ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 624.333:630.28
DOI: 10.22213/2410-9304-2024-4-24-31

Тематическое моделирование текстового документа с использованием матрицы терминов обработанного документа

М. М. Аббаси, кандидат технических наук, Удмуртский государственный университет, Ижевск, Россия
А. П. Бельтюков, доктор физико-математических наук, профессор, Удмуртский государственный университет, Ижевск, Россия

Тематическое моделирование – это метод определения темы текстового документа путем анализа семантики и синтаксиса текста. При анализе текста метод определяет внутреннюю структуру документа или набора документов и использует эту информацию для классификации или группировки похожих слов по темам. Это также помогает выявить основные тенденции интересов в текстовом документе. Например, многие люди интересуются онлайн-покупками, политикой, спортом, экономикой, обществом и т. д. Существуют различные онлайн- и офлайн-методы интеллектуального анализа данных и алгоритмы, используемые для определения темы текста. Большинство из них используют определенный механизм, основанный на семантических характеристиках языка и тематике текста. В данном исследовании основная идея заключается в разработке методологии, которую можно эффективно использовать для тематического моделирования текста на разных языках. Модель сначала предварительно обрабатывает текст, который включает в себя токенизацию слов, удаление из него стоп-слова (STOPWORDS), выполнение лемматизации. Предварительная обработка текста и фильтрация несоответствующих элементов уменьшает размер текста и повышает производительность его классификации. Алгоритм предполагает наличие 'n' тем в текстовом документе и, основываясь на этом предположении, генерирует матрицы терминов обработанного документа (PDTM) для текстового документа. Матрица терминов обработанного документа (PDTM) представляет собой двумерную матрицу, которая присваивает конкретное числовое значение каждому слову в тексте на основе частоты его появления в документе, а затем соотносит это слово с каждой темой, предполагавшейся ранее. Матрица терминов обработанного документа (PDTM) генерируется для хранения токенизированных слов. Предлагаемая модель и ее результаты подробно описаны в разделах методологии и обсуждения этой статьи.

Ключевые слова: тема, анализ, текст, тематическое моделирование, токенизация, лемматизация.

Введение

Текст становится основным источником информации и обеспечивает механизм для эффективной коммуникации человека. Увеличение использования интернета изменило способ, которым люди приобретают знания, информацию и распространяют их дальше. В интернете существует огромное количество доступного текста. Но не весь текст в интернете правильно классифицирован. В некоторых случаях сложно определить интересующую тему по текстовому документу. Тематическое моделирование – это механизм для анализа внутренней структуры слов и семантики текста и их классификации по темам. Данный механизм работает на том принципе, что текстовый документ представляет собой совокупность различных тем, и эти темы генерируют слова на основе их распределения в тексте. При обработке текста на естественном языке он используется для анализа статистических закономерностей, скрытых в текстовых данных. Как контролируемые, так и неконтролируемые модели используются для создания тем в текстовых данных. Интересно наблюдать различие слов в теме текста и отношения между разными словами и темой.

Тематическое моделирование используется в различных областях технологий, таких как биоинформатика, для извлечения биологической информации из различных данных. Здесь это работает как классификация или подход кластеризации. Механизм моделирует биологический объект в терминах скрытых тем, которые в большей степени отражают неизвестный биологический смысл. Другим важным использованием модели является обобщение мнения людей, выраженного в тексте. Различные компании заинтересованы в предоставлении более качественных услуг и понимании мнения людей об их политике. Речь идет не только о том, какое мнение является положительным или отрицательным, но также и о том, чтобы определить события, связанные с такими мнениями. Точно так же подготовка документов во время встречи требует усилий для создания итогового документа.

Эффективный подход к моделированию темы может помочь найти тему и краткое содержание встречи. Это может улучшить процесс анализа настроений, выраженных в тексте, доступном в интернете. Метод способен анализировать настроения и обобщать их по теме. Аналогичным образом, используя тематическое моделирование, можно создать систему рекомендаций, которая может предлагать авторам книги, подобные тем, которые он уже прочитал. Он используется в фильтрации спама, категоризации текста, отслеживании тем и даже для улучшения взаимодействия человека с машиной.

