Topic Modeling with LDA (Latent Dirichlet Allocation)

https://quantdev.ssri.psu.edu/sites/qdev/files/topic\_modeling\_tutorial-Gutenberg-chapter\_as\_document.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9120935/

Egger R., Yu J. A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts //Frontiers in Sociology. – 2022. – Т. 7.

Богатство данных социальных сетей открыло новые возможности для исследований в области социальных наук, позволяющих получить представление о человеческом поведении и опыте. В частности, новые подходы, основанные на данных и опирающиеся на тематические модели, открывают совершенно новые перспективы интерпретации социальных явлений.

ля эффективного извлечения функций из большого массива текстовых данных были введены многочисленные подходы к интеллектуальному анализу текста (Li Q., Li S., Zhang S., Hu J., Hu J. (2019). A review of text corpus-based tourism big data mining. Appl. Sci. 9:3300. 10.3390/app9163300), среди которых тематическое моделирование служит наиболее часто применяемой техникой (Hong L., Davison B. D. (2010). Empirical study of topic modeling in Twitter. Proc. First Workshop Soc. Media Analyt. 2010, 80-88. 10.1145/1964858.1964870). В двух словах, модель темы - это форма статистического моделирования, используемая в машинном обучении и НЛП, как обсуждалось ранее, которая идентифицирует скрытые тематические шаблоны в коллекции текстов (Guo Y., Barnes S. J., Jia Q. (2017). Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. Tour. Manag. 59, 467–483. 10.1016/j.tourman.2016.09.009). Те, которые рассматриваются как наиболее устоявшиеся методы, включают LDA, латентный семантический анализ (LSA) и вероятностный LSA (Albalawi R., Yeap T. H., Benyoucef M. Using topic modeling methods for short-text data: A comparative analysis //Frontiers in Artificial Intelligence. – 2020. – Т. 3. – С. 42.). Однако совсем недавно недавно разработанные алгоритмы, такие как NMF, Corex, Top2Vec и BERTopic, также привлекли и продолжают привлекать все большее внимание исследователей (Obadimu A., Mead E., Agarwal N. (2019). “Identifying latent toxic features on YouTube using non-negative matrix factorization,” in The Ninth International Conference on Social Media Technologies, Communication, and Informatics (Valencia: ), 1–6.; Sánchez-Franco M. J., Rey-Moreno M. (2022). Do travelers' reviews depend on the destination? An analysis in coastal and urban peer-to-peer lodgings. Psychol. Market. 39, 441–459. 10.1002/mar.21608). В социальных науках тематические модели ранее применялись, например, для выявления неявных предпочтений потребителей (Vu H. Q., Li G., Law R. Discovering implicit activity preferences in travel itineraries by topic modeling //Tourism Management. – 2019. – Т. 75. – С. 435-446.

; Egger R., Pagiri A., Prodinger B., Liu R., Wettinger F. (2022). “Topic modelling of tourist dining experiences based on the GLOBE model,” in ENTER22 e-Tourism Conference (Berlin: Springer; ), 356–368. 10.1007/978-3-030-94751-4\_32 [), выявления семантических структур в Instagram (Egger R., Pagiri A., Prodinger B., Liu R., Wettinger F. (2022). “Topic modelling of tourist dining experiences based on the GLOBE model,” in ENTER22 e-Tourism Conference (Berlin: Springer; ), 356–368. 10.1007/978-3-030-94751-4\_32

Egger R., Yu J. Identifying hidden semantic structures in Instagram data: a topic modelling comparison //Tourism Review. – 2021.) и улучшения систем рекомендаций (Shafqat W., Byun Y. C. A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis //Sustainability. – 2019. – Т. 12. – №. 1. – С. 320.). Несмотря на надежность алгоритмов моделирования темы, существующая литература опирается в основном на одну единственную модель, причем LDA является доминирующим методом (Gallagher R. J. et al. Anchored correlation explanation: Topic modeling with minimal domain knowledge //Transactions of the Association for Computational Linguistics. – 2017. – Т. 5. – С. 529-542.) и обычно рассматривается как стандартный подход.

Porturas T., Taylor R. A. Forty years of emergency medicine research: Uncovering research themes and trends through topic modeling //The American Journal of Emergency Medicine. – 2021. – Т. 45. – С. 213-220.

Определение темы может облегчить сбор знаний, выявить тематические взаимосвязи, тенденции и предсказать будущее направление. Мы стремились определить с помощью неконтролируемого подхода машинного обучения к тематическому моделированию наиболее распространенные темы исследований в области неотложной медицины за последние 40 лет и обобщить их тенденции и характеристики.

Chen X. et al. A decade of sentic computing: topic modeling and bibliometric analysis //Cognitive Computation. – 2022. – Т. 14. – №. 1. – С. 24-47.

Наше исследование дает полный обзор sentic computing, выявляет основные проблемы среди ученых в течение последнего десятилетия и предлагает понимание будущих направлений исследований sentic computing.

Xie Y., Ning C., Sun L. The twenty-first century of structural engineering research: A topic modeling approach //Structures. – Elsevier, 2022. – Т. 35. – С. 577-590.

