
АВТОМАТИЧЕСКОЕ ВЫДЕЛЕНИЕ ТЕРМИНОВ ДЛЯ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Никитина Мария Александровна
nikitina.mariia@phystech.edu

Консультант: Потапова Полина Сергеевна
potapov.polina@gmail.com

Эксперт: Доктор ф-м наук, Воронцов Константин Вячеславович

30 марта 2023 г.

АННОТАЦИЯ

В данной статье рассматривается задача автоматического выделения терминов в коллекции документов. Новые научные термины появляются каждый день. Ручное извлечение терминов с привлечением узкоспециализированных специалистов является трудозатратным. Цель настоящей работы — обнаружение таких терминов в коллекциях документов в автоматическом режиме. Для решения данной задачи используется метод выделения коллокаций (TopMine) в сочетании с модульной технологией тематического моделирования (с использованием библиотеки BigARTM) и современные методы, основанные на нейросетевых моделях языка. Производится сравнение рассматриваемых решений.

Ключевые слова: тематическое моделирование · TopMine · BigARTM · Automatic Term Extraction

1 Введение

Поиск научных терминов в коллекции документов вручную практически невозможен из-за слишком больших объёмов работы. Для экономии ресурсов и времени предлагается рассмотреть задачу автоматического выделения терминов. К её решению можно подойти с разных сторон. Например, использовать сочетание метода выделения коллокаций с технологией математического моделирования [3]. *Коллокация* — слово или словосочетание, имеющее признаки синтаксически и семантически целостной единицы

Тематическое моделирование – это технология обработки естественного языка, направленная на определение тем, к которым относится текстовый документ из коллекции, и какие слова каждую тему образуют. Иначе говоря, тематическая модель осуществляет *мягкую кластеризацию*, выбирая для документа кластеры-темы.

Вероятностная тематическая модель определяет вероятности тем в каждом документе и вероятности слов в каждой теме. Большим отличием такой модели от глубоких нейронных сетей типа BERT [2] или GPT-3 [1] является простота организации и свойство интерпретируемости в ущерб качеству предсказания вероятности появления слов в документе. Векторное представление тяжёлой нейросети всё ещё не удалось интерпретировать, в то время как тематический эмбединг – это вектор вероятностей тем.

Новизной данной статьи является сравнение этих двух подходов. Нейросеть используется готовая. Для построения же тематической модели требуется подбор *регуляризаторов* – критериев, учитывающих специфические особенности данных или предметной области, от подбора которых значительно зависит качество определения основных тем документов. В данной работе используется модель *аддитивной регуляризации тематической модели*, ARTM [7]. Для построения тематической модели с аддитивной регуляризацией используется библиотека BigARTM [4] с открытым кодом.

Перед выполнением кластеризации необходимо выделить из коллекции документов ключевые слова и словосочетания и отбросить те, что не несут основной смысловой нагрузки. Поиск составных терминов является нетривиальной и трудоёмкой задачей. Для её решения используется метод поиска коллокаций TopMine, использующий информацию о частоте и совстречаемости слов в коллекции [8].

С учётом интерпретируемости и простоты тематическая модель является хорошей заменой нейросети. Предшествующие исследования предлагаемого подхода показали хорошие результаты как по полноте, так и по вычислительной эффективности. В работе сравниваются тематическая модель и сложная нейросетевая модель, анализируется их качество для рассматриваемой задачи.

2 Постановка задачи

Основная задача – построение модели АТЕ (Automatic Term Extraction) для автоматического выделения словосочетаний, являющихся терминами предметной области, в текстах научных статей. Предлагается использовать эффективные методы выделения коллокаций и тематические модели для определения «тематичности» словосочетания. Модель должна обучаться без учителя.

Для решения поставленной задачи применяются алгоритмы поиска коллокаций TopMine с последующей фильтрацией по критерию тематичности, подбор гиперпараметров тематической модели и критерия тематичности.