Научные труды, связанные с данной работой

Работы по анализу текста и его категоризация начались еще до появления интернета и методов машинного обучения. Анализ и классификация текста, написанного вручную, впервые привлекли внимание в конце 70-х годов. В 1978 году Кэри и Бартлетт предложили модель, которая использовала статистический паттерн для определения совпадений терминов в тексте, написанном детьми, и использовала механизм для моделирования темы, которая помогает в развитии языка детей [1]. В 1989 году Лю использовал математическую модель BFGS для анализа текста и его классификации по системе с ограниченным ресурсом памяти [2]. После этого произошла небольшая задержка в исследовании темы моделирования из текстового документа.

В 2003 году Andrew и Ng опубликовали статью об автоматической классификации текста по темам с использованием алгоритма скрытого распределения Дирихле. LDA – это итеративный алгоритм, который использовался для случайного присвоения слов из текста другой теме. Они наблюдали высокую классификационную способность алгоритма классифицировать слова по темам [3]. В 2005 году Griffiths проанализировал текст и его элемент в виде временного ряда. Он создал документ в хронологическом порядке для анализа текста, а затем для его тематического моделирования [4]. В 2006 году Уоллах предложил расширение в моделировании тем, проанализировав отношения, существующие между словами в документе [5]. В том же году Вэй и Крофт использовали моделирование тем для выполнения информационно-поисковых задач. Они использовали внешнюю модель и автоматически выполнили моделирование темы [6].

В 2008 году Ansuncion провел опрос для сравнения производительности и обучающей способности алгоритмов моделирования тем. Он пришел к выводу, что путем настройки и сглаживания гиперпараметров текста алгоритмы работают лучше эмпирически [7]. В 2009 году Яо использовал математическую модель для расчета итерации алгоритма LDA в длинном текстовом документе. Он заметил, что в случае, если текст написан плохо, число итераций значительно возрастает [8]. Точно так же Chang предложил подход к моделированию тем, который анализирует внутренние характеристики текста под контролем человека-агента. Человек-агент выносит суждения о производительности алгоритма тематического моделирования во время каждой итерации алгоритма [9]. Позже, в 2010 году, Гриффитс Блей и Джордан предлагают модель для классификации слов в неограниченном количестве тем. Они применили моделирование тем в тексте о ресторанах и использовали непараметрический вывод для определения иерархии тем в тексте [10].

В 2012 году Blei предложил новый механизм использования тематического моделирования не только для текстовых данных, но и для анализа данных в биоинформатике и в компьютерном зрении. Согласно им, тематическое моделирование делает три важных предположения, которые включают в себя принятие фиксированного количества тем, обмен словами в документах и обмен документами [11]. В 2016 году Qiang использовал технику искусственного интеллекта для идентификации темы в гетерогенном тексте. Он собрал текст из разных источников на разные темы и применил методы искусственного интеллекта для определения основной темы текста [12].

В 2017 году Wang предложил бимодальный механизм моделирования тем, который анализирует внешнюю характеристику текстового документа. Он использует показатель корреляции для определения различий между классификацией слов человеком-агентом и алгоритмом машинного обучения. В последнее время популярность получила тенденция к выполнению тематического моделирования в коротких текстах. Было отмечено, что обучающая способность алгоритма для выполнения тематического моделирования в коротком тексте является эффективной по сравнению с большими текстами [13]. В 2018 году Йэн использовал кластеризацию на основе моделей для коротких текстовых потоков. Использовался текст из онлайн-новостей, к которому применялась кластеризация, чтобы классифицировать его по различным темам [14]. Текст и его характеристики были проанализированы с использованием алгоритмов Word-space. Предложенные алгорит-

мы определили связь между разными словами, существующими в тексте [15].

В 2018 году Ши применил метод неотрицательной факторизации для тематического моделирования в коротком тексте. Он проанализировал локальный контекст слова и их корреляции в теме [16]. М. М. Аббаси и А. П. Бельтюков проанализировали логические характеристики текста и выявили семантические и синтаксические зависимости, существующие между разными словами в тексте [17–20]. В 2019 году Цян провел опрос, чтобы проанализировать методы, используемые для тематического моделирования, и привел сравнение их эффективности применительно к короткому тексту [21].