Стремясь раскрыть общий исследовательский ландшафт строительной инженерии в XXI веке, это исследование применяет скрытое распределение Дирихле (LDA), подход к тематическому моделированию, для анализа 51 346 тезисов статей из 23 престижных журналов по строительной инженерии с периодом публикации с 2000 по 2020 год.

Kim M., Kim D. A Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results //Applied Sciences. – 2022. – Т. 12. – №. 6. – С. 3118.

для индексации результатов академических исследований

Abba M. et al. One Hundred Years of Hypertension Research: Topic Modeling Study //JMIR Formative Research. – 2022. – Т. 6. – №. 5. – С. e31292.

Kumar V., Srivastava A. Trends in the thematic landscape of corporate social responsibility research: A structural topic modeling approach //Journal of Business Research. – 2022. – Т. 150. – С. 26-37.

Kwon H. J. et al. Topic modeling and sentiment analysis of online review for airlines //Information. – 2021. – Т. 12. – №. 2. – С. 78.

Kirilenko A. P., Stepchenkova S. O., Dai X. Automated topic modeling of tourist reviews: does the Anna Karenina principle apply? //Tourism Management. – 2021. – Т. 83. – С. 104241.

Параллельно с растущими дебатами по поводу устойчивой энергетики и решений в области искусственного интеллекта, мир в настоящее время обсуждает этику искусственного интеллекта и его возможные негативные последствия для общества и окружающей среды.

Shang Z., Luo J. M. Topic modeling for hiking trail online reviews: Analysis of the Mutianyu Great Wall //Sustainability. – 2022. – Т. 14. – №. 6. – С. 3246.

Doogan C., Buntine W. Topic model or topic twaddle? re-evaluating semantic interpretability measures //Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2021. – С. 3824-3848.

Gregoriades A. et al. Supporting digital content marketing and messaging through topic modelling and decision trees //Expert systems with applications. – 2021. – Т. 184. – С. 115546.

Анализ и поддержка маркетинговых компаний

Lim M. K., Li Y., Song X. Exploring customer satisfaction in cold chain logistics using a text mining approach // Industrial Management & Data Systems, Vol. 121 No. 12, pp. 2426-2449. https://doi.org/10.1108/IMDS-05-2021-0283

Lim, M.K., Li, Y. and Song, X. (2021), "Exploring customer satisfaction in cold chain logistics using a text mining approach", Industrial Management & Data Systems, Vol. 121 No. 12, pp. 2426-2449. https://doi.org/10.1108/IMDS-05-2021-0283

Automated Content Analysis with R

КОД

https://content-analysis-with-r.com/6-topic\_models.html

Two important parameters exist in topic discovery and LDA: alpha and beta, also known as hyperparameters. In many studies, the default value of the alpha is 50/number of topics, and the beta is set to 0.01 [[**71**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B71-sustainability-14-03246),**[72](https://www.mdpi.com/2071-1050/14/6/3246/htm" \l "B72-sustainability-14-03246" \o ")**,[**73**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B73-sustainability-14-03246)]. However, some studies also adjusted the parameters to obtain more interpretable topics. For example, Jacobi, van Atteveldt and Welbers [[**71**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B71-sustainability-14-03246)] compared the performance of LDA when the alpha was set to 50/number of topics and five/number of topics, and found that the latter alpha value can yield better outcomes. We adopted the mallet library’s hyperparameter optimization option to set a proper value in the current study. Given these optimal hyperparameters, the LDA analysis process and the topic coherence calculation are repeated 10 times for each topic. The final topic coherence curves of the English and Chinese datasets are shown in [**Figure 1**](https://www.mdpi.com/2071-1050/14/6/3246/htm#fig_body_display_sustainability-14-03246-f001). According to the topic coherence curves, the improvements in the topic coherence of the English and Chinese datasets tend to slow down when the number of topics is five. Therefore, we qualitatively interpreted the five topics in the corresponding dataset. Then, we gradually increased the number of topics to find the proper number of topics. Finally, the optimal number of topics was set at 10 for the English dataset and nine for the Chinese dataset. This study adopts the same methodology of naming topics as that used in Guo, Barnes and Jia [[**22**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B22-sustainability-14-03246)] as well as Taecharungroj and Mathayomchan [[**74**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B74-sustainability-14-03246)]. A researcher identified a topic according to the logical connection between the top 15 words in each topic–word distribution, and another researcher confirmed and refined the topic [[**22**](https://www.mdpi.com/2071-1050/14/6/3246/htm#B22-sustainability-14-03246),**[74](https://www.mdpi.com/2071-1050/14/6/3246/htm" \l "B74-sustainability-14-03246" \o ")**]. [**Table 4**](https://www.mdpi.com/2071-1050/14/6/3246/htm#table_body_display_sustainability-14-03246-t004) and [**Table 5**](https://www.mdpi.com/2071-1050/14/6/3246/htm#table_body_display_sustainability-14-03246-t005) provide the final topic names and top 15 words in each topic. For interpretation, we further grouped these topics (see [**Table 6**](https://www.mdpi.com/2071-1050/14/6/3246/htm#table_body_display_sustainability-14-03246-t006)) based on the summarized attributes as described in the “Hiking Experience” section and [**Table 1**](https://www.mdpi.com/2071-1050/14/6/3246/htm#table_body_display_sustainability-14-03246-t001).

