НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2023 Том 23 № 1 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2023 Vol. 23 No 1 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2023-23-1-112-120
УДК 004.912

Программный фреймворк для оптимизации гиперпараметров тематических моделей с аддитивной регуляризацией

Мария Андреевна Ходорченко[1]✉, Николай Алексеевич Бутаков[2], Денис Александрович Насонов[3], Михаил Юрьевич Фирулик[4]

[1,2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[4] ООО «Оператор Газпром ИД», Санкт-Петербург, 191028, Российская Федерация

[1] mariyaxod@yandex.ru✉, https://orcid.org/0000-0001-5446-5311
[2] alipoov.nb@gmail.com, https://orcid.org/0000-0002-2705-1313
[3] denis.nasonov@gmail.com, https://orcid.org/0000-0002-2740-0173
[4] firulik@gmail.com, https://orcid.org/0000-0001-7114-1052

Аннотация

Предмет исследования. Обработка неструктурированных данных, таких как тексты на естественном языке, является одной из актуальных задач при разработке интеллектуальных продуктов. В свою очередь, тематическое моделирование как метод работы с неразмеченными и частично размеченными текстовыми данными активно используется для анализа корпусов документов и создания векторных представлений. В связи с этим особенно важно обучение качественных тематических моделей за короткое время, что возможно с помощью предложенного фреймворка. **Метод.** Разработанный фреймворк реализует эволюционный подход к оптимизации гиперпараметров моделей с аддитивной регуляризацией и высокими результатами по метрикам качества (когерентность, NPMI). Для уменьшения вычислительного времени представлен режим работы с суррогатными моделями, который обеспечивает ускорение вычислений до 1,8 раз без потери качества. **Основные результаты.** Эффективность фреймворка продемонстрирована на трех наборах данных с разными статистическими характеристиками. Получены результаты, превосходящие аналогичные решения в среднем на 20 % по когерентности и 5 % по качеству классификации для двух из трех наборов. Создана распределенная версия фреймворка для проведения экспериментальных исследований тематических моделей. **Практическая значимость.** Полученный фреймворк может быть использован пользователями без специальных знаний в области тематического моделирования, благодаря выстроенному пайплайну работы с данными. Результаты работы могут применяться исследователями для проведения анализа тематических моделей и расширения функционала.

Ключевые слова

AutoML фреймворк, тематическое моделирование, неструктурированные данные, аддитивная регуляризация, эволюционный подход, суррогатные модели

Благодарности

Исследование выполнено при финансовой поддержке Российского научного фонда, проект № 20-11-20270.

Ссылка для цитирования: Ходорченко М.А., Бутаков Н.А., Насонов Д.А., Фирулик М.Ю. Программный фреймворк для оптимизации гиперпараметров тематических моделей с аддитивной регуляризацией // Научно-технический вестник информационных технологий, механики и оптики. 2023. Т. 23, № 1. С. 112–120. doi: 10.17586/2226-1494-2023-23-1-112-120

Software framework for hyperparameters optimization of models with additive regularization

Maria A. Khodorchenko[1]✉, Nikolay A. Butakov[2], Denis A. Nasonov[3], Mikhail Yu. Firulik[4]

[1,2,3] ITMO University, Saint Petersburg, 197101, Russian Federation
[4] LLC "Operator Gazprom ID", Saint Petersburg, 191028, Russian Federation

[1] mariyaxod@yandex.ru✉, https://orcid.org/0000-0001-5446-5311
[2] alipoov.nb@gmail.com, https://orcid.org/0000-0002-2705-1313
[3] denis.nasonov@gmail.com, https://orcid.org/0000-0002-2740-0173
[4] firulik@gmail.com, https://orcid.org/0000-0001-7114-1052

Abstract
The processing of unstructured data, such as natural language texts, is one of the urgent tasks in the development of intelligent products. In turn, topic modeling as a method of working with unmarked and partially marked text data is a natural choice for analyzing document bodies and creating vector representations. In this regard, it is especially important to train high-quality thematic models in a short time which is possible with the help of the proposed framework. The developed framework implements an evolutionary approach to optimizing hyperparameters of models with additive regularization and high results on quality metrics (coherence, NPMI). To reduce the computational time, a mode of working with surrogate models is presented which provides acceleration of calculations up to 1.8 times without loss of quality. The effectiveness of the framework is demonstrated on three datasets with different statistical characteristics. The results obtained exceed similar solutions by an average of 20 % in coherence and 5 % in classification quality for two of the three datasets. A distributed version of the framework has been developed for conducting experimental studies of topic models. The developed framework can be used by users without special knowledge in the field of topic modeling due to the default data processing pipeline. The results of the work can be used by researchers to analyze topic models and expand functionality.

Keywords
AutoML framework, topic modeling, unstructured data, additive regularization, evolutionary approach, surrogate models

Acknowledgements
This research is financially supported by The Russian Science Foundation, Agreement No. 20-11-20270.

For citation: Khodorchenko M.A., Butakov N.A., Nasonov D.A., Firulik M.Yu. Software framework for hyperparameters optimization of models with additive regularization. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2023, vol. 23, no. 1, pp. 112–120 (in Russian). doi: 10.17586/2226-1494-2023-23-1-112-120

Введение

В настоящее время ведется активная разработка и внедрение интеллектуальных продуктов в разнообразные бизнес-процессы и клиентские услуги. При этом большую популярность получает работа с неструктурированными данными, такими как тексты на естественном языке (документы, отзывы, комментарии и т. д.). Для их качественной обработки необходимо проведение разведочного анализа, основным инструментом которого, для текстовых данных, является такой метод обучения без учителя как тематическое моделирование. Использование этого метода позволяет получать скрытые компоненты (под ними понимаются темы), и их представленность в корпусе документов. Таким образом, текст превращается в интерпретируемое представление, которое можно использовать для решения актуальных проблем, таких как построение профилей интересов пользователей [1, 2], извлечение сентимента [3], разделение данных на тематические и структурно близкие подмножества [4], обогащение контекстуализированных представлений документов [5] и другие.