Методология

В этой исследовательской работе предлагается оптимальный метод для тематического моделирования. Тематическое моделирование является важной частью области обработки естественного языка. Оно помогает классифицировать огромное количество текста по различным темам. Для моделирования тем документ должен состоять из списка разнородных тем, а каждая тема представляет собой набор слов. Методология состоит из трех основных этапов.

1. Предварительная обработка текста.

2. Создание матрицы терминов обработанного документа (PDTM) из текста.

3. Распределение слов в соответствии с определенной темой.

Каждый из них далее подразделяется на подэтапы, как описано ниже. Вначале определяется текст для анализа из таких источников, как книги, журналы, или из онлайн-блогов, сайтов социальных сетей. Кроме того, номер темы в тексте принимается за *n* номеров тем.

Алгоритм разделяет текст на набор предложений. На следующем этапе предложения токенизируются в список слов. Маркированные слова хранятся в форме матрицы, называемой «матрица терминов документа» (Document Term Matrix (DTM)). В матрице терминов документа каждый элемент представлен набором числовых значений. DTM [i] [j], что означает количество вхождений [j], слова [i] в документе. [i] представляет индекс первого вхождения слова в тексте.

DTM далее обрабатывается и удаляются STOPWORDS из списка вместе с пунктуацией в нем. В результате генерируется обработанный DTM, называемый «матрица терминов обработанного документа» (Processed Document Term Matrix (PDTM)). Слова в PDTM лемматизированы в свою начальную или корневую форму. Например, слово «играл», «играют» меняется в слово «играть».

В результате количество слов в PDTM уменьшается и частота встречаемости всех форм добавляется к частоте корневого слова. Типичный пример PDTM представлен ниже, в табл. 1.

Таблица 1. **Матрица терминов обработанного документа (PDTM) и ее индексы**

Table 1. **Processed document term matrix (PDTM) and its indexes**

Матрица терминов обработанного документа (PDTM)		
Слово	Индекс первого вхождения слова в текст [i]	Частота встречаемости в тексте [j]
Любовь	20	15
Зрительный зал	101	7
Исследования	75	12
Обед	300	18
Участники	512	16
Конференция	480	5
Заканчивать	321	27
Университет	32	16
Организация	61	4

Эксперимент

Для эксперимента был выбран текст на тему студенческой жизни в университете. Текстовые документы из разных онлайн-блогов, таких как «как проходят университетские будни: рассказывают студенты» (https://thegirl.ru/articles/kak-prohodyat-universitetskie-budni-rasskazyivayut-studentyi/), «9 привычек суперуспешных студентов» (https://studyexp.ru/blog-student/9-privychek-super-uspeshnyh-studentov/), «им вообще, кроме тиктоков, что-то интересно?», «чем на самом деле живут современные студенты» (https://mel.fm/blog/universitet-itmo/3967-im-voobshche-krome-tiktokov-chto-to-interesno-chem-na-samom-dele-zhivut-sovremennyye-studenty), «сессия, столовая, свобода: университет глазами студентов Москвы» (https://www.m24.ru/articles/obshchestvo/25012019/154501) и т. д., были загружены, и затем сгенерирован один файл.

Текст был использован для проведения эксперимента и для его классификации по различным темам. Как упомянуто выше в разделе методологии и в табл. 1, был подготовлен текст для использования в матрице терминов документа (Document Term Matrix (DTM)), который представляет частоту каждого слова в тексте.

Алгоритм выполнял предварительную обработку текста и удалял STOPWORDS, знаки препинания и наименее часто встречающиеся слова из матрицы терминов обработанного документа (Processed Document Term Matrix (PDTM)). Для инициализации в качестве темы были выбраны слова из пяти пар предложений. Каждая пара была сборником первых двух предложений абзаца текста. Вероятность появления каждого слова PDTM наблюдалась в заранее определенных темах. Если слова не было в теме, то оно относится к той теме, к которой относится ближайшее слово-сосед PDTM. Пример таких расчетов приведен ниже, в табл. 2.