В обнаружении темы и LDA существуют два важных параметра: альфа и бета, также известные как гиперпараметры. Во многих исследованиях значение альфа по умолчанию равно 50/количество тем, а бета установлено на 0,01 [71,72,73 Jacobi C., Van Atteveldt W., Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling //Digital journalism. – 2016. – Т. 4. – №. 1. – С. 89-106. Lin C., He Y. Joint sentiment/topic model for sentiment analysis //Proceedings of the 18th ACM conference on Information and knowledge management. – 2009. – С. 375-384.; Ren G., Hong T. Investigating online destination images using a topic-based sentiment analysis approach //Sustainability. – 2017. – Т. 9. – №. 10. – С. 1765. ]. Тем не менее, некоторые исследования также скорректировали параметры, чтобы получить более интерпретируемые темы. Например, Jacobi, van Atteveldt и Welbers [Jacobi C., Van Atteveldt W., Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling //Digital journalism. – 2016. – Т. 4. – №. 1. – С. 89-106.] сравнили эффективность LDA, когда альфа была установлена на 50/количество тем и пять/количество тем, и обнаружили, что последнее значение альфа может дать лучшие результаты.

Мы приняли опцию оптимизации гиперпараметров библиотеки mallet, чтобы установить правильное значение в текущем исследовании. Учитывая эти оптимальные гиперпараметры, процесс анализа LDA и расчет согласованности темы повторяются 10 раз для каждой темы. Окончательные кривые согласованности тем для наборов данных на английском и китайском языках показаны на рис. 1. Согласно кривым согласованности тем улучшения согласованности тем для наборов данных на английском и китайском языках имеют тенденцию к замедлению, когда число тем равно пяти. Поэтому мы качественно интерпретировали пять тем в соответствующем наборе данных. Затем мы постепенно увеличивали количество тем, чтобы найти нужное количество тем. Наконец, оптимальное количество тем было установлено на уровне 10 для набора данных на английском языке и девяти для набора данных на китайском языке. В этом исследовании используется та же методология именования тем, что и у Го, Барнса и Цзя [22], а также у Таечарунгроя и Матхайомчана [74]. Исследователь определил тему в соответствии с логической связью между 15 верхними словами в каждом распределении тема-слов, а другой исследователь подтвердил и уточнил тему [22,74]. В Таблице 4 и Таблице 5 представлены окончательные названия тем и 15 лучших слов в каждой теме. Для интерпретации мы дополнительно сгруппировали эти темы (см. Таблицу 6) на основе обобщенных характеристик, как описано в разделе «Пешеходный опыт» и в Таблице 1.

Тематическое моделирование относится к ряду алгоритмов обработки естественного языка, которые дают нам представление о «скрытых» семантических темах или шаблонах в коллекции документов. Эти модели совпадения слов используются для определения скрытых «тем», присутствующих в корпусе. Тематическое моделирование успешно используется для поиска информации, классификации документов, их обобщения и исследовательского анализа больших корпусов текстов. В этом обзоре рассматриваются различные алгоритмы, которые использовались для тематического моделирования с течением времени, включая TF-IDF, скрытый алгоритм Дирихле (LDA), кластеризацию на вложениях BERT на уровне предложений и новый гибридный подход к созданию контекстных тем с использованием комбинации векторов LDA и BERT. . В этом обзоре будут проанализированы преимущества и ограничения этих алгоритмов.

# Latent Dirichlet Allocation and t-Distributed Stochastic Neighbor Embedding Enhance Scientific Reading Comprehension of Articles Related to Enterprise Architecture

https://www.mdpi.com/2673-2688/2/2/11/htm

Моделирование в основном зависит от количества тем, которые должны быть определены [[**39**](https://www.mdpi.com/2673-2688/2/2/11/htm#B39-ai-02-00011)]. Чтобы определить идеальное количество тем, эмпирические исследования часто полагаются на измерение логарифмического правдоподобия [[**40,41**](https://www.mdpi.com/2673-2688/2/2/11/htm#B40-ai-02-00011)].

Для валидации результатов, определенных алгоритмом LDA, применяется кросс-валидация [[**42**](https://www.mdpi.com/2673-2688/2/2/11/htm#B42-ai-02-00011)]. В рамках этой процедуры корпус обучающего текста делится на две части. Первая и обычно более обширная часть текста используется для обучения, а вторая - для тестирования. Тестовый корпус *w* состоит из документов, которые не были рассмотрены. Обучающая модель описывается как матрица тем с Φ как распределение тем-слов. Параметр Θ не учитывается, поскольку он представляет распределение документа по темам обучающего набора и поэтому непригоден для оценки.