Существующие методы тематического моделирования можно отнести к нескольким большим семействам: на основе матричных разложений (неотрицательная матричная факторизация (Non-negative Matrix Factorization, NMF) [6]), классические вероятностные (например, вероятностный латентный семантический анализ (Probabilistic Latent Semantic Analysis, PLSA) [7] и латентное распределение Дирихле (Latent Dirichlet Allocation, LDA) [8]), с аддитивной регуляризацией (аддитивная регуляризация тематических моделей (Additive Regularization of Topic Models, ARTM) [9]) и нейронные модели [10–12]. Долгое время наиболее предпочтительным методом, показывавшим высокие результаты качества, являлось LDA, но его использование на корпусах документов, обладавших спецификой (например, наличие терминологии или малое количество слов в документе) затруднялось необходимостью модификации аппарата обучения модели [13, 14], что требовало от разработчика соответствующих знаний. В свою очередь, развивающиеся нейронные модели демонстрируют высокую степень приспособленности под заданную оценку качества, но при этом, в случае работы с контекстуализированными представлениями, являются вычислительно затратными. К оптимальным и перспективным вариантам можно отнести подход с аддитивной регуляризацией, который обобщает понятие тематического моделирования и обеспечивает простую оптимизацию с использованием алгоритма максимизации ожидания (expectation-maximization, EM-алгоритм).

При работе с моделями, имеющими гиперпараметры, возникает необходимость в их настройке. Существуют различные программные решения для автоматической оптимизации тематических моделей различных семейств. В genism [15] реализованы «классические методы» тематического моделирования, такие как LDA, и возможность настройки их параметров. Библиотека OCTIS [16] предоставляет возможность

настройки ряда тематических моделей, включая нейронные, с помощью байесовской оптимизации, что обеспечивает достаточно эффективную и быструю работу. К недостаткам можно отнести отсутствие специфичных алгоритмов оптимизации под разные семейства моделей. TopicNet работает с семейством моделей с аддитивной регуляризацией посредством жадного алгоритма настройки. При этом автоматическая настройка осложняется отсутствием единых метрик измерения качества [17].

В данной работе представлен фреймворк для настройки моделей с аддитивной регуляризацией, позволяющий за ограниченное время получать тематические модели высокого качества сразу по ряду существующих метрик.

Основные технологии

Рассмотрим базовые технологии, которые лежат в основе разработанного AutoML фреймворка для задачи настройки моделей тематического моделирования.

BigARTM [18] — библиотека для обучения моделей с аддитивной регуляризацией. Обеспечивает быструю подготовку и инференс из готовых моделей. К недостаткам библиотеки относятся ее сложность для рядового пользователя, в том числе необходимость понимания основных механизмов настройки моделей с аддитивной регуляризацией в связи с отсутствием автоматической оптимизации гиперпараметров.

Эволюционный подход к оптимизации гиперпараметров моделей с аддитивной регуляризацией [19] включает в себя способ представления гиперпараметров в виде стратегии обучения, генетический алгоритм настройки и оценку качества. С целью получения более стабильной работы фреймворка внесены изменения в эволюционный подход.

В инициализацию индивидов добавлены базовые модели без сильной регуляризации, что отображено на рис. 1, *a*, где итерации обучения n_2–n_4 заменены нулями, это позволяет улучшить скорость сходимости модели за счет коррекции направления оптимизации при инициализации сильно разреженных моделей. Остальные гиперпараметры, такие как количество тем — B_n, декорреляция фоновых и основных тем матрицы $\boldsymbol{\varphi}$ — D^B_φ, D^S_φ, сглаживание фоновых тем — P^B_φ, P^B_θ и разреживание основных тем P^S_θ, P^S_φ, сэмплировались в соответствии с правилами, определенными в [19].

Процедура мутации реализована в два шага (рис. 1, *b*). На шаге 1 происходит определение, будет ли произведена мутация индивида с вероятностью m_α, затем каждый элемент из соответствующей категории (которым соответствуют цвета стрелок) мутирует с вероятностью m_β, т. е. заменяется на случайный элемент из категории. На шаге 2 происходит замена параметров мутации с вероятностью m_β на случайное значение из равномерного распределения.

Приведем модифицированную функцию приспособленности, которая обеспечивает разреженность матрицы темы-документы в диапазоне 0,2–0,8, что способствует получению более разнообразных результатов:

$$\alpha(mean(coh_{50}) + \min(coh_{50})),$$

$$\alpha = \begin{cases} 1, \text{ если } 0{,}2 \le Sp_{\boldsymbol{\theta}} \le 0{,}8 \\ 0{,}7, \text{ иначе} \end{cases},$$

где $\boldsymbol{\theta}$ — матрица распределения вероятностей тем над документами; $Sp_{\boldsymbol{\theta}}$ — разреженность матрицы $\boldsymbol{\theta}$.

Суррогатное моделирование для ускорения оптимизации [20]. Фреймворк реализует модуль работы с суррогатными моделями для уменьшения времени поиска решения для больших корпусов данных. В качестве наилучшей суррогатной модели функции фитнеса использована модель на основе гауссовских процессов.

Описание разработанного фреймворка

При разработке фреймворка приняты во внимание следующие правила.

— *Простота использования*. Предлагаемый базовый пайплайн не требует дополнительной настройки со стороны пользователя, так как в нем уже определены все необходимые шаги и заданы гиперпараметры, обеспечивающие хорошее качество «в среднем» на наборах данных с различными ста-

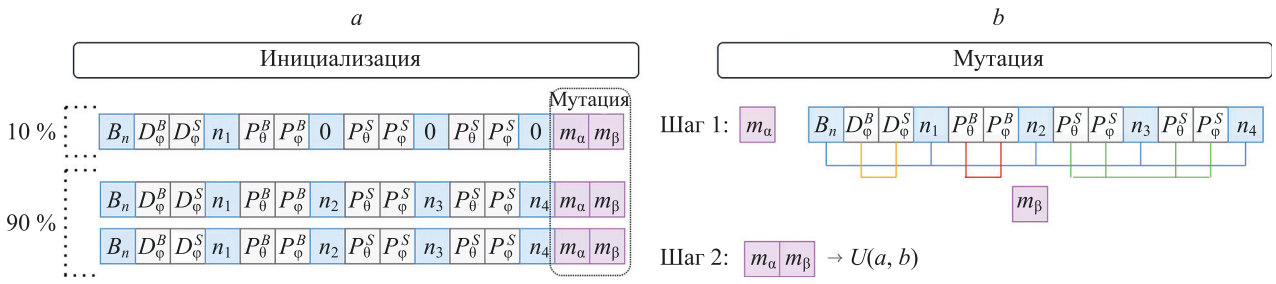

Рис. 1. Инициализация гиперпараметров, где обозначены 10 % базовых индивидов и 90 % индивидов со случайной инициализацией (*a*) и процесс мутации (*b*).

