Terra Linguistica. 2023, Том 14, № 1. С. 70–87.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 70–87.

Научная статья
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.14107

ДИНАМИЧЕСКОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ РУССКОЯЗЫЧНОГО КОРПУСА ЮРИДИЧЕСКИХ ДОКУМЕНТОВ

О.А. Митрофанова ✉ , **М.М. Атугодаге**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ o.mitrofanova@spbu.ru

Аннотация. Статья посвящена анализу результатов динамического тематического моделирования законодательных актов Российской Федерации, указов высших должностных лиц и постановлений Верховного и Конституционного Судов за 2008–2022 годы, входящих в исследовательский корпус русскоязычных юридических документов. В статье описаны процедуры формирования и предобработки корпуса, эксперименты по обучению тематических моделей на данном корпусе. Рассматривается как стандартная тематическая модель, так и динамическая тематическая модель, учитывающая изменение тем корпуса во времени. После обучения моделей в различных условиях были определен набор оптимальных параметров обучения. В качестве основного инструмента тематического моделирования использовалась библиотека BERTopic на языке программирования Python, комбинирующая алгоритмы построения тематических моделей и нейросетевые контекстуализированные модели распределенных векторных вложений. Исследовательские данные могут представлять интерес не только для специалистов в области компьютерной лингвистам, но и для социологов, политологов, юристов, работающих с законодательными документами.

Ключевые слова: тематическое моделирование, динамическая тематическая модель, BERTopic, корпус русскоязычных юридических документов, Российская Газета.

Финансирование: НИП СПбГУ № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта», грант РНФ № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики».

Для цитирования: Митрофанова О.А., Атугодаге М.М. Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. Т. 14. № 1. С. 70–87. DOI: 10.18721/JHSS.14107

Research article
DOI: https://doi.org/10.18721/JHSS.14107

DYNAMIC TOPIC MODELLING OF THE RUSSIAN LEGAL TEXT CORPUS

O.A. Mitrofanova ✉ , **M.M. Athugodage**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ o.mitrofanova@spbu.ru

Abstract. The article is devoted to the dynamic topic modelling analysis of legislative acts, decrees of senior officials and resolutions of the Supreme and Constitutional Courts dated 2008–2022, included into the research corpus of Russian legal documents. The article describes the procedures of corpus construction and preprocessing, training of topic models on this corpus. We consider both standard topic model and a dynamic topic model that takes into account changes in topics over time. After training the models in various conditions, a set of optimal training parameters was determined. The BERTopic library was used as the main tool for topic modelling, combining algorithms for constructing topic models and contextualized neural network models of distributed vectors. The research data may be of interest both for specialists in the field of computational linguistics as well as for sociologists, political scientists, lawyers working with legislative documents.

Keywords: topic modelling, dynamic topic model, BERTopic, Russian corpus of legal documents, Russian gazette.

Acknowledgements: Research Program of St. Petersburg State University No. 75254082 "Modeling the communicative behavior of residents of a Russian metropolis in the socio-speech and pragmatic aspects with the use of artificial intelligence methods", RSF grant No. 21-78-10148 "Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics."

Citation: O.A. Mitrofanova, M.M. Athugodage, Dynamic topic modelling of the russian legal text corpus, Terra Linguistica, 14 (1) (2023) 70–87. DOI: 10.18721/JHSS.14107

