УДК 621.396.41

С.А. Землянский, С.В. Аксёнов, И.А. Лызин, О.Г. Берестнева

Тематическое моделирование в контексте медицинских текстов

Анализ текста – это важная область исследований, которая включает в себя несколько направлений, таких как информационный поиск, извлечение информации и категоризация текста. Анализ текста широко используется в области медицинских исследований из-за количества ежедневно публикуемых исследований, которые могут быть обработаны с такой скоростью только с помощью вычислительных ресурсов. В данной работе представлены результаты эксперимента по тематическому моделированию корпуса статей из базы данных PubMed с 2000 по 2020 г.

Ключевые слова: науки о здоровье, анализ текстов, латентное распределение Дирихле, научно-исследовательские тенденции, картирование знаний, обобщение знаний, PubMed.
DOI: 10.21293/1818-0442-2021-24-4-58-64

Ежегодно в области здравоохранения публикуется огромное количество исследований [1]. Однако нельзя сказать, что эта область полностью сформирована и может быть описана в терминах фиксированных определений, концепций и областей исследования [2] – это сильно затрудняет обработку и категоризацию всей генерируемой информации.

Чтобы преодолеть эти ограничения, в настоящее время для частичной автоматизации этого процесса используются современные вычислительные методы, такие как машинное обучение, а именно: интеллектуальный анализ текста (text mining), классификация текстов (text classification) и тематическое моделирование (topic modelling) [3]. Последнее предлагает вычислительный инструмент для автоматического поиска релевантных тем путем выявления значимой структуры среди коллекций документов [1]. В данной работе применен метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) для автоматического определения тем в коллекции исследований из базы данных PubMed [4] за период с 2000 по 2020 г.

Тематическое моделирование

Огромное количество биомедицинских текстовых документов может служить важным источником информации для биомедицинских исследований. Биомедицинские текстовые документы характеризуются огромными объемами неструктурированной и разреженной информации в различных формах, таких как научные статьи, биомедицинские наборы данных и отчеты [5, 11].

В то же время выявление релевантных исследований для включения в систематические обзоры или для категоризации полученных знаний является сложной, трудоемкой и дорогостоящей задачей [3, 11]. Однако недавно ряд исследований [1, 2] показал, что использование методов машинного обучения и анализа текста для автоматического определения релевантных исследований и извлечения тем может значительно сократить объем ручной работы и существенно повысить ее качество. По этой причине автоматический тематический анализ в настоящее время набирает популярность в области анализа текста.

Анализ текста направлен на выявление ценной информации из неструктурированных текстовых документов с использованием инструментов и методов из нескольких дисциплин, таких как машинное обучение, информационный поиск и вычислительная лингвистика. Использование текстового анализа является одним из наиболее перспективных инструментов в биомедицинской области, который привлекает большой исследовательский интерес [5–9].

Тематическая модель описывает связи между словами и темами и таким образом выступает в качестве инструмента для обобщения и систематизации информации из больших текстовых коллекций, эта модель позволяет выявить скрытые структуры и неявные зависимости в данных. Тематическое моделирование широко используется при решении задач информационного поиска, автоматического аннотирования и индексирования документов, пополнения тоновых словарей, поиска классов переводческих эквивалентов, определения сопоставимости текстов в многоязычных корпусах текстов [6].

Анализ текста в биомедицинской области может успешно применяться для решения широкого круга задач, включая выявление знаний о конкретных заболеваниях, диагностику, лечение и профилактику рака, определение состояния ожирения у пациентов, выявление факторов риска сердечных заболеваний, аннотирование экспрессии генов и выявление мишеней и кандидатов в лекарственные препараты [5, 11].

Тематическая модель обычно определяется как подход для обнаружения скрытой информации в корпусе (hidden pattern) текстов [3].