Задача называется корректно поставленной по Адамару, если её решение существует, единственно и устойчиво. В общем случае построение тематической модели – некорректно поставленная задача по Адамару, поэтому её нужно дополнить регуляризаторами.

1. Как вариант, можно уже добавить GPT-4, тем более у них технический отчет вышел, на который можно сослаться
2. я бы в скобках так же пояснил термин на русском языке
3. Можно дать ссылку на TopMine. Вообще у Вас в тексте TopMine несколько раз упоминается, но нигде не описывается. Лучше кратко описать.
4. Уже писал, возможно имеет смысл дать ссылку на обоснование некорректности задачи тематического моделирования. В любом случае, сейчас связи между абзацами нет. Нужно как-то связать текст.

выборки?

Таблица 1: Описание датасета ACL RD-TEC

Датасет	Год	Количество документов	Количество слов	Количество терминов
ACL RD-TEC	2014	10,922	36,729,513	82,000

В практических задачах автоматической обработки текстов существует очень много критериев и ограничений.

Пусть p_{wd} – вероятность появления термина w в документе d , φ_{wt} – вероятность того, что терм w относится к теме t , θ_{td} – вероятность встречи темы t в документе d . Тогда $P = (p_{wd})_{W \times D}$ – матрица частот термов в документах, $\Phi = (\varphi_{wt})_{W \times T}$ – матрица термов тем, $\Theta = (\theta_{td})_{T \times D}$ – матрица тем документов. W , D , T – множества всех термов, документов и тем соответственно.

Аддитивная регуляризация тематических моделей основана на максимизации логарифма правдоподобия и регуляризаторов $R_i(\Phi, \Theta)$ с неотрицательными коэффициентами регуляризации τ_i , $i = 1, \dots, k$ [7]:

$$\sum_{d \in D} \sum_{w \in W} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}; \quad R(\Phi, \Theta) = \sum_{i=1}^k \tau_i R_i(\Phi, \Theta); \quad (1)$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w \in W} \varphi_{wt} = 1; \quad \varphi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0. \quad (2)$$

Решается задача нахождения разложения $P = \Phi\Theta$ при достижении максимума. Для её решения применяется ЕМ-алгоритм. Таким образом, основной проблемой построения тематической модели становится поиск регуляризаторов $R_i(\Phi, \Theta)$, подходящих под нашу задачу поиска терминов в коллекции документов.

3 Вычислительный эксперимент

Для обучения модели используется открытый датасет ACL RD-TEC [6], в котором собраны статьи на английском языке с 1965 по 2006 год из области компьютерной лингвистики. Его описание представлено в таблице 1. Для проведения эксперимента из него удаляются документы, содержащие менее 20 терминов. В результате получается датасет из 9,095 статей. Распределение терминов по документам представлено на графике 1.

Документы представлены в текстовом формате и были получены путём считывания текста из pdf-файлов. Для проведения первого эксперимента из статей удаляются числа, заголовки и ссылки на литературу, затем эксперимент разделяется на два направления: выделение терминов с помощью библиотеки BigARTM и поиск кандидатов в термины с помощью TopMine. Поиск терминов сначала осуществляется в предположении, что они состоят из одного слова.

5. При таком заявлении точно нужны ссылки на примеры регуляризации
9 (поздно увидел, поэтому нумерация поехала): эксперимент идет после теории