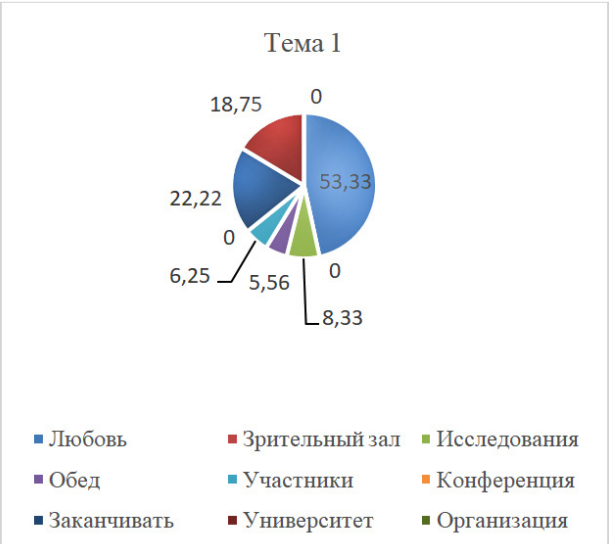
Таблица 2. **Вероятность появления слов в разных темах**

Table 2. **The probability of occurrence of words in different topics**

Слово	Частота встречаемости в тексте	Вероятность появления слов в разных темах, %				
		Тема 1	Тема 2	Тема 3	Тема 4	Тема 5
Любовь	15	53,33	20,00	–	13,33	13,33
Зрительный зал	7	–	28,57	57,14	–	14,29
Исследования	12	8,33	16,67	58,33	16,67	–
Обед	18	5,56	16,67	55,56	–	22,22
Участники	16	6,25	18,75	56,25	6,25	12,50
Конференция	5	–	20,00	60,00	–	20,00
Заканчивать	27	22,22	14,81	55,56	3,70	7,41
Университет	16	18,75	43,75	–	12,50	25,00
Организация	4	–	75,00	–	25,00	–

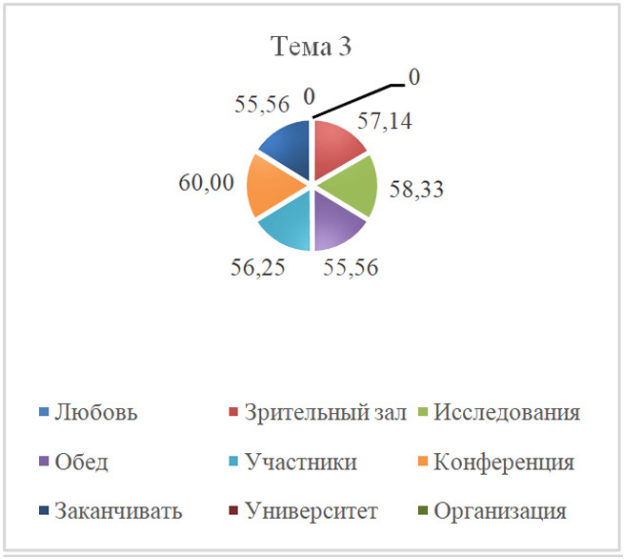
В приведенной выше табл. 2 представлена общая частота каждого слова в документе и вероятность его появления в каждой теме, упомянутой в таблице.

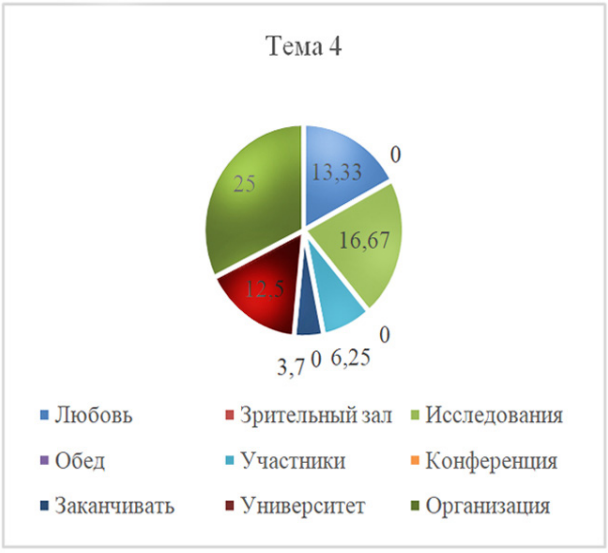
Алгоритм выделил слово для темы, в которой оно имеет высокую вероятность появления. Это можно наблюдать в каждой теме на рисунке ниже.