Большинство методов обнаружения скрытой информации в совокупности текстов используют автоматическую или полуавтоматическую классификацию текстов [12]. Классификация текстов обычно выполняется с использованием модели «мешка слов» (bag-of-words, BOW). Эта модель предполагает, что слова в документах используются в качестве признаков для классификации, но их порядок игнорируется. Одна из проблем модели BOW заключается в том, что количество уникальных слов, которые появляются в полном корпусе документов,

может быть очень большим; использование такого большого количества признаков может быть проблематичным для некоторых алгоритмов. С другой стороны, существуют методы, позволяющие более компактно представлять документы. Например, LSA и LDA, а также улучшенная версия LDA-PTM (Parsimonious Topic Modeling) [5, 12].

Так, латентно-семантический анализ (LSA) использует сингулярное разложение (Singular Value Decomposition, SVD) для обнаружения семантической информации в корпусе текстов [3]. Этот метод сначала значительно снижает размерность матрицы слов до двух или трех измерений, затем определяет выпуклое множество этих слов. Слова в опорных точках рассматриваются как темы (topic, latent topic). Модель LDA (Latent Dirichlet Allocation, LDA), в свою очередь, использует модель распределения вероятностей для генерации тем [7]. Этот метод рассматривает отдельные темы как вероятностные распределения терминов, присутствующих в корпусе, или кластеры, которые определяют веса этих терминов [7].

Последняя описанная модель является своего рода стандартом во многих недавних исследованиях [2, 6, 8, 12, 15] и была выбрана нами для данного эксперимента.

Латентное размещение Дирихле

При проведении тематического моделирования наиболее широко используемыми в литературе методами обработки являются вероятностные тематические модели, включая LDA [9, 12]. LDA, или метод латентного размещения Дирихле, – это автоматический метод идентификации тем (в отличие от ручного или полуручного подхода) [9].

LDA является примером вероятностного метода моделирования тем, который предполагает, что документ охватывает ряд тем и каждое слово в документе выбирается из вероятностных распределений с различными параметрами. Каждое слово в LDA обозначается уникальным латентным вектором, указывающим на распределение, из которого оно взято [3]. Таким образом, документ содержит набор тем с различными вероятностями, а тема содержит набор слов с различными вероятностями.

Важным допущением модели LDA является гипотеза «взаимозаменяемости» (exchangeability), или «мешка слов», которая означает, что важность слова не связана с порядком его появления в документе, а скорее с частотой его появления [7]. Другим важным допущением этой модели является ненулевая вероятность принадлежности к теме для любого отдельного документа, что в сочетании с большим количеством скрытых параметров модели может приводить к переобучению [12].

Основным недостатком тематической модели LDA является то, что она не способна автоматически определять оптимальное количество тем для коллекции текстов; другими словами, количество тем является настраиваемым гиперпараметром и точное количество тем должно определяться пользователем модели.

Метод LDA осуществляет мягкую кластеризацию и предполагает, что каждое слово в документе порождено некоторой латентной темой, определяемой распределением вероятности на множестве всех слов в тексте.

В нашем исследовании, основываясь на высокой оценке эффективности этого метода в других исследованиях и общей универсальности метода, решено использовать именно этот алгоритм (LDA) для проведения тематического моделирования.

Выбор количества тем

Вероятностные модели, такие как LDA, предлагают алгоритмы для сопоставления коллекций сообщений с наборами ключевых слов, которые представляют основные темы. В этих подходах, однако, выяснение того, сколько конкретных тем представляют наборы ключевых слов, остается отдельной задачей [9]. В то же время определение количества тем имеет решающее значение для тематического моделирования [7].

Существуют различные метрики, используемые при выборе количества тем, наиболее популярными из которых можно считать сложность (perplexity), изолированность (isolation), стабильность (stability) и согласованность (coherence). Можно сказать, что любая хорошая метрика, используемая для выбора оптимального количества тем, при применении должна создавать тематическую модель со следующими свойствами: хорошая предсказательная сила, высокая изолированность между темами, отсутствие перекрывающихся тем и воспроизводимость. В данной работе авторами используется метрика согласованности для выбора оптимального количества тем для каждого корпуса текстов (для каждого года отдельно).