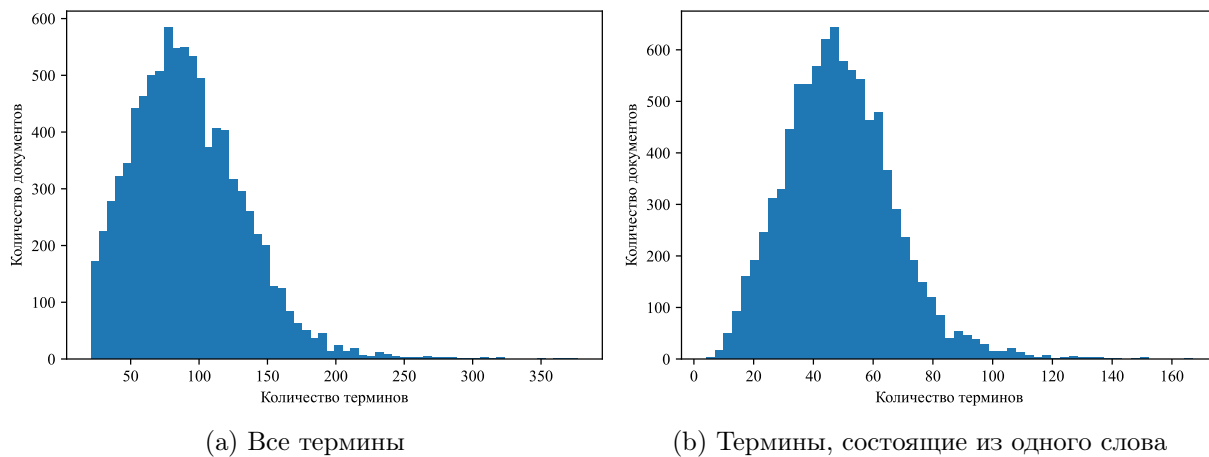


Рис. 1: Распределение терминов по документам

Для использования BigARTM необходимо привести слова к нормальной форме. Так как датасет состоит из статей, написанных на английском языке, для этого используется *стемминг* – отбрасывание окончаний и других изменяемых частей слова. Также для улучшения результата из текста выбрасываются стоп-слова, например: «the», «a», «however». Затем полученные документы собираются в файл формата Vowpal Wabbit, который уже обрабатывается с помощью библиотеки BigARTM.

Для подбора подходящих гиперпараметров эксперимент запускается с различными комбинациями гиперпараметров *сглаживания*, *разреживания* и *декоррелирования*. Подробное их описание есть в [5]. Сглаживание применяется для фоновых тем, куда собираются слова, не имеющие определённой темы. Сами по себе они не несут никакой смысловой нагрузки. Обычно выбираются 1-2 фоновые темы. Разреживание применяется для предметных тем. В отличие от гиперпараметра сглаживания влияние разреживания рекомендуется увеличивать постепенно в процессе обучения модели. Декорреляция отвечает за степень различия между темами.

Результат работы алгоритма – матрица Φ , состоящая из вероятностей принадлежности некоторого термина к некоторой теме. Следует выделить из этих термов термины. В большинстве случаев термин отличается от обычного слова или словосочетания тем, что имеет большую вероятность появления только в 1-2 темах. Таким образом выделяются все кандидаты в термины для коллекции документов.

Вторая часть эксперимента – выделение коллокаций с помощью TopMine. После обработки документов без чисел, заголовков и литературы получается список кандидатов в термины для каждого документа отдельно.

После пересечения списка терминов, полученных с помощью BigARTM, с коллокациями, выделенными алгоритмом TopMine, получается окончательный список терминов для каждого документа.

6. Размер текста на графиках должен соответствовать размеру текста в документе
7. Если это эксперимент, то здесь следует писать не сколько тем выбирают обычно, а сколько тем использовалось в эксперименте
8. Если обе части эксперимента независимо друг от друга несут полезную информацию, имеет смысл разбить на подсекции, чтобы было удобнее читать

10

4 Теоретическая часть

Для решения задачи (1)-(2) используется ЕМ-алгоритм.