Другой возможностью для оценки наилучшего количества тем является R-фреймворк ‘ldatuning’, разработанный Мурзинцевым [[**47**](https://www.mdpi.com/2673-2688/2/2/11/htm#B47-ai-02-00011)]. Фреймворк использует несколько метрик для измерения. Мы используем три из них для нашего исследования. Griffiths 2004 [[**48**](https://www.mdpi.com/2673-2688/2/2/11/htm#B48-ai-02-00011)] представляет собой подход, при котором количество тем является оптимальным, когда логарифмическое правдоподобие для данных становится максимальным. CaoJuan 2009 [[**49**](https://www.mdpi.com/2673-2688/2/2/11/htm#B49-ai-02-00011)] метрика измеряет стабильность структуры темы, используя среднее косинусное расстояние между каждой парой тем. Arun 2010 [[**50**](https://www.mdpi.com/2673-2688/2/2/11/htm#B50-ai-02-00011)] находит оптимальное количество тем, применяя симметричную дивергенцию Кульбака-Лейблера к распределениям, генерируемым из матриц topic-word и document-topic, поскольку они рассматривали модель темы как матричную факторизацию [[**51**](https://www.mdpi.com/2673-2688/2/2/11/htm#B51-ai-02-00011)].

Все три показателя (см. [**рисунок 3**](https://www.mdpi.com/2673-2688/2/2/11/htm#fig_body_display_ai-02-00011-f003)) предполагают, что оптимальное количество тем составляет от 40 до 60. Низкое значение предпочтительно на верхнем графике, в то время как на нижнем графике предпочтительно высокое значение. Это подтверждает анализ недоумения, проведенный ранее. Таким образом, для этого исследования необходимо определить 50 тем.

The optimum number of topics for the application is between 40 and 100 iterations (see [**Figure 2**](https://www.mdpi.com/2673-2688/2/2/11/htm#fig_body_display_ai-02-00011-f002)). Another possibility for evaluating the best possible number of topics is the R-framework ‘ldatuning’, developed by Murzintcev [[**47**](https://www.mdpi.com/2673-2688/2/2/11/htm#B47-ai-02-00011)]. The framework uses several metrics for measurement. We use three of them for our study. Griffiths 2004 [[**48**](https://www.mdpi.com/2673-2688/2/2/11/htm#B48-ai-02-00011)] represents an approach where the number of topics is optimal when the log-likelihood for the data becomes maximum. CaoJuan 2009 [[**49**](https://www.mdpi.com/2673-2688/2/2/11/htm#B49-ai-02-00011)] metric measures the stability of topic structure using an average cosine distance between every pair of topics. Arun 2010 [[**50**](https://www.mdpi.com/2673-2688/2/2/11/htm#B50-ai-02-00011)] finds the optimal number of topics by applying a symmetric Kulback-Leibler divergence on the distributions generated from topic-word and document-topic matrices, as they viewed a topic model as matrix factorization [[**51**](https://www.mdpi.com/2673-2688/2/2/11/htm#B51-ai-02-00011)].

All three metrics (see [**Figure 3**](https://www.mdpi.com/2673-2688/2/2/11/htm#fig_body_display_ai-02-00011-f003)) suggest that the optimal number of topics is between 40 and 60. A low value is preferred in the upper graph, while in the lower graph, a high value is favored. This confirms the perplexity analysis carried out previously. Thus, for this study, 50 topics are to be determined.

#### **3.2. t-Distributed Stochastic Neighbor Embedding for Topic Model Visualization**

Алгоритм tSNE, введенный ван дер Маатеном и Хинтоном [[**56**](https://www.mdpi.com/2673-2688/2/2/11/htm#B56-ai-02-00011)], используется для решения проблемы представления многомерных данных и имеет широкий спектр применений, например, в науках о жизни, анализе сетей глубокого обучения [[**57,58,59**](https://www.mdpi.com/2673-2688/2/2/11/htm#B57-ai-02-00011)].

В общем, tSNE уменьшает размерность данных и создает 2D или 3D вложения, сохраняя локальные структуры в многомерных данных. Типичные задачи, выполняемые пользователями tSNE, основаны на выявлении связей между точками данных и их происхождением. Задачи часто включают идентификацию визуальных кластеров и их верификацию, например, с использованием параллельных координатных графиков [[**60**](https://www.mdpi.com/2673-2688/2/2/11/htm#B60-ai-02-00011)].

Алгоритм tSNE вычисляет два совместных распределения вероятностей *P* для представления общего расстояния между точками данных в многомерном пространстве и *Q*, описывающих сходство в маломерном пространстве. Целью этой логики является достижение верного представления *P* в низкоразмерном пространстве по *Q.* Это может быть достигнуто путем минимизации функции стоимости *C*, заданной дивергенцией Кулбака-Лейблера между совместными распределениями вероятностей *P* и *Q* для оптимизации позиций в низкоразмерном пространстве [[**60**](https://www.mdpi.com/2673-2688/2/2/11/htm#B60-ai-02-00011)]. Минимизация дивергенции Кульбака-Лейблера и изменение положения низкоразмерных точек для каждого шага градиентного спуска определяется как [[**61**](https://www.mdpi.com/2673-2688/2/2/11/htm#B61-ai-02-00011)]:

ПАРАМЕТРЫ И ОПИСАНИЕ

Quantitative analysis of large amounts of journalistic texts using topic modelling

<https://habr.com/ru/post/417167/>

Topic modeling и LDA.

Для начала, рассмотрим, что вообще делает LDA и в каких задачах используется.  
Наиболее часто LDA применяется для Topic Modeling(Тематическое моделирование) задач. Под такими задачами подразумеваются задачи кластеризации или классификации текстов — таким образом, что каждый класс или кластер содержит в себе тексты со схожими темами.

