модель позволяет выделить линейные области функции плотности распределения слов и переходные области, соответствующие минимумам энтропии Реньи. В-пятых, реализован метод агрегации тематических моделей на основе процедуры ренормализации, который позволяет в сотни раз ускорить поиск оптимального размера смеси распределений в датасете. В данной части работы показано, что наиболее эффективной процедурой ренормализации, как с точки зрения поиска корректного количества тем, так и с точки зрения скорости расчета, является ренормализация на основе объединения тем с минимальными значениями энтропии Реньи. В-шестых, реализована гранулированная тематическая модель на основе процедуры сэмплирования Гиббса, в которой процедура регуляризации тематической модели задается с помощью локальной функции распределения тем. Предложенная модель демонстрирует высокий уровень стабильности в сравнении с другими тематическими моделями.

Хотя данная работа является законченной, предложенные модели могут быть использованы для дальнейшего развития области машинного обучения следующим образом: 1. Энтропии Реньи и Шарма−Миттала являются вариантами параметризованной функцией логарифма, где параметры существенно меняют его поведение. Исходя из этого, можно сформулировать класс математических моделей в области машинного обучения на основе поиска максимума параметризованного логарифма правдоподобия (одно- или двухпараметризованный варианты логарифма). 2. Процедура ренормализации может быть включена внутрь существующих алгоритмов тематического моделирования, что может существенно ускорить работу тематических моделей. 3. Принцип поиска минимума параметризованных энтропий можно использовать для оптимизации существующих алгоритмов кластеризации (включая иерархические процедуры кластеризации). 4. Принцип поиска минимума параметризованной энтропии можно использовать для определения оптимального количества слоев в нейронных сетях. Предварительные эксперименты на сетях ограниченных и глубоких машин Больцмана показывают, что поведение энтропии Реньи как функции от числа слоев в таких сетях аналогично поведению энтропии Реньи в тематических моделях как функции от числа тем. 5. Гранулированный вариант процедуры сэмплирования можно использовать для разработки тематических моделей, где процедура гранулированного сэмплирования будет учитывать не ближайшие слова в документе, а ближайшие слова по ‘word embedings’.

Финансирование

Исследования, вошедшие в диссертацию, а также подготовка диссертации были поддержаны грантами ВШЭ за 2014-2021:

- 2014 ТЗ-83 Социально-политические процессы в Интернете: структура и содержание социальных взаимодействий
- 2015 ТЗ-78 Междисциплинарные исследования интернета
- 2016 ТЗ-75 Интернет-пользование и интернет-пользователи: межстрановые и межрегиональные сравнения
- 2017 ТЗ-68 Интернет как социо-технический феномен
- 2018 ТЗ-67 Измерение социальных и текстовых свойств аккаунтов пользователей социальных сетей
- 2019 ТЗ-61 Социальные сети как социально-психологический и текстовый феномен
- 2020 ТЗ-57 Онлайн-коммуникация: когнитивные лимиты и методы автоматического анализа
- 2021 ТЗ-75 Моделирование поведения и социально-психологических характеристик индивидов на основе мультимодальных цифровых следов