Метрика согласованности может быть описана следующим образом (1):

$$c(t,W_t) = \sum_{\omega_1,\omega_2 \in W_t} \ln \frac{d(\omega_1,\omega_2)+e}{d(\omega_1)}, \qquad (1)$$

где t – это тема, W_t – набор ключевых слов темы, $d(\omega_i, \omega_j)$ – количество документов, в которых встречаются ключевые слова ω_i, ω_j; $d(\omega_i)$ – количество документов, в которых встречается ключевое слово ω_i; ω_1 – первое ключевое слово темы, ω_2 – второе ключевое слово темы.

Методика выбора оптимального количества тем для набора текстов на основе этой метрики заключается в последовательном обучении моделей с разным количеством тем. Для каждой модели рассчитывается метрика согласованности терминов для каждого выбранного кластера текстов (тем), а затем полученные метрики усредняются по всем темам. Полученное значение запоминается. После обучения всех моделей со всеми желаемыми наборами параметров выбирается лучшая модель из тех, которые имеют наибольшую среднюю согласованность внутри тем.

В данной работе обучено 160 тематических моделей с разным количеством тем (от 2 до 10) по 8 для каждого корпуса текстов (всего 20 корпусов,

по 1 на каждый год), для каждой из этих моделей подсчитана метрика согласованности (coherence score) и выбрано оптимальное количество тем для каждой коллекции текстов. На рис. 1 ниже показано значение метрики связности для всех моделей и всех значений гиперпараметров.

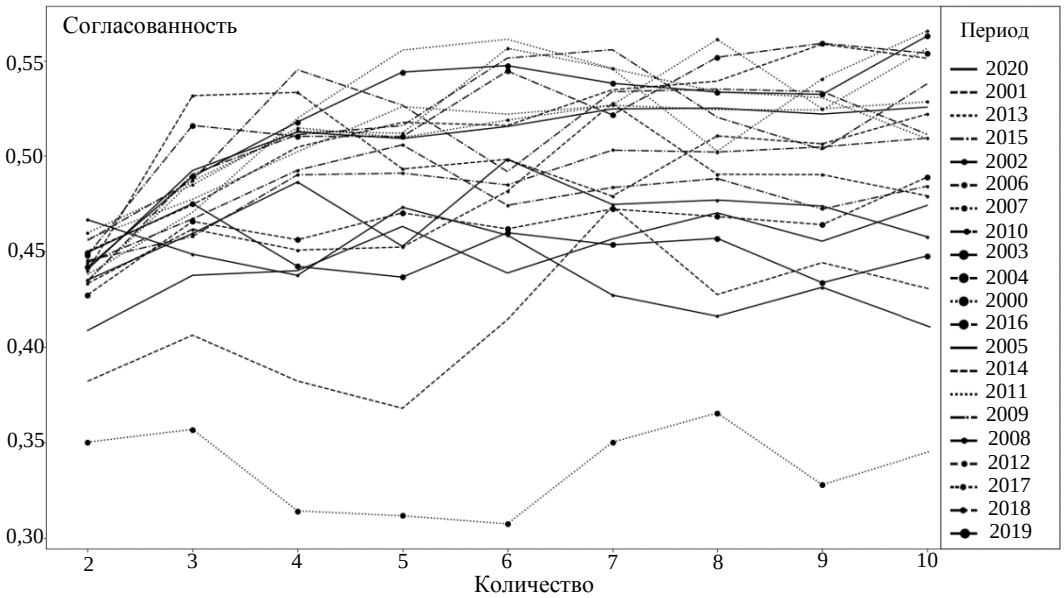

Рис. 1. Значение метрики согласованности для количества тем в модели для каждой коллекции текстов с 2000 по 2020 г.

На рис. 2 показано определенное количество тем в корпусах текстов.

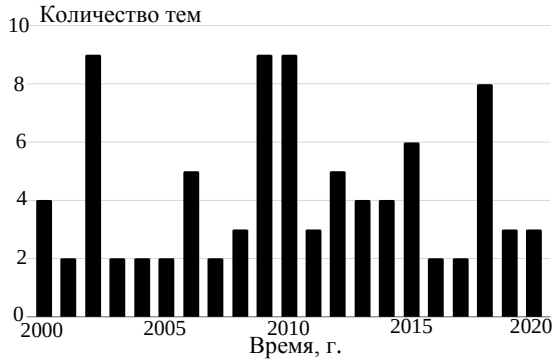

Рис. 2. Количество найденных тем в коллекциях текстов в период с 2000 по 2020 г.