Ячейки хромосом обозначены: оптимизируемые целочисленные (синим цветом) и действительные (серым цветом) гиперпараметры; параметры мутации, изменяющиеся в рамках индивида (фиолетовым цветом)

Fig. 1. Hyperparameters initialization where 10 % of base individuals and 90 % of randomly initialized individuals (*a*) and mutation procedure (*b*) are provided. Blue color indicates integer hyperparameters, white — real ones, violet — mutation parameters that is self-learned with individual development

тистическими характеристиками. Задается только желаемое количество тем.

— *Реализация полного пайплайна, включая предобработку данных*. Часто библиотеки и фреймворки опускают процедуру подготовки данных, оставляя ее за пользователем.
— *Возможность расширения и реализации собственных методов*. В первую очередь предполагается модификация оценок качества.
— *Обеспечение скорости обработки текстов*. Фреймворк предполагает реализацию использования суррогатных моделей, которые позволят повысить скорость оптимизации до 1,8 раз, существенно не влияя на качество [20]. Также исследователям будет предоставлена возможность проведения быстрых экспериментов при использовании распределенной версии фреймворка.

Фреймворк содержит набор модулей, реализующих общий пайплайн: предобработки данных (очищение, лемматизация, подготовка); оценки качества; оптимизации, содержащей реализации эволюционных алгоритмов; суррогатных моделей. При необходимости возможно расширение имеющихся модулей либо реализация новых.

Общий принцип работы фреймворка (рис. 2) можно описать следующим образом. На вход поступает корпус текстовых данных и производится его предобработка с целью очистки и нормализации текстовых данных, а также подготовки данных в формате, требуемом для работы библиотеки BigARTM. Затем происходит определение метрики оценки качества целевой тематической модели (ТМ), которую можно либо оставить предложенной по умолчанию, либо выбрать из существующих. Задаются общие настройки для процедуры поиска оптимальных гиперпараметров (Гиперп), реализуемой библиотекой, такие как используемый эволюционный алгоритм, размер, число итераций и т. п. После этого начинает работу алгоритм поиска гиперпараметров, в течение которого происходят итерации отчета метрик качества для каждого индивида в популяции, представляющего набор параметров и стратегию их применения.

В случае, если используется оригинальная версия предложенного метода (т. е. без суррогатов) каждый набор параметров проходит через процедуру построения тематической модели по задаваемой стратегии (рис. 2, правая часть «Обучение тематической модели») путем обучения тематической модели с первым набором параметров в течении некоторого количества итераций, затем смены на второй набор в стратегии и продолжении обучения ранее созданной модели с новым набором, и т. д. до получения конечной модели. Отметим, что таких сменяющих друг друга наборов параметров может быть несколько или стратегия может состоять только из одного набора (так как значение количества итераций обучения может быть равно нулю).