Список литературы

1. Chauhan, Uttam and Apurva Shah. “Topic Modeling Using Latent Dirichlet allocation.” *ACM Computing Surveys (CSUR)* 54 (2022): 1 - 35.
2. Koltsov, Sergei, Vera Ignatenko, Maxim Terpilovskii and Paolo Rosso. “Analysis and tuning of hierarchical topic models based on Renyi entropy approach.” *PeerJ Computer Science* 7 (2021): n. pag.
3. Catherine, Ş A. and Ugar. “Finding the number of clusters in a data set : An information theoretic approach C.” (2003).
4. Stephens, Greg J., Thierry Mora, Gašper Tkačik and William Bialek. “Statistical thermodynamics of natural images.” *Physical review letters* 110 1 (2013): 018701.
5. Mirkin, Boris G.. “Clustering for data mining - a data recovery approach.” *Computer science and data analysis series* (2005).
6. Tibshirani, Robert, Guenther Walther and Trevor J. Hastie. “Estimating the number of clusters in a data set via the gap statistic.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2000): n. pag.
7. Fujita, André, Daniel Y. Takahashi and Alexandre Galvão Patriota. “A non-parametric method to estimate the number of clusters.” *Comput. Stat. Data Anal.* 73 (2014): 27-39.
8. Aldana-Bobadilla, Edwin and Ángel Fernando Kuri Morales. “A Clustering Method Based on the Maximum Entropy Principle.” *Entropy* 17 (2015): 151-180.
9. Ramírez-Reyes, Abdiel, Alejandro Raúl Hernández-Montoya, Gerardo Herrera-Corral and Ismael Domínguez-Jiménez. “Determining the Entropic Index q of Tsallis Entropy in Images through Redundancy.” *Entropy* 18 (2016): 299.
10. Milligan, Glenn W. and Martha Cooper. “An examination of procedures for determining the number of clusters in a data set.” *Psychometrika* 50 (1985): 159-179.
11. SH Cha, Taxonomy of nominal type histogram distance measures, Proceedings of the American conference on applied mathematics, 325-330, 2008.
12. Rose, Gurewitz and Fox. “Statistical mechanics and phase transitions in clustering.” *Physical review letters* 65 8 (1990): 945-948.
13. Hofmann, Thomas. “Probabilistic Latent Semantic Indexing.” *ACM SIGIR Forum* 51 (2017): 211 - 218.
14. Blei, David M., A. Ng and Michael I. Jordan. “Latent Dirichlet Allocation.” *J. Mach. Learn. Res.* 3 (2003): 993-1022.
15. Griffiths, Thomas L. and Mark Steyvers. “Finding scientific topics.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004): 5228 - 5235.
16. Vorontsov, Konstantin V., Anna Potapenko and Alexander Plavin. “Additive Regularization of Topic Models for Topic Selection and Sparse Factorization.” *SLDS* (2015).
17. Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. “Hierarchical Dirichlet Processes.” *Journal of the American Statistical Association* 101 (2006): 1566 - 1581.
18. Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. “Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes.” *NIPS* (2004).
19. Mimno, David, Wei Li and Andrew McCallum. “Mixtures of hierarchical topics with Pachinko allocation.” *ICML '07* (2007).
20. Belyy A. V., Seleznova M. S., Sholokhov A. K., Vorontsov K. V. Quality Evaluation and Improvement for Hierarchical Topic Modeling, Computational Linguistics and Intellectual Technologies. Dialogue 2018. pp. 110-123
21. Dieng, Adji B., Francisco J. R. Ruiz and David M. Blei. “Topic Modeling in Embedding Spaces.” *Transactions of the Association for Computational Linguistics* 8 (2020): 439-453.
22. Miao, Yishu, Edward Grefenstette and Phil Blunsom. “Discovering Discrete Latent Topics with Neural Variational Inference.” *ArXiv* abs/1706.00359 (2017): n. pag.