Пусть $R_i(\Phi, \Theta)$ – непрерывно дифференцируемы. Точка локального экстремума задачи (1)-(2) удовлетворяет системе уравнений, если из решения удалить нулевые столбцы матриц Φ и Θ [7]:

$$p_{td\omega} = \text{norm}_{t \in T}(\varphi_{\omega t} \theta_{td}); \quad (3)$$

$$\varphi_{\omega t} = \text{norm}_{\omega \in W} \left(n_{\omega t} + \varphi_{\omega t} \frac{\partial R}{\partial \varphi_{\omega t}} \right); \quad n_{\omega t} = \sum_{d \in D} n_{d\omega} p_{td\omega}; \quad (4)$$

$$\theta_{td} = \text{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{\omega \in d} n_{d\omega} p_{td\omega}, \quad (5)$$

где $p_{td\omega} = p(t|d, \omega)$,

$n_{td\omega}$ – число троек, в которых терм ω документа d связан с темой t ,

$n_{\omega t} = \sum_d n_{td\omega}$ – число троек, в которых терм ω связан с темой t ,

$n_{d\omega}$ – число вхождений терма ω в документ d ,

$n_{td} = \sum_{\omega} n_{td\omega}$ – число троек, в которых терм документа d связан с темой t .

Оператор norm определяется так:

$$\text{norm}_{i \in I}(x_i) = \frac{(x_i)_+}{\sum_{k \in I} (x_k)_+}, \quad \text{где } (x)_+ = \max\{0, x\}.$$

Обобщённый регуляризатор сглаживания-разреживания:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{\omega \in W} \beta_{\omega t} \ln \varphi_{\omega t} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \quad (6)$$

Подставив этот регуляризатор в (4)-(5), получим для М-шага:

$$\varphi_{\omega t} = \text{norm}_{\omega \in W}(n_{\omega t} + \beta_{\omega t}) \quad (7)$$

$$\theta_{td} = \text{norm}_{t \in T}(n_{td} + \alpha_{td}) \quad (8)$$

При положительных β_{ω} , α_{td} регуляризатор увеличивает правдоподобие и происходит сглаживание, при отрицательных – уменьшение правдоподобия и разреживание.

Декоррелирование тем:

10. Анализ свойств предложенного метода (название рабочее. Консультанты подскажут название лучше)

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{\omega \in W} \varphi_{\omega t} \varphi_{\omega s} \quad (9)$$

Подставив этот регуляризатор в (4)-(5), получим для М-шага:

$$\theta_{td} = \text{norm}_{\omega \in W} (n_{\omega t} - \tau \varphi_{\omega t} \sum_{s \in T \setminus t} \varphi_{\omega s}) \quad (10)$$

Систему (3)-(5) можно решить методом простых итераций, чередуя Е-шаг (3) и М-шаг (4)-(5).

5 Анализ ошибки

91

Для анализа ошибки используются Precision – точность, Recsll – полнота, а также F1 – их среднее гармоническое:

$$\text{Precision} = \frac{TP}{TP + FP}; \quad \text{Recall} = \frac{TP}{TP + FN}; \quad (11)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (12)$$

где TP – истинно-положительное решение, FP – ложноположительное решение, FN – ложноотрицательное решение.

В таблице 2 представлены некоторые результаты (от худшего до лучшего), полученные без выделения фоновых тем. В таблице 2 представлены результаты, полученные с выделением 2 фоновых тем.

Кол-во тем – количество тем, на которые разбиваются слова в датасете. Во второй таблице количество тем везде равно 150. Их меньше, чем в первой из-за выделения фоновых тем. Это удобно, так как размер выходной матрицы вероятностей Φ сильно уменьшается.

τ_{dec} – значение коэффициента декорреляции. В случае выделения фоновых тем он задаётся отдельно для предметных и отдельно для фоновых тем.

τ_{Φ} – значение коэффициента сглаживания-разреживания матрицы Φ . Во второй таблице представлено его значение для предметных и фоновых тем отдельно.

τ_{Θ} – значение коэффициента сглаживания-разреживания матрицы Θ . Во второй таблице представлено его значение для предметных и фоновых тем отдельно.

Порог – пороговое значение вероятности для отбора кандидатов в термины.

Кол-во тем для термина – разрешённое количество тем, для которых термин имеет большую вероятность. Заметно, что при наличии выделения фоновых тем для лучшего результата требуется меньшее значение этого параметра.