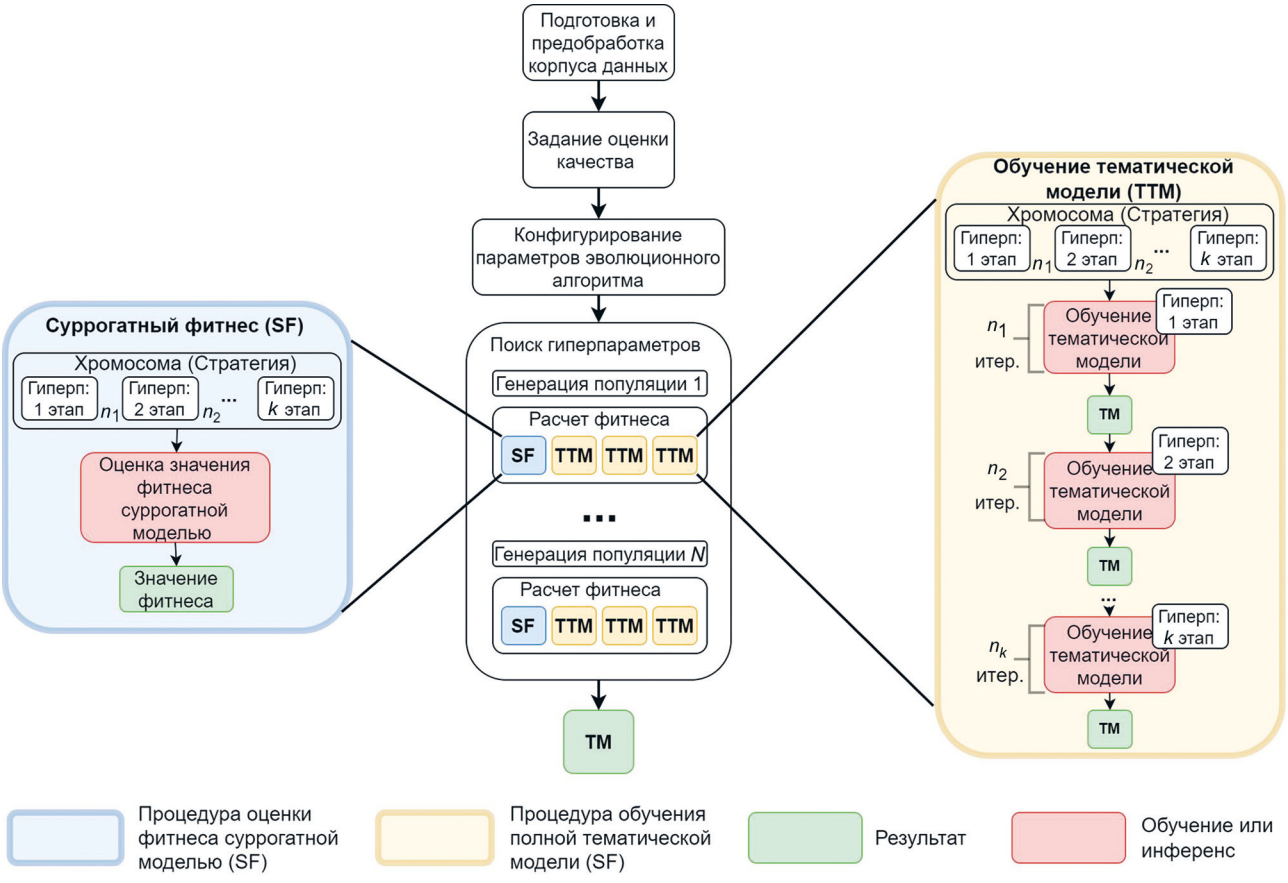

Рис. 2. Общая схема работы фреймворка

Fig. 2. The base scheme of the proposed framework

При использовании модификации метода с применением суррогатов, для некоторого подмножества индивидов в популяции вместо непосредственного вычисления метрики, описанного выше, производится ее прогнозирование с помощью суррогатной модели. Прогнозирование позволяет получить оценку метрики, используя только сами значения гиперпараметров и порядок в стратегии.

По окончании процесса поиска гиперпараметров, эволюционный алгоритм выдает наилучший найденный набор гиперпараметров и стратегию его применения. Фреймворк, в свою очередь, обучает итоговую тематическую модель и возвращает ее пользователю вместе с набором гиперпараметров. При необходимости пользователь может расширить функционал и добавить собственные методы расчета.

Эффективность использования разработанного фреймворка можно повысить за счет ускорения построения тематических моделей. Заметим, что популяционные подходы можно успешно распараллелить на этапе вычисления фитнеса для отдельных особей. В связи с этим предложено распределенное расширение фреймворка за счет включения возможности расчета популяции особей на наборе вычислительных узлов (рис. 3).

Принцип работы распределенной версии имеет несколько значимых отличий от базового варианта. В данном случае расчеты производятся на наборе вычислительных узлов, где один из них ведущий, а остальные — расчетные.

Подготовленные данные на первом шаге обработки (Подготовка и предобработка корпуса данных) сохраняются в хранилище (Storage), доступном на всех узлах вычислительного кластера для последующего использования при расчетах метрики качества. На этапе поиска гиперпараметров текущая популяция эволюционного алгоритма рассылается на вычислительные узлы для подсчета фитнеса с помощью построения тематических моделей согласно стратегии и гиперпараметрам каждого из индивидов. Результаты оценки метрики качества затем передаются обратно на ведущий узел. Обученные модели, их метрики и логи могут сохраняться для всех индивидов или выборочно в хранилище результатов (MLFlow и Storage).

Если используется модификация метода с суррогатными моделями, прогнозирование метрики качества для фитнеса производится локально, без пересылки на удаленные узлы, так как данная процедура не занимает много времени из-за простоты суррогатных моделей. Однако в будущем такой порядок работы может быть изменен, если в этом появится необходимость ввиду роста вычислительной сложности суррогатных моделей.

Основные компоненты распределенного фреймворка (рис. 4): autotmlib-distr – интеграционная прослойка,