Для того, чтобы применять к датасету текстов(далее корпус текстов) LDA, необходимо преобразовать корпус в term-document matrix(Терм-документная матрица).

Терм-документная матрица — это матрица которая имеер размер N×W, где  
N — количество документов в корпусе, а W — размер словаря корпуса т.е. количество слов(уникальных) которые встречаются в нашем корпусе. В i-й строке, j-м столбце матрицы находится число — сколько раз в i-м тексте встретилось j-е слово.

LDA строит, для данной Терм-документной матрицы и T заранее заданого числа тем — два распределения:

1. Распределение тем по текстам.(на практике задается матрицей размера N×T)
2. Распределение слов по темам.(матрица размера T×W)

Значения ячеек данных матриц — это соответственно вероятности того, что данная тема содержится в данном документе(или доля темы в документе, если рассматривать документ как смесь разных тем) для матрицы 'Распределение тем по текстам'.

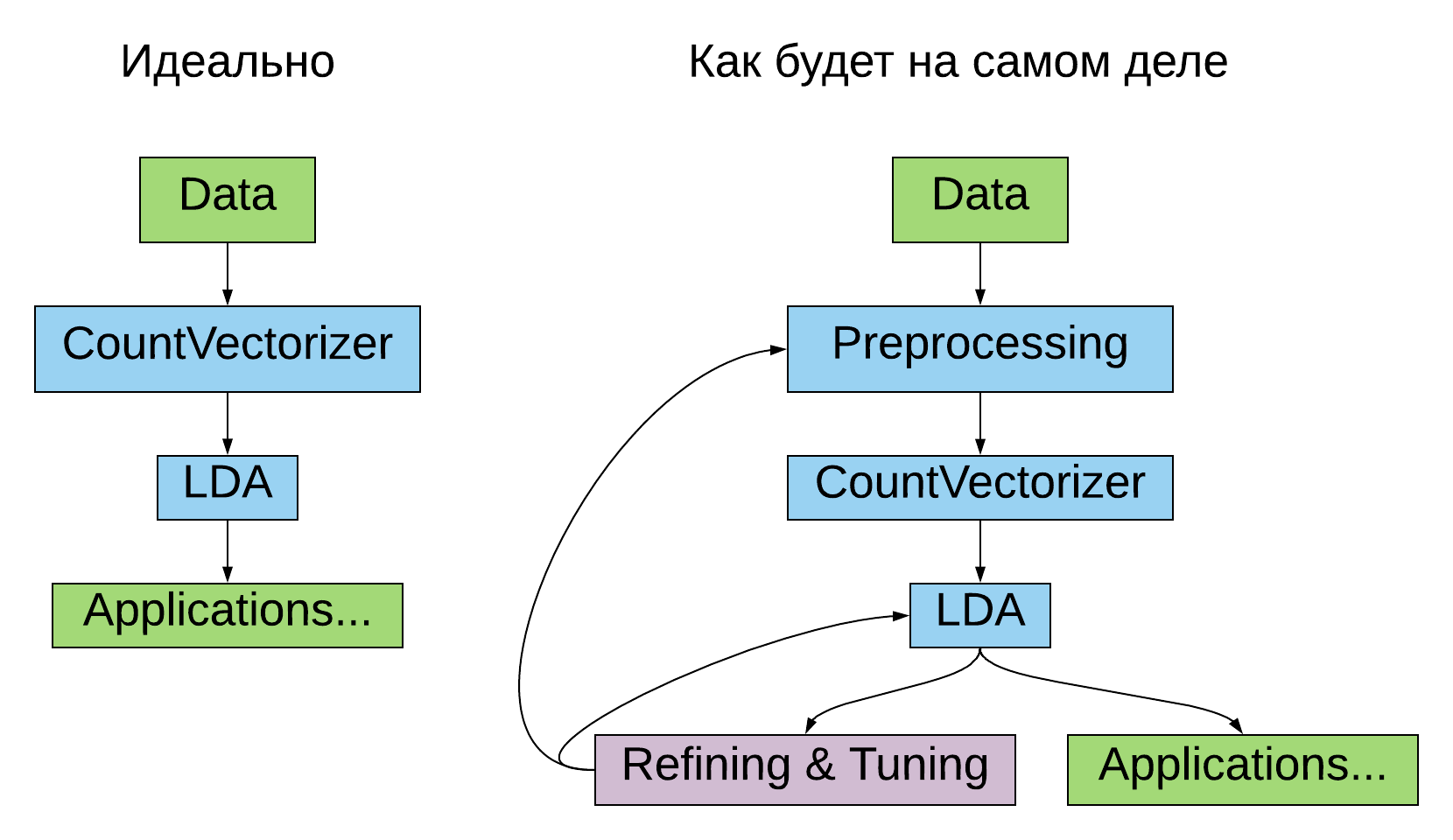
Для матрицы 'Распределение слов по темам' значения — это соотв-но вероятность встретить в тексте с темой i слово j, качествено, можно рассматривать эти числа как коэффициенты характеризующие, то насколько данное слово характерно для данной темы.