23. Daud, Ali, Juan-Zi Li, Lizhu Zhou and Faqir Muhammad. “Knowledge discovery through directed probabilistic topic models: a survey.” *Frontiers of Computer Science in China* 4 (2009): 280-301.
24. Asuncion, Arthur U., Max Welling, Padhraic Smyth and Yee Whye Teh. “On Smoothing and Inference for Topic Models.” *UAI* (2009).
25. Cao, Juan, Tian Xia, Jintao Li, Yongdong Zhang and Sheng Tang. “A density-based method for adaptive LDA model selection.” *Neurocomputing* 72 (2009): 1775-1781.
26. Arun, R., V. Suresh, C. E. Veni Madhavan and M. Narasimha Murty. “On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.” *PAKDD* (2010).
27. Roberts, Margaret E., Brandon M Stewart and Dustin Tingley. “Navigating the Local Modes of Big Data: The Case of Topic Models.” *Computational Social Science* (2016).
28. Koltsov, Sergei, Sergei I. Nikolenko, Olessia Koltsova and Svetlana Bodrunova. “Stable topic modeling for web science: granulated LDA.” *Proceedings of the 8th ACM Conference on Web Science* (2016): n. pag.
29. Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov and David Mimno. “Evaluation methods for topic models.” *ICML '09* (2009).
30. Foulds, James R. and Padhraic Smyth. “Annealing Paths for the Evaluation of Topic Models.” *UAI* (2014).
31. Zhu, Jun, Amr Ahmed and Eric P. Xing. “MedLDA: maximum margin supervised topic models.” *J. Mach. Learn. Res.* 13 (2012): 2237-2278.
32. Sristy, Nagesh Bhattu and Durvasula V. L. N. Somayajulu. “Entropy Regularization for Topic Modelling.” *I-CARE 2014* (2014).
33. Kadanoff, Leo P.. “Statistical Physics: Statics, Dynamics and Renormalization.” (2000).
34. Wilson, Kenneth G.. “Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture.” *Physical Review B* 4 (1971): 3174-3183.
35. Olemskoi, A. Synergetics of Complex Systems: Phenomenology and Statistical Theory, KRASAND Publ. House, Moscow, 2009, 384 p. (in Russian).
36. Carpinteri, Alberto, Bernardino Chiaia and Giuseppe Andrea Ferro. “Size effects on nominal tensile strength of concrete structures: multifractality of material ligaments and dimensional transition from order to disorder.” *Materials and Structures* 28 (1995): 311-317.
37. Essam, John W.. “Potts models, percolation, and duality.” *Journal of Mathematical Physics* 20 (1979): 1769-1773.
38. Tikhonov, A.N. and Arsenin, V.Y. Solutions of Ill-Posed Problems. Winston, New York, (1977).
39. Belford, Mark, Brian Mac Namee and Derek Greene. “Stability of topic modeling via matrix factorization.” *Expert Syst. Appl.* 91 (2018): 159-169.
40. Greene, Derek, Derek O'Callaghan and Pádraig Cunningham. “How Many Topics? Stability Analysis for Topic Models.” *ECML/PKDD* (2014).
41. De Waal, A., Barnard, E.: Evaluating topic models with stability. In: 19th Annual Symposium of the Pattern Recognition Association of South Africa (2008).
42. Koltsov, Sergei, Sergei I. Nikolenko, Olessia Koltsova, Vladimir Filippov and Svetlana Bodrunova. “Stable Topic Modeling with Local Density Regularization.” *INSCI* (2016). *Lecture Notes in Computer Science series* Vol. 9934. Switzerland : Springer, (2016)
43. Derbanosov, R. Stability of topic modeling via modality regularization [Текст] / R. Derbanosov, M. Bakhanova // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue. (2020)
44. Agrawal, Amritanshu, Wei Fu and Tim Menzies. “What is wrong with topic modeling? And how to fix it using search-based software engineering.” *Inf. Softw. Technol.* 98 (2018): 74-88.