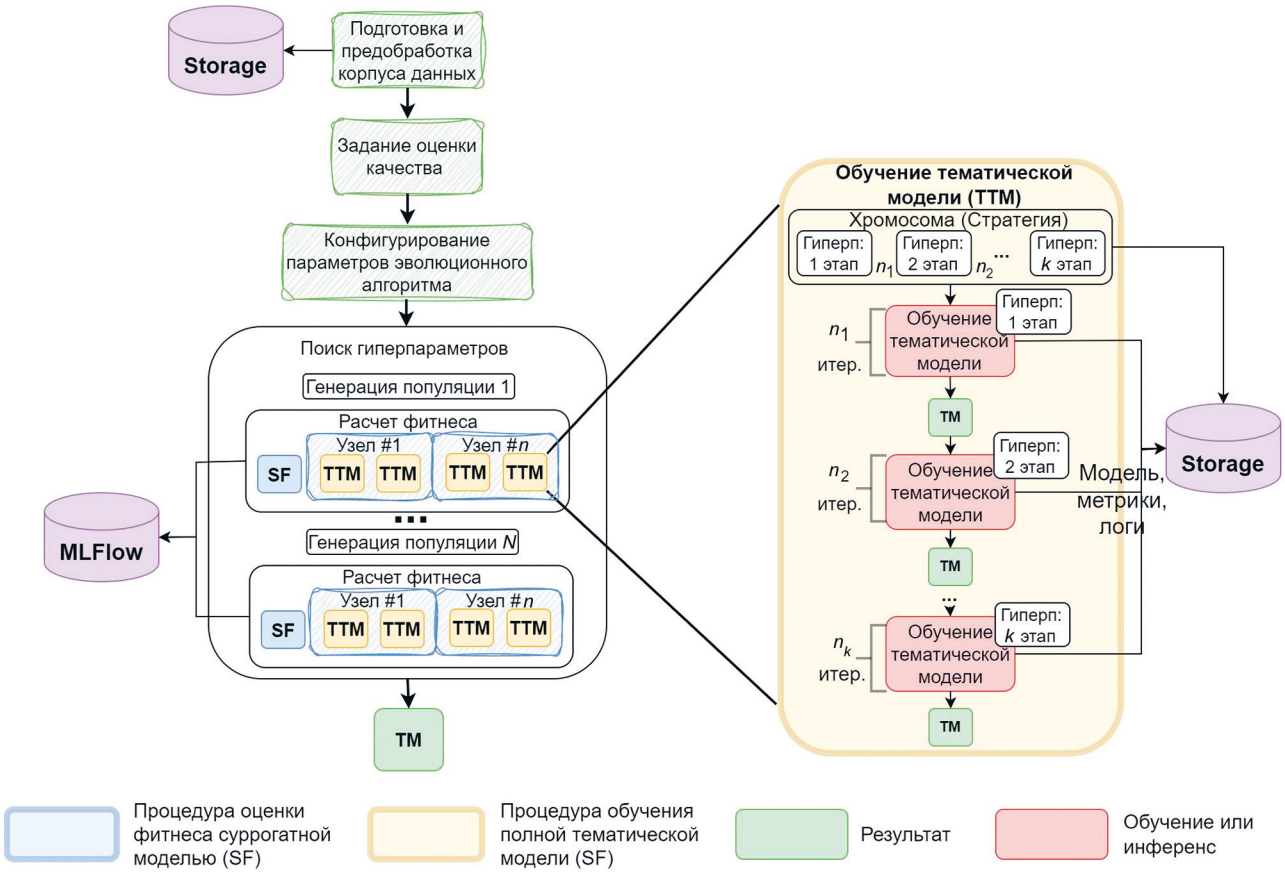

Рис. 3. Схема работы распределенной версии фреймворка.

Шаги предобработки и конфигурирования выполняются на ведущем узле, вычисление фитнеса — на расчетных узлах

Fig. 3. Workflow of the distributed framework version. The preprocessing and configuration steps are performed on the master node; the fitness calculation is performed on the calculation nodes

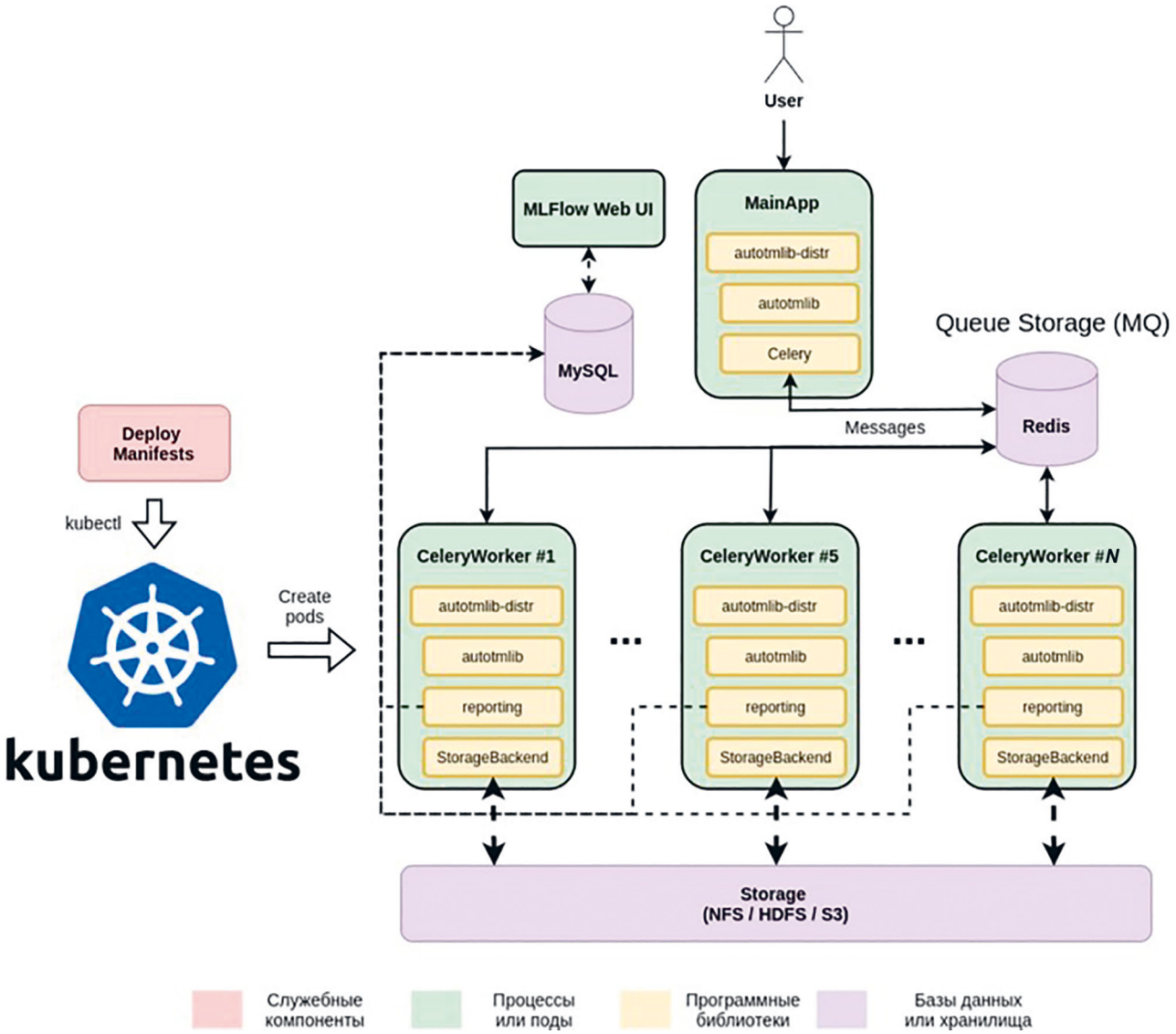

Рис. 4. Схема общей архитектуры фреймворка

Fig. 4. Scheme of the framework architecture

содержащая основной код фреймворка и представляющая вспомогательный материал репортинга и сохранения данных; Celery — брокер задач на обучении экземпляров тематических моделей для подбора гиперпараметров; Queue Storage — хранилище очереди задач на основе базы данных Redis для брокера Celery; MLFlow — система регистрации результатов расчетов и процессов оптимизации гиперпараметров; Storage — долговременное хранилище для промежуточных и конечных тематических моделей; менеджер ресурсов Kubernetes, обеспечивающий выделение ресурсов для обработчиков (CeleryWorker) брокера Celery; модуль манифестов Deploy manifests, отвечающий за развертывание и конфигурирование всех остальных компонентов фреймворка в кластерных средах под управлением Kubernetes.

Включение в архитектуру фреймворка компонентов Storage и MLFlow позволило удобно и эффективного провести регистрацию результатов вычислений фитнеса: построенные модели; полученные наборы тематик; использованный для построения набор гиперпараметров и стратегию их применения (как хромосому); метрики качества; итоговое значение фитнеса; логи обучения тематической модели, и другую служебную информацию.