Используемые данные

Входом практически любой тематической модели является корпус текстов, каждый из которых представляет собой отдельный документ. Результатом работы модели является список тем, выявленных в корпусе и представленных списком первых наиболее характерных слов для каждой рассматриваемой темы [6].

Для постановки эксперимента в данной работе были собраны аннотации статей из открытой базы данных PubMed за период с 2000 по 2020 г. При формировании корпуса текстов (в данной работе каждый временной период представлен как отдельный корпус текстов) использовался метод перекрестного включения [3]. Согласно этому методу, в корпусе текстов учитываются не только документы за исследуемый период, но и тексты, на которые ссылается работа. Таким образом, удалось собрать значительное количество записей (с повторами). Количество уникальных записей, в свою очередь, за каждый период времени показано ниже на рис. 3.

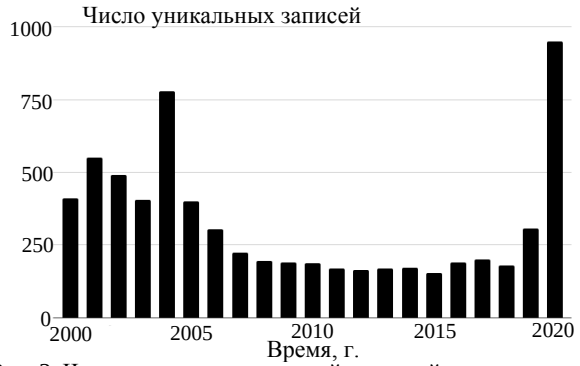

Рис. 3. Число уникальных записей в каждой из коллекции текстов в период с 2000 по 2020 г.

Предварительная обработка данных

Для того чтобы провести тематическое моделирование, собранные тексты были преобразованы в несколько этапов. Сначала были удалены знаки препинания и другие служебные символы. Далее была проведена лемматизация слов (lemmatisation), т.е. все словоформы всех наборов были приведены к лемме, другими словами, к словарной форме слова. После лемматизации были удалены часто используемые слова (также известные как «стоп-слова», stopwords), которые не вносят никакой качественной информации в рассматриваемый документ. Наконец, записи корпуса текстов были дополнены смысловыми биграммами, т.е. пары слов, образующие словосочетания, были идентифицированы и сгруппированы в единый набор значений (лексем).

Все перечисленные этапы обработки текста являются стандартными [12] и хорошо описанными этапами подготовки неструктурированных текстовых данных. Преобразование словоформ в леммы значительно снижает размерность, а удаление часто используемых слов также снижает уровень шума. Преобразование пар лексем, образующих словоформу, в одну лексему также считается обязательной процедурой для обеспечения более точного семантического представления [3]. Несмотря на глубину разработки проблемы предобработки текстовых данных [7, 9–10], многие вопросы до сих пор остаются открытыми. Например, в [6] исследователи предлагают не только объединять пары слов, составляющих фразу, в единый семантический токен, но и более широко использовать метод *n*-грамм, другими словами, использовать при построении модели биграммы и триграммы всех пар слов. Несмотря на значительные теоретические преимущества такого подхода, нельзя сказать, что использование наборов *n*-грамм решает проблему «мешка слов» в моделях вероятностного тематического моделирования, но значительно увеличивает размерность отдельной записи.

В данной работе было решено следовать только стандартной процедуре обработки для подготовки собранных аннотаций статей к дальнейшему тематическому моделированию.

Обучение тематической модели

Как уже упоминалось выше, для проведения качественного тематического моделирования с использованием метода латентного размещения Дирихле перед проведением обучения необходимо установить количество тем в исследуемом наборе данных. Однако, как уже отмечалось, область научных исследований, особенно медицинских, не может быть определена конечным набором направлений исследований. Таким образом, количество тем в наборе текстов не может быть определено заранее. Решением проблемы выбора количества тем становится метод последовательного оценивания модели при подборе гиперпараметров.