Введение

Тематическое моделирование – это способ построения семантической модели коллекции текстовых документов, который определяет, к какой теме относится каждый из документов. Результатом тематического моделирования является установление соответствия между текстами корпуса и скрытыми факторами, темами (кластерами слов-тематизаторов), при этом каждый документ соотносится с некоторой вероятностью с одной или нескольким темам, а сами темы могут пересекаться: определенное слово может быть с разными вероятностями отнесено к нескольким темам [1, 2 и т.д.]. Тематические модели способствуют повышению результативности процедур извлечения информации из естественно-языковых текстов, таких как, например, автоматическая рубрикация, кластеризация и классификация документов, сентимент-анализ и т.д., а также вносят весомый вклад в обучение систем искусственного интеллекта [3–6]. Сфера применения тематических моделей широка, она охватывает корпусы текстов разных типов и жанров: новостные корпусы [7–9], корпусы социальных сетей [10–14], корпусы медицинских текстов [15], корпусы по финансам и банковскому делу [16, 17], корпусы текстов по разным областям научного знания [18], художественные корпусы [19–25] и т.д. Тематическое моделирование юридических документов – это новая исследовательская область, где удается получать ценные результаты [26]. Наше исследование призвано решить задачу изучения динамики тем в законодательных текстах, поэтому в фокусе нашего внимания находится такая модификация алгоритмов тематического моделирования как динамическое тематическое моделирование.

Автоматическая обработка текстов юридических документов представляет большой интерес как для лингвистов, так и для юристов, правоведов, социологов. В фокусе внимания исследователей находится юридическая терминология [27–29], необходимость ее гармонизации [30–32], машинный перевод [33], неоднозначность [34], сложность юридических текстов [35–38] и другие вопросы. В связи с нуждами прикладных исследований формируются специализированные корпусы юридических документов: многоязычные корпусы – Europarl (a Parallel Corpus for Statistical Machine Translation)[1], Параллельный корпус документов ООН (United Nations Parallel Corpus)[2] и т.д., для русского языка – корпус законов CorCodex, корпус решений конституционного суда CorDec, корпус локальных актов CorRIDA[3] и т.д. Однако для цели нашего исследования требуется особенный корпусной ресурс.

В ходе динамического тематического моделирования восстанавливается структура тем корпуса, документы в котором имеют хронологическую метаразметку и распределены по сегментам, соответствующим периодам времени (чаще всего, по месяцам, годам или десятилетиям). Результатом является выделение нишевых тем, характеризующих изучаемые хронологические промежутки [39, 40]. В случае применения динамического тематического моделирования оказывается возможным проведение анализа того, как формируется повестка дня в законодательных актах (как федеральных, так и региональных), указах Президента Российской Федерации и постановлениях Верховного Суда и Конституционного Суда Российской Федерации и других юридических документах России.

Изменения, происходящие в российской юриспруденции и влияющие на уровень правовой культуры в России, могут свидетельствовать о «юридическом богатстве общества», явно или косвенно представлять состояние его правовой жизни и характеризовать коллективный юридический опыт [41] в тот или иной хронологический период. Законодательные тексты своевременно отражают нововведения в общественно-политической и социальной сфере, тем самым, являются маркерами значимых событий в государстве и обществе. Представляет исследовательский интерес взаимосвязь между содержанием общественных дискуссий и их результатом в виде законодательных актов: эта взаимосвязь определима в ходе лингвистической экспертизы текстов, включающей также и процедуры тематического моделирования. Юридические документы являются источником данных о политической ситуации в стране, в том числе, об изменениях политического строя, уровня политических и гражданских свобод и т.п., о кадровых изменениях в составе руководящих органов власти, фиксируемых указом Президента или Председателя Правительства. Тем самым, исследование динамики тем в корпусе русскоязычных юридических документов может показать, как развивалась с течением времени различные аспекты деятельности государства.

Исследовательский корпус русскоязычных юридических документов

Для проведения процедур динамического тематического моделирования русскоязычных юридических текстов был использован исследовательский корпус сопоставимых документов[4], исходное предназначение которого связано с решением задачи упрощения юридических текстов. Корпус содержит юридический документ, опубликованный на сайте Российской Газеты, и его упрощенный вариант, или комментарий, написанный специалистами из Российской Газеты. Российская Газета[5] – это официальный печатный орган Правительства Российской Федерации; законы вступают в силу только после публикации в Парламентской газете, Российской газете, Собрании законодательства Российской Федерации или на Официальном интернет-портале правовой информации [42]. Интернет-портал «Российской газеты» RG.RU, который использовался при сборе документов, существует с 1999 г. и также наделён официальным статусом.