Таким образом, разработанный фреймворк содержит все необходимые компоненты для проведения оптимизации гиперпараметров тематических моделей, включая предобработку данных. Предложен распределенный вариант библиотеки для проведения экспериментальных исследований.

Исследование эффективности фреймворка

Для сравнения с существующими решениями определим основной критерий, по которому выполним сравнение фреймворков и библиотек — достижение высокого качества за ограниченное время оптимизации.

Для экспериментов выберем три набора данных с различными характеристиками: датасет Lentaru, с

новостными постами, собранными за 20 лет; Amazon еда — отзывы о продукции, реализованные Amazon и 20 newsgroups — классический датасет с 20 категориями постов на разные темы. У каждого из датасетов отобран сэмпл в 10 000 документов.

Сравнение произведено с библиотеками genism (настройка алгоритма LDA), Octis (оптимизация гиперпараметров контекстуализированной модели (Contextualized TM), а именно, использовались веса предобученной модели RoBERTa для русского и английского языков) и topicNet (использовался предлагаемый базовый пайплайн). Для сравнения выбрана базовая версия фреймворка с использованием суррогатного моделирования фитнеса.

Когерентность (Coh) [21] и нормализованная мера попарной взаимной информации (NPMI) [22] вычислены для 20 наиболее вероятных токенов в каждой теме. Для расчета качества полученных моделей использованы оценки качества тем, имеющих доказанную корреляцию с человеческим восприятием (когерентность и NPMI). Вычислено качество классификации при использовании полученных представлений документов (для датасета Amazon метки классов отсутствуют, поэтому качество классификации не замерялось). Для оценки использован метод *k*-ближайших соседей с 5 соседями и рассчитана средняя взвешенная f1-мера на 5 фолдах (Cls (f1)).

Из таблицы видно, что разработанный фреймворк показал лучшие результаты даже при ограничении времени оптимизации. При этом наблюдается соблюдение баланса между качеством тем и эффективностью применения эмбеддингов для прикладной задачи классификации. Отметим, что алгоритмы TopicNet и Gensim за выделенное в 3 мин время сходятся к стабильному решению, в то время как предложенный фреймворк еще может повысить качество до 10 % при обучении на большем количестве итераций.

На рис. 5 представлен пример визуализации с помощью метода t-SNE (стохастическое вложение соседей с t-распределением) получаемых тематических кластеров при обучении на 25 темах, где отображено их соответствие с метками классов датасета. Заметно, что существуют как тематически обособленные кластеры, так и близкие друг к другу темы.

Таблица. Полученное разными фреймворками качество при среднем времени оптимизации 3 мин. Результаты представлены в безразмерных единицах

Table. Quality obtained by different frameworks with average optimization time (3 minutes). The results are presented in dimensionless units

Фреймворк	Датасет							
	20 newsgroups (10 тем)			Lentaru (50 тем)			Amazon еда (25 тем)	
	Coh	NPMI	Cls (f1)	Coh	NPMI	Cls (f1)	Coh	NPMI
TopicNet	–2,50	0,03	0,05	–4,93	0,02	0,39	–2,34	–0,10
Gensim (LDA)	–3,20	0,01	0,36	–6,66	–0,05	0,48	–0,21	–0,09
Octis (Contextualized TM)	–2,65	–0,05	0,32	–4,56	0,06	0,47	–1,23	–0,23
Разработанный фреймворк	**–2,30**	0,01	**0,37**	**–3,77**	0,05	**0,49**	–4,95	**–0,09**

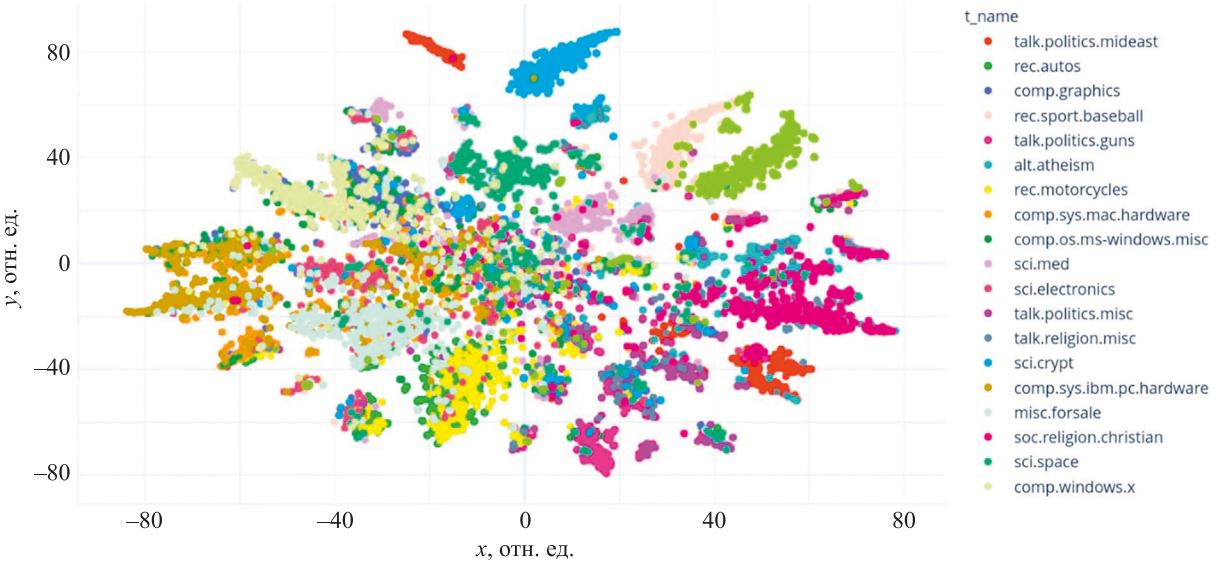

Рис. 5. Кластеризация представлений текстов датасета 20 newsgroups с помощью t-SNE

Fig. 5. Embeddings clustering for 20 newsgroups dataset with t-SNE

Заключение

Предложенный фреймворк автоматического подбора гиперпараметров для моделей с аддитивной регуляризацией позволил получить высококачественные модели за малое количество времени, что продемонстрировано результатами сравнения с аналогичными решениями. Благодаря этим особенностям он очень удобен для решения задач разведочного анализа, экономит время специалистов и позволяет проводить экспериментальные исследования в режиме распределенной обработки. Полученные эмбеддинги текстов можно использовать для решения разнообразных прикладных задач.