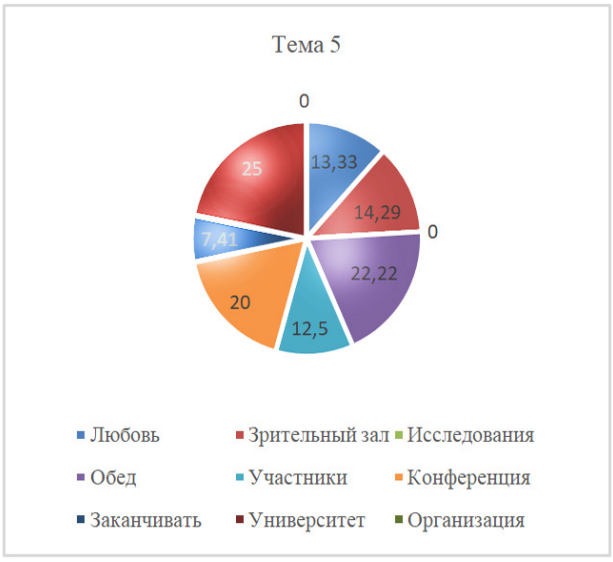


Диаграммы из пяти тем, представляющих слова в каждой теме (окончание на с. 28)

Diagrams from five topics, representing the words from each topic







Диаграммы из пяти тем, представляющих слова в каждой теме (окончание, начало на с. 27)

Diagrams from five topics, representing the words from each topic

Из рисунка можно заметить, что нет конкретного имени, назначенного алгоритмом для темы. Это задача человека – наблюдать за диаграммой в каждой теме и назначать подходящую тему.

Название каждой темы было подписано в зависимости от характера слов, представленных в теме. Задание было выполнено в конце классификации слов по разным темам, и оно зависит от человеческого понимания природы слов в каждой теме.

Результаты категоризации алгоритма были проверены и сопоставлены матрицы ошибок и ее показателей, таких как точность измерений, отзыв, точность результата измерений и F-мера. Матрица ошибок – это контролируемый метод обучения, при котором значения матрицы заполняются вручную, чтобы определить эффективность классификации программы или алгоритма.

Результаты матрицы ошибок были использованы для определения способности программы правильно классифицировать неоднозначное слово в его соответствующем значении в тексте. Это метод, используемый для определения истинной и ложной категоризации слова в соответствующей теме. Используя данный метод, точность алгоритма вычисляется вместе с его отзывом и F-мерами. Результаты матрицы ошибок по каждой теме подробно описаны в табл. 3.

Таблица 3. **Статистика производительности алгоритма, использующего матрицу ошибок**

Table 3. **Performance measure of the algorithms using confusion matrix**

Номер темы	Точность измерений (Accuracy), %	Отзыв (Recall), %	Точность результата измерений (Precision), %	F-Мера (F-Measure), %
Тема 1	89,66	85,71	75,00	80,00
Тема 2	81,82	71,43	71,43	71,43
Тема 3	91,30	80,00	80,00	80,00
Тема 4	80,65	76,92	76,92	76,92
Тема 5	77,08	82,35	77,78	80,00

Исходя из данных табл. 3, можно наблюдать высокую точность и производительность алгоритма по всем темам. В среднем показатель точности алгоритма составляет более 80 %. Представленный выше вариант хорошо подходит для моделирования слов в разных темах.

Заключение

В ходе исследования предыдущие работы по тематическому моделированию были подробно изучены и проанализированы. Предварительная обработка текста и фильтрация несоответствующих элементов текста уменьшили размер текста и повысили производительность его классификации. Программа обеспечивает гибкость выбора любого количества тем. Инициализация темы в виде слов из пары предложений в каждом абзаце происходит в соответствии с правилом языка, где первое предложение выбирает тему последующего отрывка. Характеристики соседства описывают отношения между словами в тексте. Высокая точность показала производительность предлагаемой модели. В дальнейшем алгоритм будет доработан для работы с большими объемами текста. Текст будет более разнородным по природе.

Библиографические ссылки

1. Carey S., Bartlett E. Acquiring a single new word // Papers and Reports on Child Language Development. 1978. 15 (1), pp. 17-29. URL: https://api.semanticscholar.org/CorpusID:50145091.

2. LiuD.C.,Nocedal J. On the limited memory bfgs method for large scale optimization // Mathematical programming. 1989. 45 (1), pp. 503-528. URL: https://doi.org/10.1007/BF01589116.