Для проведения тематического моделирования был использован пакет для интеллектуального анализа текстов Gensim [10, 12], обучали каждую тематическую модель отдельно от других «по сетке» из 8 значений гиперпараметров от 2 до 10 возможных тем для набора. После обучения всех моделей выбиралась лучшая модель для данного набора данных на основе учета метрики согласованности.

Другие параметры модели были одинаковыми для всех итераций для всех коллекций документов, а именно: количество эпох обучения составляло десять итераций, использовалось итеративное, а не пакетное обучение, также использовался параметр, позволяющий модели самой настраивать параметры априорного распределения вероятностей. Все остальные параметры были оставлены по умолчанию.

При построении графиков «облака слов» ключевые слова были отфильтрованы по частоте во время обучения модели, так что наиболее часто используемые слова с частотой более 50% во всех документах корпуса не учитывались.

Результаты

В вероятностном тематическом моделировании предполагается, что документы представляют собой мультиномиальную смесь латентных тем, а темы представлены в виде распределения вероятности по количеству слов.

Таким образом, LDA разделяет связанные слова на наборы, которые рассматриваются как темы. Однако определение основной концепции, связанной с наборами слов, полученными автоматически, обычно требует дополнительного – возможно, ручного – анализа [9, 11]. Поэтому в данной работе, как и во многих других [1, 8], необходимо вручную дополнить полученные результаты названиями тем после выполнения тематического моделирования. Список определенных тем в соответствии с ключевыми словами, рассчитанными обученной моделью для каждого периода, можно найти в таблице.

Распределение тем среди коллекций

Период	Темы, определённые из набора ключевых слов
2000	positive trends (increased rates), patient survival, data analysis/modeling, vaccine research
2001	positive trends (increased rates), patient survival
2002	positive trends (increased rates), water control, patient survival, data analysis/modeling, disability, clinical trials, surgery, mortality, production costs
2003	positive trends (increased rates), data analysis / modeling
2004	positive trends (increase in rates), patient survival
2005	positive trends (increased rates), patient survival,
2006	positive trends (increased rates), patient survival, data analysis/modeling, vaccine research, plants, cancer
2007	positive trends (increased rates), patient survival, water and plants
2008	positive trends (increased rates), patient survival, water, data analysis/modeling, cancer
2009	plants, national programs, positive trends (increased rates), production costs, data analysis/modeling, mortality, infection, vaccine research
2010	data analysis/modeling, positive trends (increased rates), water and plants, health workers/health system, clinical trials, mortality
2011	data analysis/modeling, positive trends (increased rates), mortality, viruses/ infection, plants
2012	data analysis/modeling, positive trends (increased rates), plants,viruses/ infection, population number
2013	infection, mortality, cancer, surgery, injury, production costs, databases
2014	health system, drugs, infection, data analysis
2015	health system, infection, data analysis, statistics, clinical trials
2016	positive trends (increased rates), plants and water, infection
2017	plants, population, positive trends (increased rates), data analysis
2018	health system, vaccine research, cancer drugs, population/region, data analysis, cancer, infection/virus, plants, clinical trials
2019	positive trends (increased rates), data analysis, patient survival, infection, covid, vaccine, population, health system
2020	positive trends (increased rates), population/region, surgery, data analysis, mortality, virus

Примечание: При выполнении работы использованы данные из открытой базы PubMed (в этой базе данные на англ. яз.).

Выводы

Метод тематического моделирования очень эффективен для извлечения знаний, содержащихся в неструктурированных больших данных, и широко используется в анализе биологических и медицинских документов [1–4, 8, 9, 11, 12]. Однако, как и при всех подходах к анализу текста, возникают трудности с интерпретацией и субъективной проверкой, поскольку вся информация, содержащаяся в этих документах (которая может содержаться), и количество релевантных тем априори неизвестны. Была определена лучшая модель, используя метод, основанный на вычислении метрики согласованности. Однако общее количество тем остается неизвестным и зависит от разумных умозаключений.