В дальнейшем планируется расширение фреймворка с помощью включения в него поддержки работы с различными модальностями и разработки схемы обработки данных с частичной разметкой во время обучения модели. Также возможно ускорение работы фреймворка за счет параллелизации на шаге предобработки данных.

Литература

1. Khanthaapha P., Pipanmaekaporn L., Kamonsantiroj S. Topic-based user profile model for POI recommendations // Proc. of the 2nd International Conference on Intelligent Systems, Metaheuristics Swarm Intelligence. 2018. P. 143–147. https://doi.org/10.1145/3206185.3206203
2. Peña F.J., O'Reilly-Morgan D., Tragos E.Z., Hurley N., Duriakova E., Smyth B., Lawlor A. Combining rating and review data by initializing latent factor models with topic models for Top-N recommendation // Proc. of the 14th ACM Conference on Recommender Systems. 2020. P. 438–443. https://doi.org/10.1145/3383313.3412207
3. Sokhin T., Butakov N. Semi-automatic sentiment analysis based on topic modeling // Procedia Computer Science. 2018. V. 136. P. 284–292. https://doi.org/10.1016/j.procs.2018.08.286
4. Nevezhin E., Butakov N., Khodorchenko M., Petrov M., Nasonov D. Topic-driven ensemble for online advertising generation // Proc. of the 28th International Conference on Computational Linguistics. 2020. P. 2273–2283. https://doi.org/10.18653/v1/2020.coling-main.206
5. Zamiralov A., Khodorchenko M., Nasonov D. Detection of housing and utility problems in districts through social media texts // Procedia Computer Science. 2020. V. 178. P. 213–223. https://doi.org/10.1016/j.procs.2020.11.023
6. Shi T., Kang K., Choo J., Reddy C.K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations // Proc. of the World Wide Web Conference (WWW 2018). 2018. P. 1105–1114. https://doi.org/10.1145/3178876.3186009
7. Hofmann T. Probabilistic latent semantic indexing // Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99). 1999. P. 50–57. https://doi.org/10.1145/312624.312649
8. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation // Journal of Machine Learning Research. 2003. V. 3. P. 993–1022.
9. Vorontsov K., Potapenko A., Plavin A. Additive regularization of topic models for topic selection and sparse factorization // Lecture Notes in Computer Science. 2015. V. 9047. P. 193–202. https://doi.org/10.1007/978-3-319-17091-6_14
10. Card D., Tan C., Smith N.A. Neural models for documents with metadata // Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018. P. 2031–2040. https://doi.org/10.18653/v1/p18-1189
11. Cao Z., Li S., Liu Y., Li W., Ji H. A novel neural topic model and its supervised extension // Proceedings of the AAAI Conference on Artificial Intelligence. 2015. V. 29. N 1. P. 2210–2216. https://doi.org/10.1609/aaai.v29i1.9499
12. Bianchi F., Terragni S., Hovy D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence // Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021. P. 759–766. https://doi.org/10.18653/v1/2021.acl-short.96
13. Ye J., Jing X., Li J. Sentiment Analysis Using Modified LDA // Lecture Notes in Electrical Engineering. 2018. V. 473. P. 205–212. https://doi.org/10.1007/978-981-10-7521-6_25
14. Bodrunova S., Koltsov S., Koltsova O., Nikolenko S., Shimorina A. Interval semi-supervised LDA: Classifying needles in a haystack // Lecture Notes in Computer Science. 2013. V. 8265. P. 265–274. https://doi.org/10.1007/978-3-642-45114-0_21

References

1. Khanthaapha P., Pipanmaekaporn L., Kamonsantiroj S. Topic-based user profile model for POI recommendations. *Proc. of the 2nd International Conference on Intelligent Systems, Metaheuristics Swarm Intelligence*, 2018, pp. 143–147. https://doi.org/10.1145/3206185.3206203
2. Peña F.J., O'Reilly-Morgan D., Tragos E.Z., Hurley N., Duriakova E., Smyth B., Lawlor A. Combining rating and review data by initializing latent factor models with topic models for top-n recommendation. *Proc. of the 14th ACM Conference on Recommender Systems*, 2020, pp. 438–443. https://doi.org/10.1145/3383313.3412207
3. Sokhin T., Butakov N. Semi-automatic sentiment analysis based on topic modeling. *Procedia Computer Science*, 2018, vol. 136, pp. 284–292. https://doi.org/10.1016/j.procs.2018.08.286
4. Nevezhin E., Butakov N., Khodorchenko M., Petrov M., Nasonov D. Topic-driven ensemble for online advertising generation. *Proc. of the 28th International Conference on Computational Linguistics*, 2020, pp. 2273–2283. https://doi.org/10.18653/v1/2020.coling-main.206
5. Zamiralov A., Khodorchenko M., Nasonov D. Detection of housing and utility problems in districts through social media texts. *Procedia Computer Science*, 2020, vol. 178, pp. 213–223. https://doi.org/10.1016/j.procs.2020.11.023
6. Shi T., Kang K., Choo J., Reddy C.K. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. *Proc. of the World Wide Web Conference (WWW 2018)*, 2018, pp. 1105–1114. https://doi.org/10.1145/3178876.3186009
7. Hofmann T. Probabilistic latent semantic indexing. *Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, 1999, pp. 50–57. https://doi.org/10.1145/312624.312649
8. Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, vol. 3, pp. 993–1022.
9. Vorontsov K., Potapenko A., Plavin A. Additive regularization of topic models for topic selection and sparse factorization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9047, pp. 193–202. https://doi.org/10.1007/978-3-319-17091-6_14
10. Card D., Tan C., Smith N.A. Neural models for documents with metadata. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2031–2040. https://doi.org/10.18653/v1/p18-1189
11. Cao Z., Li S., Liu Y., Li W., Ji H. A novel neural topic model and its supervised extension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015, vol. 29, no. 1, pp. 2210–2216. https://doi.org/10.1609/aaai.v29i1.9499
12. Bianchi F., Terragni S., Hovy D. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, pp. 759–766. https://doi.org/10.18653/v1/2021.acl-short.96
13. Ye J., Jing X., Li J. Sentiment Analysis Using Modified LDA. *Lecture Notes in Electrical Engineering*, 2018, vol. 473, pp. 205–212. https://doi.org/10.1007/978-981-10-7521-6_25
14. Bodrunova S., Koltsov S., Koltsova O., Nikolenko S., Shimorina A. Interval semi-supervised LDA: Classifying needles in a haystack. *Lecture Notes in Computer Science*, 2013, vol. 8265, pp. 265–274. https://doi.org/10.1007/978-3-642-45114-0_21