3. Blei D., Ng. A., Jordan M. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. 3 (4), pp. 993–1022. URL: https://dl.acm.org/doi/10.5555/944919.944937.

4. Griths T., Steyvers M., Blei D.,Tenenbaum J. Integrating topics and syntax // Proc. of Neural Information Processing Systems .Vancouver, British Columbia, Canada. 2004. URL: https://papers.nips.cc/paper_files/paper/2004/hash/ef0917ea498b1665ad6c701057155abe-Abstract.html.

5. Wallach H. Topic modeling: beyond bag-of-words // Proc. ofTwenty-Third International Conference ICML, Pittsburgh, Pennsylvania, USA, 2006, June 25-29, 2006. DOI:10.1145/1143844.1143967.

6. Wei X., Croft B. LDA-based document models for ad-hoc retrieval // Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA. 2006. DOI:10.1145/1148170.1148204.

7. Asuncion A.,Welling M., Smyth P., The Y.W. On smoothing and inference for topics models // Proc. of the 25th Conference on Uncertainty in Artificial Intelligence, June, 2009, pp. 27-34. URL: https://dl.acm.org/doi/10.5555/1795114.1795118.

8. Yao L., Mimno D., McCallum A. Ecient methods for topic model inference on streaming document collections // Proc. of the 15thACM SIGKDD International Conference on Knowledge discovery and data mining. 2009. Pp. 937–946. URL: https://doi.org/10.1145/1557019.1557121.

9. Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. M. Reading tea leaves: How humans interpret topic models // Proc.of 23rd Annual Conference on Neural Information Processing Systems // Advances in Neural Information Processing Systems. 2009. 32 (1). Pp. 288-296.

10. Blei D.M., Griths T., Jordan M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies // ACM. 2010. 57 (2). Pp. 1–30. URL: https://cocosci.princeton.edu/tom/papers/ncrp.pdf.

11. Blei D.M. Probabilistic topic models // Communications of the ACM. 2012. 55 (4). Pp.77–84. Doi.org/10.1145/2133806.2133826.

12. Qiang J.,Chen P.,Ding W.,Wang T.,Xie F.,Wu X. Topic discovery from heterogeneous texts //In: Tools with Artificial Intelligence (ICTAI), IEEE 28thInternational Conference on IEEE. 2016. Pp. 196–203. Doi: 10.1109/ICTAI.2016.0039.

13. Wang S., Roller S.,Erk K. Distributional model on a diet: One-shot word learning from text only. CoRR. 2017. URL: https://arxiv.org/abs/1704.04550v4.

14. Yin D., Chao Z., Liu W., Zhang X., Yu J. Wang. Model-based clustering of short text streams // Proc. of the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2018, pp. 2634–2642. Doi: 10.1145/3219819.3220094.

15. Abbasi M. M., Beltiukov A. P. Analyzing emotions from text corpus using word space CSIT`2018 // Proc. of the 20thInternational Workshop on Computer Science and Information Technologies ,Varna- Bulgaria, pp. 90-94, Industry 4.0, 2018, 3 (4), pp. 161-164. URL: https://stumejournals.com/journals/i4/2018/4/161.

16. Shi T., Kang K., Choo J., ReddyC. K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations // Proc. oftheWorld Wide Web Conferences Steering Committee, 2018, pp. 1105–1114. DOI: 10.1145/3178876.3186009.

17. Beltiukov A. P., Abbasi M. M. Logical analysis of Emotions in Text from Natural language // Vestnik Udmurtskogo Universiteta. Matematika. Mekhanika. Komp'yuternye Nauki, Ижевск. 2019. 1 (29). Pp. 106-116. URL: https://doi.org/10.20537/vm190110.

18. Abbasi M. M., Beltiukov A. P. Identifying the strength of emotions in relation with the topic of text using Word space // Proc. of the 21thinternational workshop on computer science and information technologies, Austria, Vienna // Journal of Atlantis Highlights in Computer Sciences, 2019, 3 (1), pp. 1-5. DOI: 10.2991/csit-19.2019.1.

19. Abbasi M. M., Beltiukov A. P., Hussain L., Abbasi A. Q. Analysis of emotions from texts for managing society // Infocommunication technologies Journal,

Academy of Telecommunications and Informatics, 2019, 2 (17), pp.246-254.