В данной работе использовался алгоритм LDA – вычислительный подход к тематическому моделированию без учителя (unsupervised task) для автоматического определения тем в коллекции текстов исследований. Это позволило извлечь ключевые термины и темы из огромного массива литературы в основном автоматическим способом. Таким образом, в данной работе можно проследить эволюцию тем исследований от прошлого к настоящему в портале PubMed, представлены результаты вместе с использованным подходом. В таблице показаны темы, извлеченные из набора ключевых слов для каждого корпуса текстов, на рис. 4, 5 представлены графики облаков ключевых слов за весь период и за 2018–2020 гг.

Рис. 4. Облако ключевых слов за весь период

На основании полученных результатов можно сделать вывод, что вероятностные автоматические модели для извлечения тем из корпуса текстов имеют один недостаток: большинство моделей сильно зависят от настройки гиперпараметров, а не от содержания корпуса. В то же время можно получить одинаковые результаты при различных уровнях частотной фильтрации. Тем не менее из полученных в результате экспериментов данных можно сделать выводы о соответствии определенных глобальных тем известным событиям (например, см. ключевое слово – «covid»); также можно наблюдать микроэволюцию тем, например «анализ данных и моделирование», которые со временем трансформируются в «анализ данных и базы данных», что можно связать с ростом информации в целом.

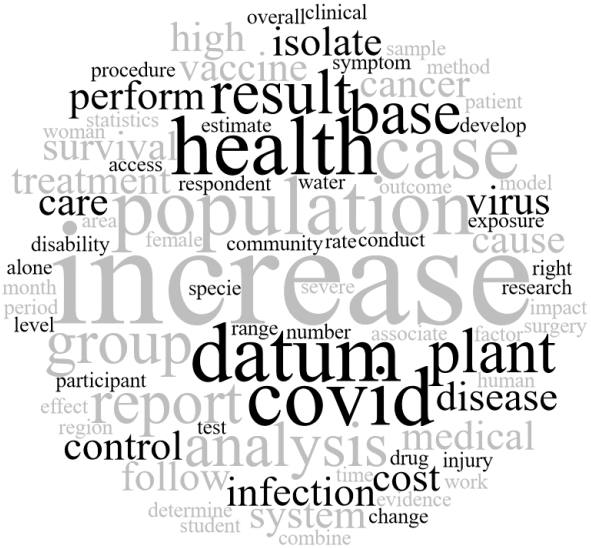

Рис. 5. Облако ключевых слов за 2018–2020 гг.

Исследование выполнено при частичной финансовой поддержке РФФИ в рамках выполнения научного проекта № 19-37-90005.

Литература

1. Wang S.-H. Text mining for identifying topics in the literatures about adolescent substance use and depression / S.-H. Wang, Y. Ding, W. Zhao, Y.-H. Huang, R. Perkins, W. Zou, J. J. Chen // BMC public health. – 2016. – No. 16. – P. 279.

2. Cho S.M. The evolution of social health research topics: A data-driven analysis / S.M. Cho, C. Park, M. Song. – Social Science & Medicine. – 2020. – (265). – P. 113299.

3. Mo Y. Supporting systematic reviews using LDA-based document representations / Y. Mo, G. Kontonatsios, S. Ananiadou // Systematic reviews. – 2015. – No. 4. – P. 172.

4. PubMed: National Library of Medicine [Электронный ресурс]. – Режим доступа: https://pubmed.ncbi.nlm.nih.gov/, свободный (дата обращения: 16.12.2021).

5. Onan A. Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling // Computational and mathematical methods in medicine. – 2018. – P. 1–22.

6. Седова А. Тематическое моделирование русскоязычных текстов с опорой на леммы и лексические конструкции / А. Седова, О. Митрофанова // Компьютерная лингвистика и вычислительные онтологии. – 2018. – С. 132–144.

7. Gan J. Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example / J. Gan, Y. Qi // Entropy (Basel, Switzerland). – 2021. – № 10 (23). – P. 1301.

8. Chandrasekaran R. Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study / R. Chandrasekaran V. Mehta, T. Valkunde, E. Moustakas // Journal of medical Internet research. – 2020. – № 10 (22). – P. e22624.