11. Лучше подсекцией эксперимента

Таблица 2: Результаты работы алгоритма для нахождения терминов, состоящих из одного слова, без выделения фоновых тем

N	Кол-во тем	τ_{dec}	τ_{Φ}	τ_{Θ}	Порог	Кол-во тем для термина	Precision	Recall	F1
1	200	0.025	0.025	0.025	1,00E-04	1 — 2	0.023	0.031	0.026
2	200	0.025	0.025	0.025	5,00E-04	1 — 4	0.058	0.176	0.087
3	200	0.025	0.025	0.025	7,00E-04	1 — 5	0.078	0.270	0.121
4	200	0.025	0.025	0.025	1,00E-03	1 — 6	0.098	0.354	0.154
5	200	0.025	0.025	-0.25	1,00E-03	1 — 6	0.091	0.332	0.142
6	200	0.025	0.1	0.025	5,00E-04	1 — 4	0.104	0.345	0.160
7	200	0.025	0.1	0.025	1,00E-03	1 — 6	0.156	0.432	0.229
8	200	0.025	0.1	0.1	1,00E-03	1 — 6	0.155	0.432	0.228

Таблица 3: Результаты работы алгоритма для нахождения терминов, состоящих из одного слова, с выделением фоновых тем

N	τ_{dec}	τ_{Φ}	τ_{Φ} (фон)	τ_{Θ}	τ_{Θ} (фон)	Порог	Кол-во тем для термина	Precision	Recall	F1
1	0.1	-0.1	0.1	0.1	-0.1	0.01	1 — 4	0.204	0.358	0.260
2	0.1	-0.1	0.1	0.1	-0.1	0.3	1 — 2	0.553	0.096	0.164
3	0.1	-0.1	0.2	0.1	-0.1	0.05	1 — 4	0.458	0.213	0.290
4	0.01	-0.1	0.1	-0.1	0.2	0.03	1 — 2	0.395	0.255	0.310
5	0.01	-0.1	0.2	-0.1	0.2	0.03	1 — 2	0.407	0.263	0.319
6	0.01	-0.1	0.05	-0.1	0.2	0.03	1 — 2	0.389	0.240	0.297
7	0.01	-0.2	0.1	-0.1	0.2	0.03	1 — 2	0.363	0.266	0.307
8	0.01	-0.2	0.2	-0.1	0.2	0.03	1 — 2	0.365	0.268	0.309
9	0.01	-0.2	0.05	-0.1	0.2	0.03	1 — 2	0.369	0.270	0.312
10	0.01	-0.05	0.1	-0.1	0.2	0.03	1 — 2	0.415	0.243	0.306
11	0.01	-0.05	0.2	-0.1	0.2	0.03	1 — 2	0.417	0.254	0.316
12	0.01	-0.05	0.05	-0.1	0.2	0.03	1 — 2	0.394	0.230	0.291
13	0.01	-0.1	0.2	-0.1	0.2	0.03	1 — 4	0.413	0.286	0.338
14	0.01	-0.1	0.2	-0.1	0.1	0.03	1 — 4	0.413	0.286	0.338
15	0.01	-0.1	0.2	-0.2	0.1	0.03	1 — 4	0.426	0.281	0.338
16	0.01	-0.1	0.2	-0.2	0.2	0.03	1 — 4	0.424	0.281	0.338
17	0.01	-0.1	0.2	-0.1	0.05	0.03	1 — 4	0.424	0.281	0.338
18	0.01	-0.1	0.2	-0.05	0.1	0.03	1 — 4	0.402	0.276	0.327
19	0.01	-0.1	0.2	-0.05	0.2	0.03	1 — 4	0.400	0.276	0.327
20	0.01	-0.1	0.2	-0.05	0.05	0.03	1 — 4	0.402	0.276	0.327
21	0.1	-0.1	0.1	-0.1	0.1	0.03	1 — 4	0.498	0.229	0.313

12. Лучшие показатели выделяют жирным

Лучший результат, полученный при вариации гиперпараметров: $F1 = 0.338$. Отметим отдельно результат работы алгоритма TopMine без использования модульного тематического моделирования – таблица 4. Параметр F1 в 3 раза меньше, чем у лучшего результата объединения двух алгоритмов.