20. Abbasi M.M., Beltiukov A.P. Summarizing emotions from text using Plutchik wheel of emotion // Proc. of the 7thAll Russian Conference on Information technology for intelligent decision making support (ITIDS), Ufa, Russian Federation, 2019, 7 (166), pp. 291-294. DOI: 10.2991/itids-19.2019.52.

21. Qiang J., Qian Z., Li Y., Yuan Y., Wu X. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey // Journal of latex class files, Published in IEEE Transactions on Knowledge and Data Engineering, 2019, 34(1), pp.1427-1445. DOI:10.1109/tkde.2020.2992485.

References

1. Carey S., Bartlett E. Acquiring a single new word. In Papers and Reports on Child Language Development. 1978. 15 (1), pp. 17-29. URL: https://api.semanticscholar.org/CorpusID:50145091.

2. Liu D.C., Nocedal J. On the limited memory bfgs method for large scale optimization. In Mathematical programming. 1989. 45 (1), pp. 503-528. URL: https://doi.org/10.1007/BF01589116.

3. Blei D., Ng. A., Jordan M. Latent Dirichlet Allocation. In Journal of Machine Learning Research. 2003. 3 (4), pp. 993–1022. URL: https://dl.acm.org/doi/10.5555/944919.944937.

4. Griths T., Steyvers M., Blei D.,Tenenbaum J. Integrating topics and syntax // Proc. of Neural Information Processing Systems .Vancouver, British Columbia, Canada. 2004. URL: https://papers.nips.cc/paper_files/paper/2004/hash/ef0917ea498b1665ad6c701057155abe-Abstract.html.

5. Wallach H. Topic modeling: beyond bag-of-words. In Proc. ofTwenty-Third International Conference ICML, Pittsburgh, Pennsylvania, USA, 2006, June 25-29, 2006. DOI:10.1145/1143844.1143967.

6. Wei X., Croft B. LDA-based document models for ad-hoc retrieval. In Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA. 2006. DOI:10.1145/1148170.1148204.

7. Asuncion A.,Welling M., Smyth P., The Y.W. On smoothing and inference for topics models. In Proc. of the 25th Conference on Uncertainty in Artificial Intelligence, June, 2009, pp. 27-34. URL: https://dl.acm.org/doi/10.5555/1795114.1795118.

8. Yao L., Mimno D., McCallum A. Ecient methods for topic model inference on streaming document collections. In Proc. of the 15thACM SIGKDD International Conference on Knowledge discovery and data mining. 2009. Pp. 937–946. URL: https://doi.org/10.1145/1557019.1557121.

9. Chang J., Boyd-Graber J., Gerrish S., Wang C.,Blei D. M. Reading tea leaves: How humans interpret topic models. In Proc.of 23rd Annual Conference on Neural Information Processing Systems // Advances in Neural Information Processing Systems. 2009. 32 (1). Pp. 288-296.

10. Blei D.M., Griths T., Jordan M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. In ACM. 2010. 57 (2). Pp. 1–30. URL: https://cocosci.princeton.edu/tom/papers/ncrp.pdf.

11. Blei D.M. Probabilistic topic models. In Communications of the ACM. 2012. 55 (4). Pp.77–84. Doi.org/10.1145/2133806.2133826.

12. Qiang J.,Chen P.,Ding W.,Wang T.,Xie F.,Wu X. Topic discovery from heterogeneous texts. In Tools with Artificial Intelligence (ICTAI), IEEE 28thInternational Conference on IEEE. 2016. Pp. 196–203. Doi: 10.1109/ICTAI.2016.0039.

13. Wang S., Roller S.,Erk K. Distributional model on a diet: One-shot word learning from text only. CoRR. 2017. URL: https://arxiv.org/abs/1704.04550v4.

14. Yin D., Chao Z., Liu W., Zhang X., Yu J. Wang. Model-based clustering of short text streams. In Proc. of the 22ndACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2018, pp. 2634–2642. Doi: 10.1145/3219819.3220094.