9. Yıldırım A. Identifying Topics in Microblogs Using Wikipedia / A. Yıldırım, S. Üsküdarlı, A. Özgür // PloS one. – 2016. – № 3 (11). – P. e0151885.

10. Rehurek R. Gensim-python framework for vector space modelling / R. Rehurek, P. Sojka // NLP Centre, Faculty of Informatics – Masaryk University, Brno, Czech Republic. – 2011. – № 2 (3). – P. 46–49.

11. Вафин Р.Р. Актуальные проблемы в области извлечения знаний из профессиональных медицинских текстов с применением интеллектуального анализа текста // Вестник Башкирского гос. мед. ун-та. – 2019. – С. 72–75.

12. Min J.-Y. Mining Hidden Knowledge About Illegal Compensation for Occupational Injury: Topic Model Approach / J.-Y. Min, S.-H. Song, H. Kim, K.-B. Min // JMIR medical informatics. – 2019. – № 3 (7). – P. e14763.

13. Wang H. Improved Parsimonious Topic Modeling Based on the Bayesian Information Criterion / H. Wang, D. Miller // Entropy (Basel, Switzerland). – 2020. – № 3 (22). – P. 326.

14. Liu L. An overview of topic modeling and its current applications in bioinformatics / L. Liu, L. Tang, W. Dong, S. Yao, W. Zhou // SpringerPlus. – 2016. – № 1 (5). – P. 1608.

15. Al Moubayed N. Beyond the topics: how deep learning can improve the discriminability of probabilistic topic modelling / N. Al Moubayed, S. McGough, B. Awwad Shiekh Hasan // PeerJ. Computer science. – 2020. – No. 6. – P. e252.

Землянский Сергей Александрович
Аспирант Национального исследовательского
Томского государственного университета (НИ ТГУ)
Ленина пр-т, 36, г. Томск, Россия, 634050
Тел.: +7-953-922-49-58
Эл. почта: qoelky@gmail.com

Аксёнов Сергей Владимирович
Канд. техн. наук, доцент отделения информационных
технологий Инженерной школы информационных
технологий и робототехники (ИШИТР)
Национального исследовательского
Томского политехнического ун-та (НИ ТПУ)
Ленина пр-т, 30, г. Томск, Россия, 634050
Доцент, каф. теоретических основ информатики НИ ТГУ
Ленина пр-т, 36, г. Томск, Россия, 634050
Тел.: +7-913-887-47-90
Эл. почта: axyonov@tpu.ru

Лызин Иван Александрович
Аспирант ИШИТР НИ ТПУ
Ленина пр-т, 30, г. Томск, Россия, 634050
ORCID: 0000-0003-2827-441X
Тел.: +7-923-498-70-30
Эл. почта: Lyzin@tpu.ru

Берестнева Ольга Григорьевна
Д-р техн. наук, профессор ИШИТР НИ ТПУ
Ленина пр-т, 30, г. Томск, Россия, 634050
Профессор каф. теоретических основ информатики НИ ТГУ
Ленина пр-т, 36, г. Томск, Россия, 634050
ORCID: 0000-0002-4243-0637
Тел.: +7-913-106-19-94
Эл. почта: ogb6@yandex.ru

Zemlyansky S.A., Axyonov S.V., Lyzin I.A., Berestneva O.G.

Topic Modeling in the Context of Medical Texts

Text analysis is an important area of research that includes several areas such as information retrieval, information extraction, and text categorization. Text analysis is widely used in the field of medical research because of the number of studies published daily, which can be processed at such a speed only with the help of computational resources. This paper presents the results of an experiment to thematically model a corpus of articles from the PubMed database from 2000 to 2020.

Keywords: health sciences, text analysis, latent Dirichlet distribution, research trends, knowledge mapping, knowledge synthesis, PubMed.
DOI: 10.21293/1818-0442-2021-24-4-58-64

References

1. Wang S.-H., Ding Y., Zhao W., Huang Y.-H., Perkins R. Text mining for identifying topics in the literatures about adolescent substance use and depression. *BMC public health*, 2016, (16), 279 p.

2. Cho S.M., Park C., Song M. The evolution of social health research topics: A data-driven analysis. *Social Science & Medicine*, 2020, (265), p. 113299.