Следует сказать, что под словом тема понимается не 'житейское' определение этого слова. LDA выделяет T тем, но что это за темы и соответствуют ли они каким-либо известным темам текстов, как например: 'Спорт', 'Наука', 'Политика' — неизвестно. В данном случае, уместно скорее говорить о теме, как о некой абстрактной сущности, которая задается строкой в матрице распределения слов по темам и с некоторой вероятностью соответствует данному тексту, если угодно можно представить ее, как семейство характерных наборов слов встречающихся вместе, с соответствующими вероятностями(из таблицы) в некотором определенном множестве текстов.

#### **Итог**

С затравочным 'некоторым' пользователем имеющим ~150 друзей, удалось добыть 4679 текстов — каждый характеризует некоторое сообщество ВК. Тексты сильно варьируются по размеру и написаны на многих языках — часть из них не пригодна для наших целей, но об этом мы поговорим чуть дальше.

### **Основная часть**



Пройдемся по всем блокам нашего пайплайна — сначала, по обязательным(Идеально), затем по остальным — они, как раз и представляют наибольший интерес.

#### **CountVectorizer**

Перед тем, как учить LDA, нам необходимо представить наши документы в виде Терм-документной матрицы. Это обычно включает в себя такие операции как:

* Удаление путктуации/чисел/ненужных лексем.
* Токенизация(представление в виде списка слов)
* Подсчет слов, составление терм-документной матрицы.

Все эти действия в sklearn удобно реализованы в рамках одной программной сущности — sklearn.feature\_extraction.text.CountVectorizer.

[Ссылка на документацию](http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html)

Все, что нужно сделать это:

<https://habr.com/ru/company/surfingbird/blog/170081/>

#### **Предварительная обработка текстового контента (preprocessing)**

#### **Латентное распределение Дирихле (latent Dirichlet allocation, LDA)**

Метод LDA в нашем блоге уже достаточно подробно [описывал](http://habrahabr.ru/company/surfingbird/blog/150607/) Сергей Николенко. Алгоритм предназначен для описания текстов с точки зрения их тематик. Основное предположение модели LDA состоит в том, что каждый документ имеет несколько тематик, смешанных в некоторой пропорции. LDA — это вероятностная модель порождения текста, обучение которой позволяет выявить для каждого документа вероятностное распределение по тематикам, что в дальнейшем позволяет решать ряд прикладных задач, в том числе задачу рекомендаций.

<https://habr.com/ru/company/surfingbird/blog/230103/>

### **LDA: интуиция**

Итак, начинаем разговор об LDA. Напомню, что основная цель нашего обобщения – добиться того, чтобы документ не был жёстко привязан к одной теме, и слова в документах могли приходить из нескольких тем сразу. Что это значит в (математической) реальности? Это значит, что документ представляет собой *смесь* из тем, и каждое слово может быть порождено из одной из тем в этой смеси. Проще говоря, мы сначала кидаем кубик-документ, определяя тему *для каждого слова*, а потом вытаскиваем слово из соответствующего мешка.  
  
Таким образом, основная интуиция модели выглядит примерно так:

* на свете бывают *темы* (заранее неизвестные), которые отражают то, о чём могут быть части документа;
* каждая тема – это распределение вероятностей на словах, т.е. мешок слов, из которого можно с разной вероятностью вытащить разные слова;
* каждый документ – это смесь тем, т.е. распределение вероятностей на темах, кубик, который можно кинуть;
* процесс порождения каждого слова состоит в том, чтобы сначала выбрать тему по распределению, соответствующему документу, а затем выбрать слово из распределения, соответствующего этой теме.

Поскольку интуиция – это самое главное, и здесь она не самая простая, повторю ещё раз то же самое другими словами, теперь чуть более формально. Вероятностные модели удобно понимать и представлять в виде *порождающих процессов* (generative processes), когда мы последовательно описываем, как порождается одна единица данных, вводя по ходу дела все вероятностные предположения, которые мы в этой модели делаем. Соответственно, порождающий процесс для LDA должен последовательно описывать, как мы порождаем каждое слово каждого документа. И вот как это происходит (здесь и далее я буду предполагать, что длина каждого документа задана – её тоже можно добавить в модель, но обычно это ничего нового не даёт):

* для каждой темы *t*:
  + выбрать вектор *φt* (распределение слов в теме) по распределению https://habrastorage.org/getpro/habr/post_images/cc6/a38/198/cc6a38198d29b86e67226e2bfab267fa.png;
* для каждого документа *d*:
  + выбрать вектор https://habrastorage.org/getpro/habr/post_images/5fc/443/f81/5fc443f8187edc3579708df4823f14e4.png — вектор «степени выраженности» каждой темы в этом документе;
  + для каждого из слов документа *w*:
    - выбрать тему *zw* по распределению https://habrastorage.org/getpro/habr/post_images/a95/1fb/e9b/a951fbe9bfa79ced8a7416e0773953b0.png;
    - выбрать слово https://habrastorage.org/getpro/habr/post_images/28f/29c/b16/28f29cb160cf59fa50b9924ea2b77460.png с вероятностями, заданными в β.

<https://noduslabs.com/cases/tutorial-lda-text-mining-network-analysis/>

Моделирование тем используется для обнаружения тем, которые встречаются в теле документа или в текстовом корпусе. Латентное выделение дирихле (LDA) - это подход, используемый в тематическом моделировании на основе вероятностных векторов слов, которые указывают на их отношение к текстовому корпусу.