15. Abbasi M. M., Beltiukov A. P. Analyzing emotions from text corpus using word space CSIT`2018. In Proc. of the 20thInternational Workshop on Computer Science and Information Technologies ,Varna- Bulgaria, pp. 90-94, Industry 4.0, 2018, 3 (4), pp. 161-164. URL: https://stumejournals.com/journals/i4/2018/4/161.

16. Shi T., Kang K., Choo J., ReddyC. K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In Proc. oftheWorld Wide Web Conferences Steering Committee, 2018, pp. 1105–1114. DOI: 10.1145/3178876.3186009.

17. Beltiukov A. P., Abbasi M. M. Logical analysis of Emotions in Text from Natural language. In Vestnik Udmurtskogo Universiteta. Matematika. Mekhanika. Komp'yuternye Nauki, Ижевск. 2019. 1 (29). Pp. 106-116. URL: https://doi.org/10.20537/vm190110.

18. Abbasi M. M., Beltiukov A. P. Identifying the strength of emotions in relation with the topic of text using Word space. In Proc. of the 21thinternational workshop on computer science and information technologies, Austria, Vienna. In Journal of Atlantis Highlights in Computer Sciences, 2019, 3 (1), pp. 1-5. DOI: 10.2991/csit-19.2019.1.

19. Abbasi M. M., Beltiukov A. P., Hussain L., Abbasi A. Q. Analysis of emotions from texts for managing society // Infocommunication technologies Journal, Academy of Telecommunications and Informatics, 2019, 2 (17), pp.246-254.

20. Abbasi M.M., Beltiukov A.P. Summarizing emotions from text using Plutchik wheel of emotion. In Proc. of the 7thAll Russian Conference on Information technology for intelligent decision making support (ITIDS), Ufa, Russian Federation, 2019, 7 (166), pp. 291-294. DOI: 10.2991/itids-19.2019.52.

21. Qiang J., Qian Z., Li Y., Yuan Y., Wu X. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. In Journal of latex class files, Published in IEEE Transactions on Knowledge and Data Engineering, 2019, 34(1), pp.1427-1445. DOI:10.1109/tkde.2020.2992485.

Topic Modeling of a Text Document Using Processed Document Term Matrix (PDTM)

Abbasi Mohsin Manshad, Assistant Professor, Udmurt State University, Izhevsk, Russia
A. P. Beltiukov, DSc in Physics and Mathematics, Professor, Udmurt State University, Izhevsk, Russia

Topic modeling is a method of determining the topic of a text document by analyzing the semantics and syntax of the latter. When analyzing text, the method determines the internal structure of a document or a set of documents and uses this information to classify or group similar words by topic. It also helps to identify the main trends of interests or information in a text document.

For example, many people are interested in online shopping, politics, sports, economics, society, and etc. There are various online and offline data mining methods and algorithms used to determine the topic of a text. Most of them use a certain mechanism based on the semantic characteristics of the language and the subject of the text. In this study, the main idea is to develop a methodology that can be effectively used for topic modeling of a text in different languages.

At first,the model preprocesses a text, which includes its tokenization, deletion of STOPWORDS and its lemmatization. Text preprocessing and filtering of inappropriate text elements reduces the size of the text and improves its classification performance. The algorithm also assumes the presence of 'n' topics in a text document and, based on this assumption, generates the processed document term matrix (PDTM) for a text document.

The Processed Document Term Matrix (PDTM) is a two-dimensional matrix that assigns a specific numerical value to each word in the text based on the frequency of its occurrence in the document, and then correlates this word with each topic assumed earlier. The processed document terms (PDTM) are generated to store tokenized words. The proposed model and its results are described in detail in the methodology and discussion sections of this article.

Keywords: topic, text analysis, topic modeling, tokenization, lemmatization.

Получено: 04.03.24

Образец цитирования

Аббаси М. М., *Бельтюков А. П.* Тематическое моделирование текстового документа с использованием матрицы терминов обработанного документа // Интеллектуальные системы в производстве. 2024. Т. 22, № 4. С. 24–31. DOI: 10.22213/2410-9304-2024-4-24-31.

For Citation

Abbasi M.M., Bel'tjukov A.P. [Topic modeling of a text document using the matrix of terms of the processed document]. *Intellektual'nye sistemy v proizvodstve.* 2024, vol. 22, no. 4, pp. 24-31 DOI: 10.22213/2410-9304-2024-4-24-31.