3. Mo Y., Kontonatsios G., Ananiadou S. Supporting systematic reviews using LDA-based document representations // *Systematic Reviews*. 2015 (4), p. 172.

4. PubMed: National Library of Medicine [Online]. Available at: https://pubmed.ncbi.nlm.nih.gov/, free. (Accessed: December 16, 2021).

5. Onan A. Biomedical Text Categorization Based on Ensemble Pruning and Optimized Topic Modelling // *Computational and Mathematical Methods in Medicine*, 2018. pp. 1–22.

6. Sedova A., Mitrofanova O. *Tematičeskoe modelirovanie russkojazyčnyh tekstov s oporoj na lemmy i leksičeskie konstrukcii* [Thematic modeling of Russian-language texts based on lemmas and lexical constructions] *Computational Linguistics and Computational Ontologies*, 2018, pp. 132–144 (in Russ.).

7. Gan J., Qi Y. Selection of the Optimal Number of Topics for LDA Topic Model-Taking Patent Policy Analysis as an Example. *Entropy* (Basel, Switzerland), 2021, no. 10 (23), p. 1301.

8. Chandrasekaran R., Mehta V., Valkunde T., Moustakas E. Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. *Journal of Medical Internet Research*, 2020, no. 10 (22), p. e22624.

9. Yıldırım A., Üsküdarlı S., Özgür A. Identifying Topics in Microblogs Using Wikipedia // PloS one. 2016, no. 3 (11), p. e0151885.

10. Rehurek R., Sojka P. Gensim–python framework for vector space modelling // NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011. no. 2 (3), pp. 46–49.

11. Vafin R.R. *Aktual'nye problemy v oblasti izvlečenija znanij iz professional'nyh medicinskih tekstov s primeneniem intellektual'nogo analiza teksta* [Actual problems in the field of knowledge extraction from professional medical texts using text mining]. *Bulletin of Bashkir State Medical University*, 2019, pp. 72–75 (in Russ.).

12. Min J.-Y., Song S.-H., Kim H., Min K.-B. Mining Hidden Knowledge About Illegal Compensation for Occupational Injury: Topic Model Approach // *JMIR Medical Informatics*. 2019, no. 3 (7), p. e14763.

13. Wang H., Miller D. Improved Parsimonious Topic Modeling Based on the Bayesian Information Criterion // *Entropy* (Basel, Switzerland). 2020, no. 3 (22), p. 326.

14. Liu L., Tang L., Dong W., Yao S., Zhou W. An overview of topic modeling and its current applications in bioinformatics // SpringerPlus. 2016, no. 1 (5), p. 1608.

15. Al Moubayed N., McGough S., Awwad Shiekh Hasan B. Beyond the topics: how deep learning can improve the discriminability of probabilistic topic modelling // *PeerJ Computer Science*. 2020, (6), p. e252.

Sergey A. Zemlyansky
Postgraduate student, Tomsk State University (NI TSU)
36, Lenin pr., Tomsk, Russia, 634050
Phone: +7-953-922-49-58
Email: qoelky@gmail.com

Sergey V. Axyonov
Candidate of Science in Engineering, Assistant Professor,
Department of Information Technology,
School of Engineering Information Technology and Robotics
Tomsk Polytechnic University (NI TPU)
30, Lenin pr., Tomsk, Russia, 634050
Phone: +7-913-887-47-90
Email: axyonov@tpu.ru

Ivan A. Lyzin
Postgraduate student, Department of Information Technology,
School Engineering Information Technology and Robotics,
NI TPU
30, Lenin pr., Tomsk, Russia, 634050
ORCID: 0000-0003-2827-441X
Phone: +7-923-498-70-30
Email: Lyzin@tpu.ru

Olga G. Berestneva
Doctor of Science in Engineering, Professor,
Department of Information Technology,
School of Engineering Information Technology
and Robotics, NI TPU
30, Lenin pr., Tomsk, Russia, 634050
ORCID: 0000-0002-4243-0637
Phone: +7-913-106-19-94
Email: ogb6@yandex.ru