15. Řehůřek R., Sojka P. Software framework for topic modelling with large corpora // Proc. of the LREC 2010 Workshop on New Challenges for NLP. 2010. P. 45–50.
16. Terragni S., Fersini E., Galuzzi B.G., Tropeano P., Candelieri A. OCTIS: Comparing and optimizing topic models is simple! // Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2021. P. 263–270. https://doi.org/10.18653/v1/2021.eacl-demos.31
17. Khodorchenko M., Butakov N. Developing an approach for lifestyle identification based on explicit and implicit features from social media // Procedia Computer Science. 2018. V. 136. P. 236–245. https://doi.org/10.1016/j.procs.2018.08.262
18. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open source library for regularized multimodal topic modeling of large collections // Communications in Computer and Information Science. 2015. V. 542. P. 370–381. https://doi.org/10.1007/978-3-319-26123-2_36
19. Khodorchenko M., Teryoshkin S., Sokhin T., Butakov N. Optimization of learning strategies for ARTM-based topic models // Lecture Notes in Computer Science. 2020. V. 12344. P. 284–296. https://doi.org/10.1007/978-3-030-61705-9_24
20. Khodorchenko M., Butakov N., Sokhin T., Teryoshkin S. Surrogate-based optimization of learning strategies for additively regularized topic models // Logic Journal of the IGPL. 2022. https://doi.org/10.1093/jigpal/jzac019
21. Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures // Proc. of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15). 2015. P. 399–408. https://doi.org/10.1145/2684822.2685324
22. Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries // Proc. of the 10th Annual Joint Conference on Digital Libraries (JCDL'10). 2010. P. 215–224. https://doi.org/10.1145/1816123.1816156

15. Řehůřek R., Sojka P. Software framework for topic modelling with large corpora. *Proc. of the LREC 2010 Workshop on New Challenges for NLP*, 2010, pp. 45–50.
16. Terragni S., Fersini E., Galuzzi B.G., Tropeano P., Candelieri A. OCTIS: Comparing and optimizing topic models is simple! *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021, pp. 263–270. https://doi.org/10.18653/v1/2021.eacl-demos.31
17. Khodorchenko M., Butakov N. Developing an approach for lifestyle identification based on explicit and implicit features from social media. *Procedia Computer Science*, 2018, vol. 136, pp. 236–245. https://doi.org/10.1016/j.procs.2018.08.262
18. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open source library for regularized multimodal topic modeling of large collections. *Communications in Computer and Information Science*, 2015, vol. 542, pp. 370–381. https://doi.org/10.1007/978-3-319-26123-2_36
19. Khodorchenko M., Teryoshkin S., Sokhin T., Butakov N. Optimization of learning strategies for ARTM-based topic models. *Lecture Notes in Computer Science*, 2020, vol. 12344, pp. 284–296. https://doi.org/10.1007/978-3-030-61705-9_24
20. Khodorchenko M., Butakov N., Sokhin T., Teryoshkin S. Surrogate-based optimization of learning strategies for additively regularized topic models. *Logic Journal of the IGPL*, 2022. https://doi.org/10.1093/jigpal/jzac019
21. Röder M., Both A., Hinneburg A. Exploring the space of topic coherence measures. *Proc. of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15)*, 2015, pp. 399–408. https://doi.org/10.1145/2684822.2685324
22. Newman D., Noh Y., Talley E., Karimi S., Baldwin T. Evaluating topic models for digital libraries. *Proc. of the 10th Annual Joint Conference on Digital Libraries (JCDL'10)*, 2010, pp. 215–224. https://doi.org/10.1145/1816123.1816156

Авторы

Ходорченко Мария Андреевна — младший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57207458742, https://orcid.org/0000-0001-5446-5311, mariyaxod@yandex.ru

Бутаков Николай Алексеевич — кандидат технических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56218218400, https://orcid.org/0000-0002-2705-1313, alipoov.nb@gmail.com

Насонов Денис Александрович — кандидат технических наук, старший научный сотрудник, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 56086498600, https://orcid.org/0000-0002-2740-0173, denis.nasonov@gmail.com

Фирулик Михаил Юрьевич — директор департамента, ООО «Оператор Газпром ИД», Санкт-Петербург, 191028, Российская Федерация, SC 56086498600, https://orcid.org/0000-0001-7114-1052, firulik@gmail.com

Authors

Maria A. Khodorchenko — Junior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57207458742, https://orcid.org/0000-0001-5446-5311, mariyaxod@yandex.ru

Nikolay A. Butakov — PhD, Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56218218400, https://orcid.org/0000-0002-2705-1313, alipoov.nb@gmail.com

Denis A. Nasonov — PhD, Senior Researcher, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 56086498600, https://orcid.org/0000-0002-2740-0173, denis.nasonov@gmail.com

Mikhail Yu. Firulik — Director of Department, LLC "Operator Gazprom ID", Saint Petersburg, 191028, Russian Federation, SC 56086498600, https://orcid.org/0000-0001-7114-1052, firulik@gmail.com

Статья поступила в редакцию 20.10.2022
Одобрена после рецензирования 08.12.2022
Принята к печати 24.01.2023

Received 20.10.2022
Approved after reviewing 08.12.2022
Accepted 24.01.2023