Таблица 4: Результаты работы алгоритма TopMine

Precision	Recall	F1
0.063	0.946	0.117

Пример зависимости значения F1 от значения гиперпараметров приведён на графике 2.

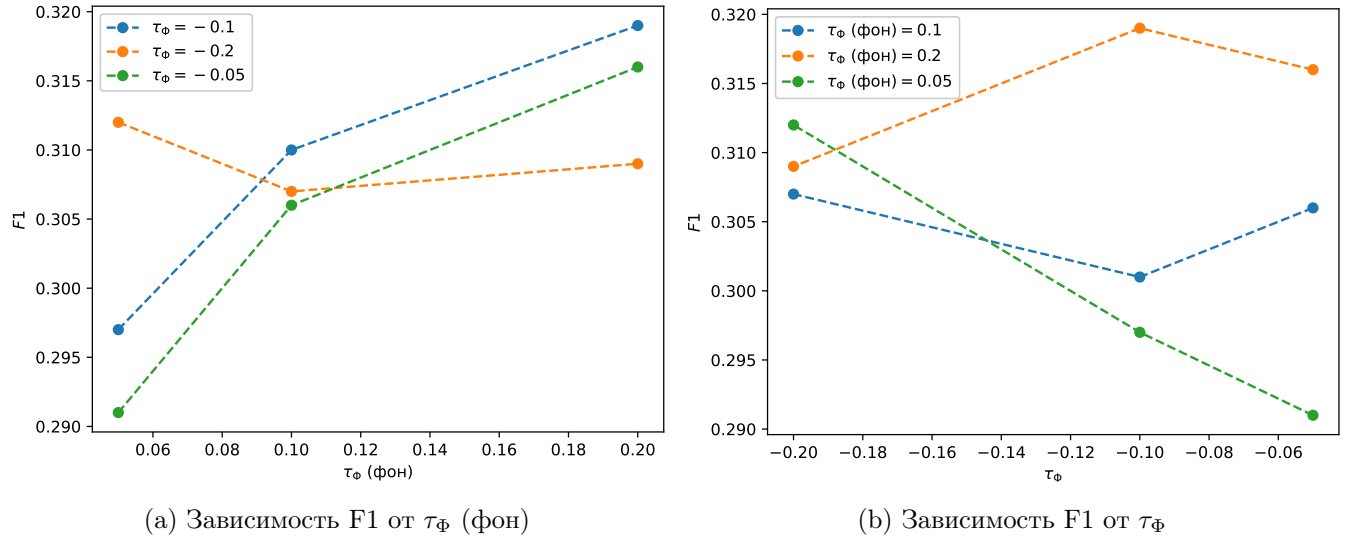


Рис. 2: Зависимость F1 от τ_Φ (фон) и τ_Φ при $\tau_\Theta = -0.1$, τ_Θ (фон) = 0.2

6 Заключение

Список литературы

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

- [3] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, November 2014.
- [4] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *Communications in Computer and Information Science*, pages 370–381. Springer International Publishing, 2015.
- [5] Konstantin Vorontsov and Anna Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In *Communications in Computer and Information Science*, pages 29–46. Springer International Publishing, 2014.
- [6] Behrang Q. Zadeh and Siegfried Handschuh. The ACL RD-TEC: A dataset for benchmarking terminology extraction and classification in computational linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology (Computerm)*. Association for Computational Linguistics and Dublin City University, 2014.
- [7] Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект bigartm. 2020.
- [8] Николай Шаталов. Методы обучения без учителя для автоматического выделения составных терминов в текстовых коллекциях. 2019.