45. Koltcov, Sergei. “A thermodynamic approach to selecting a number of clusters based on topic modeling.” *Technical Physics Letters* 43 (2017): 584-586.
46. Koltcov, Sergei. “Application of Rényi and Tsallis entropies to topic modeling optimization.” *Physica A: Statistical Mechanics and its Applications* (2018): n. pag.
47. Tsallis, Constantino. “Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World.” (2009).
48. Beck, Christian. “Generalised information and entropy measures in physics.” *Contemporary Physics* 50 (2009): 495 - 510.
49. Mora, Thierry and Aleksandra M. Walczak. “Rényi entropy, abundance distribution, and the equivalence of ensembles.” *Physical review. E* 93 5 (2016): 052418.
50. Beck, Christian and Friedrich Schögl. “Thermodynamics of chaotic systems.” (1993).
51. Klimontovich, Yu. L. Statistical Theory of Open Systems (Yanus, Moscow, 1995; Springer, Dordrecht, 1995).
52. Sharma, Bhu Dev and Asha Garg. “Nonadditive Measures of Average Charge for Heterogeneous Questionnaires.” *Inf. Control.* 41 (1979): 232-242.
53. Nielsen, Frank and Richard Nock. “A closed-form expression for the Sharma–Mittal entropy of exponential families.” *Journal of Physics A: Mathematical and Theoretical* 45 (2011): n. pag.
54. Jaccard, P.. “The distribution of the flora in the alpine zone 1.” *New Phytologist* 11: 37-50.
55. Parker, Austin J., Kelly B. Yancey and Matthew P. Yancey. “Regular Language Distance and Entropy.” *MFCS* (2017).
56. Koltsov, Sergei, Vera Ignatenko and Olessia Koltsova. “Estimating Topic Modeling Performance with Sharma–Mittal Entropy.” *Entropy* 21 (2019): n. pag.
57. Koltsov, Sergei, Vera Ignatenko and Sergei Pashakhin. “How Many Clusters? An Entropic Approach to Hierarchical Cluster Analysis.” *SAI* (2020).
58. News Dataset from Usenet. Available online: http://qwone.com/~jason/20Newsgroups/ (accessed on 31 October 2019).
59. Basu, Sugato, Ian Davidson and Kiri L. Wagstaff. “Constrained Clustering: Advances in Algorithms, Theory, and Applications.” (2008).
60. Lesche, Bernhard. “Instabilities of Rényi entropies.” *Journal of Statistical Physics* 27 (1982): 419-422.
61. Blei, David M., Thomas L. Griffiths, Michael I. Jordan and Joshua B. Tenenbaum. “Hierarchical Topic Models and the Nested Chinese Restaurant Process.” *NIPS* (2003).
62. Mimno, David, Wei Li and Andrew McCallum. “Mixtures of hierarchical topics with Pachinko allocation.” *ICML '07* (2007).
63. Chirkova, Nadezhda. “Additive Regularization for Hierarchical Multimodal Topic Modeling.”. Machine Learning and Data Analysis, 2:187–200. (2016)
64. Ignatenko, Vera, Sergei Koltsov, Steffen Staab and Zeyd Boukhers. “Fractal approach for determining the optimal number of topics in the field of topic modeling.” *Journal of Physics: Conference Series* (2019): n. pag.
65. Schroeder, Manfred. “Fractals, Chaos, Power Laws: Minutes From an Infinite Paradise.” (1991).
66. Koltcov, Sergei and Vera Ignatenko. “Renormalization Analysis of Topic Models.” *Entropy* 22 (2020): n. pag.
67. Mimno, David, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders and Andrew McCallum. “Optimizing Semantic Coherence in Topic Models.” *EMNLP* (2011).
68. Koltsov, Sergei and Vera Ignatenko. 2020. "Renormalization Analysis of Topic Models" *Entropy* 22, no. 5: 556. https://doi.org/10.3390/e22050556.
69. Rosenblatt, Murray. “Remarks on Some Nonparametric Estimates of a Density Function.” *Annals of Mathematical Statistics* 27 (1956): 832-837.
70. Epanechnikov, V. A. Nonparametric estimation of multidimensional probability density. Theory Probab. Appl. 14, 153–158, 1973.

71. Koltsov, Sergei, Sergei I. Nikolenko and E. Y. Koltsova. “Gibbs sampler optimization for analysis of a granulated medium.” *Technical Physics Letters* 42 (2016): 837-839.
72. Newman, David, Edwin V. Bonilla and Wray L. Buntine. “Improving Topic Coherence with Regularized Topic Models.” *NIPS* (2011).
73. Andrzejewski, David and Xiaojin Zhu. “Latent Dirichlet Allocation with Topic-in-Set Knowledge.” *HLT-NAACL 2009* (2009).
74. Tsallis, Constantino and Daniel A. Stariolo. “Generalized simulated annealing.” *Physica A-statistical Mechanics and Its Applications* 233 (1996): 395-406.