## Скрытое распределение Дирихле (LDA) - как это работает

Короче говоря, чтобы избежать сложной математики, мы рассмотрим, как LDA работает с некоторыми деталями, просто чтобы дать общую картину.

LDA, по своей сути, является итеративным алгоритмом, который идентифицирует набор тем, связанных с набором документов ( **[Blei 2003](http://www.jmlr.org/papers/v3/blei03a.html)**). Чтобы это работало, LDA должно заранее знать, сколько тем он ищет. Например, если у нас есть один документ, мы можем указать, что ищем четыре разные темы. LDA будет затем проходить через каждое слово, которое появляется в тексте, случайным образом приписывать его одной из четырех тем и вычислять специальный балл для этого слова на основе вероятности того, что это слово будет найдено в этой конкретной теме в наборе документов. (в нашем случае только один документ). LDA затем приписывает это же слово к другой теме и вычисляет ту же оценку. После многих итераций мы получаем список слов в каждой теме с вероятностями. Для каждой темы мы можем выбрать 5 лучших слов с наибольшей вероятностью принадлежности к этой конкретной теме, и мы получим довольно хорошее описание того, о чем эта тема, путем сочетания слов с наибольшей вероятностью. Естественно, эти слова имеют тенденцию сосуществовать вместе в одном контексте. Слова с высокой частотой займут более видное место в каждой теме ([**Блей 2012**](http://menome.com/wp/wp-content/uploads/2014/12/Blei2011.pdf) ). Отличительной особенностью скрытого распределения Дирихле является то, что все документы в коллекции имеют одинаковый набор тем, но в каждом документе эти темы представлены с разной пропорцией.

LDA, пожалуй, самый популярный и проверенный метод моделирования тем на сегодняшний день. Так как он был представлен несколько лет назад, он прошел некоторые обновления. Например, lda2vec - это модель, которая сочетает в себе LDA с word2vec, обеспечивая лучшее понимание слов и тем. Сам LDA является частным случаем pLSA - вероятностного скрытого семантического анализа, который является модификацией LSA (скрытого семантического анализа).

Один из главных недостатков LDA заключается в том, что его довольно сложно понять и что для его продолжения необходимо точное количество тем. Более того, тот факт, что это модель «мешок слов» (т. Е. Последовательность слов и структура повествования потеряны), может привести к некоторым потерям, особенно для текстов со сложной повествовательной структурой (например, несколько тем, которые встречаются в разных части текста). Мы также склонны терять отношения между темами. Более того, если определенные слова встречаются по всему тексту, они могут присутствовать во всех идентифицированных темах, делая результаты менее точными.

Вот почему были предприняты многочисленные попытки найти другие, более эффективные подходы, и анализ текстовой сети является одним из них.

<https://medium.com/@tomar.ankur287/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045>

**Что такое тематическое моделирование?**

Тематическое моделирование - это ветвь неконтролируемой обработки естественного языка, которая используется для представления текстового документа с помощью нескольких тем, которые могут наилучшим образом объяснить основную информацию в конкретном документе. Это можно рассматривать с точки зрения кластеризации, но с разницей. Теперь вместо числовых функций у нас есть набор слов, которые мы хотим сгруппировать таким образом, чтобы каждая группа представляла тему в документе.

**Зачем нам нужно моделирование темы?**

Итак, теперь возникает вопрос, зачем нам нужно моделирование темы? Если мы посмотрим вокруг, то увидим огромное количество текстовых данных, лежащих вокруг нас в неструктурированном формате в виде новостных статей, исследовательских работ, сообщений в социальных сетях и т. Д., И нам нужен способ понять, организовать и пометить эти данные, чтобы принимать обоснованные решения. Тематическое моделирование используется в различных приложениях, таких как поиск вопросов о переполнении стека, похожих друг на друга, агрегация и анализ потока новостей, системы рекомендаций и т. Д. Все они направлены на поиск скрытой тематической структуры в тексте, так как считается, что каждый текст, который мы пишем, будь то твит, пост или исследовательская работа, состоит из таких тем, как спорт, физика, авиакосмическая промышленность и т. д.

**Как сделать тему моделирования?**

В настоящее время существует много способов сделать тематическое моделирование, но в этой статье мы будем обсуждать вероятностный подход к моделированию, называемый «скрытое распределение дирихле» (LDA), разработанный профессором Дэвидом М. Блеем в 2003 году. Это расширение вероятностной скрытой семантики. Анализ (PLSA), разработанный в 1999 году Томасом Хоффманом, с очень незначительной разницей в том, как они относятся к распространению каждого документа. Итак, давайте перейдем прямо к тому, как работает LDA.

# **Скрытое распределение Дирихле**

Давайте начнем с понимания значения каждого слова в названии, так как я думаю, что оно содержит все, что нам нужно знать, чтобы понять, как работает LDA.

**Скрытый** : это относится ко всему, что мы не знаем априори и скрыты в данных. Здесь темы или темы, из которых состоит документ, неизвестны, но считается, что они присутствуют, поскольку текст генерируется на основе этих тем.

**Дирихль:**Это «распределение дистрибутивов». Да, вы правильно прочитали. Но что это значит? Давайте подумаем об этом на примере. Давайте предположим, что есть машина, которая производит кости, и мы можем контролировать, будет ли машина всегда производить кости с одинаковым весом со всех сторон, или будет какое-либо смещение для некоторых сторон. Таким образом, машина, производящая кости, является дистрибуцией, поскольку она производит кости разных типов. Кроме того, мы знаем, что сама игральная кость является распределением, поскольку мы получаем несколько значений, когда бросаем кости. Это то, что значит быть распределением распределений, и это то, чем является Дирихле. Здесь, в контексте моделирования темы, Дирихле - это распределение тем в документах и ​​распределение слов в теме. Это может быть не очень ясно в данный момент,

**Распределение** : это означает, что, как только у нас будет Dirichlet, мы будем распределять темы по документам и слова документа по темам.

Вот и все. Это то, что LDA в двух словах. Теперь давайте разберемся, как это работает в моделировании темы.

<https://nlpforhackers.io/topic-modeling/>

**Что такое тематическое моделирование?**

Тематическое моделирование в контексте обработки естественного языка описывается как метод выявления скрытой структуры в наборе текстов. Хотя это действительно так, это также довольно бесполезное определение. Давайте определим тему моделирования в более практических терминах.

**Определения:**

* C: коллекция документов, содержащих Nтексты.
* V: словарный запас (набор уникальных слов в сборнике)

**Почему моделирование тем полезно?**

Есть несколько сценариев, когда тематическое моделирование может оказаться полезным. Вот некоторые из них:

* **Классификация текста** - моделирование темы может улучшить классификацию, группируя похожие слова вместе в темах, а не используя каждое слово в качестве функции
* **Рекомендательные системы** - Используя меру сходства, мы можем создать рекомендательные системы. Если наша система будет рекомендовать статьи для читателей, она будет рекомендовать статьи со структурой тем, аналогичных тем, которые пользователь уже прочитал.
* **Раскрытие тем в текстах** - например, полезно для выявления тенденций в онлайн-публикациях

**Алгоритмы моделирования тем**

Есть несколько алгоритмов для тематического моделирования. Самые популярные из них включают

* **LDA** - скрытое распределение Dirichlet - то, на котором мы сосредоточимся в этом уроке. Его основой являются **вероятностные графические модели.**
* **LSA** или **LSI** - скрытый семантический анализ или скрытое семантическое индексирование - использует разложение по сингулярным значениям (SVD) в матрице термина документа. На основе **линейной алгебры**
* **NMF** - Факторизация неотрицательной матрицы - на основе **линейной алгебры**

<https://analyticsdefined.com/introduction-to-topic-modeling-using-lda-latent-dirichlet-allocation/>

**Скрытое распределение Дирихле (LDA)** является «генеративной вероятностной моделью» и является одним из наиболее часто используемых методов для тематического моделирования. Он строит слова для каждой темы и темы для каждого документа, который моделируется как дистрибутив Дирихле.

Вероятностная тематическая модель для тематического моделирования с использованием LDA состоит из двух таблиц. Первая таблица является таблицей терминов документа (N, T), а вторая таблица является таблицей терминов темы (T, M), где

* N - количество документов
* Т - количество тем
* М - количество уникальных терминов в словаре

Таблица Document-Term выглядит так:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | т 1 | т 2 | т 3 | т т |
| Д 1 | 0 | 1 | 1 | 0 |
| D 2 | 0 | 1 | 0 | 1 |
| D 3 | 0 | 1 | 0 | 0 |
| Д Н | 1 | 1 | 0 | 1 |

Тема-таблица выглядит так:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | W 1 | W 2 | W 3 | W M |
| т 1 | 0 | 1 | 1 | 0 |
| т 2 | 0 | 1 | 0 | 1 |
| т 3 | 0 | 1 | 0 | 0 |
| т т | 1 | 1 | 0 | 1 |

Алгоритм LDA начинается с начальных распределений тематических слов и тематических документов, а затем итеративно улучшает эти распределения с использованием методов выборки. Он просматривает каждое слово «w» в каждом документе, а затем пытается улучшить текущую тему. Затем создается новая тема «T» путем присвоения слова «w» теме с вероятностью «P», которая является произведением для двух вероятностей p 1 и p 2 , где:

Параметры LDA

Есть в первую очередь два гиперпараметра LDA, альфа и бета.

* Альфа представляет плотность документа-темы, которая представляет собой количество тем в документе, чем выше значение альфа, тем больше количество тем в документе и наоборот.
* Бета представляет собой плотность слов в теме, которая представляет собой количество слов в теме, чем выше значение беты, тем выше количество слов в теме и наоборот.

Другие параметры для LDA:

* Количество тем: Это общее количество тем, которые будут извлечены из корпуса. Существует ряд методов для определения оптимального количества тем в корпусе, таких как показатель расхождения Kullback Leibler и т. Д. Он также может быть оценен путем итеративного построения графика количества тем и показателя конвергенции и проверки числа, откуда резко падает показатель конвергенции. ,
* Количество итераций: это максимальное количество итераций для алгоритма LDA, которые должны сходиться.