[1] https://www.statmt.org/europarl/
[2] https://conferences.unite.un.org/UNCorpus
[3] https://www.plaindocument.org/corpora
[4] https://www.kaggle.com/datasets/athugodage/russian-legal-text-parallel-corpus
[5] https://rg.ru/doc

Параллельный корпус содержит не все документы, опубликованные в Российской Газете, а только лишь те, к которым прилагался комментарий (или упрощенный текст). С одной стороны, это ограничивает рамки нашего исследования, так как нам придется работать не со всеми выпущенными законами, а лишь некоторой частью, причем довольно малой. Для сравнения, в среднем Российская Газета публикует около 1,5–3 тыс. документов в год, число комментариев имеет предельное значение – 465). С другой стороны, редакция Российской газеты производит отбор текстов для комментирования и публикует комментарий только к общественно важным документам. В результате, сам источник задает критерии отбора текстов по степени важности для общества. У нас есть возможность не тратить усилия на обработку полного массива юридических текстов, а сосредоточить исследовательское внимание на наиболее важных.

В экспериментах, описываемых в настоящей статье, использовался только сегмент корпуса, содержащий полные тексты, поэтому характеристики сегмента корпуса, содержащего комментарии, в данной статье не приводится.

Корпус содержит около 3 тыс. статей, датированных от 31 декабря 2008 г. по 28 ноября 2022 г. (это дата публикации, и, соответственно, вступления в силу). Дата публикация отмечена в отдельном столбце в следующем формате: «31 декабря 2008». В нашем корпусе больше всего статей за 2018 г. (465 статей), меньше всего – за 2008 г. (1 статья). Распределение документов исследовательского корпуса по годам представлено на рис. 1. Мы приняли решение сохранить хронологические рамки корпуса и оставить в экспериментальном материале единственную статью за 2008 г., поскольку сегментация на периоды времени при построении динамической тематической модели производится не по годам, а по дням, это позволяет соблюсти требование сбалансированности данных.

Как можно увидеть из графика на рис. 1 выше, за последнюю пятилетку было опубликовано больше документов (1664 документов), чем за предыдущее десятилетие (1299 документов). Это можно связать с тем, что интернет-портал развивается и публикует все больше комментариев к законам. Следует заметить, что до 2008 г. Российская Газета вообще не публиковала комментарии.

В среднем размер документа составляет около 400–600 токенов. Самый крупный документ содержит чуть более 70 тыс. токенов. Есть документы с количеством токенов около 100 –

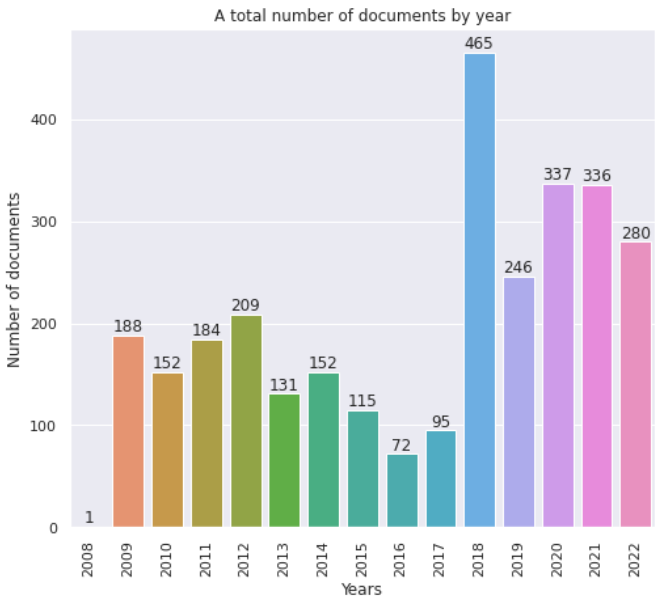

Рис. 1. Распределение документов исследовательского корпуса по годам

Fig. 1. Distribution of documents in the research corpus by year

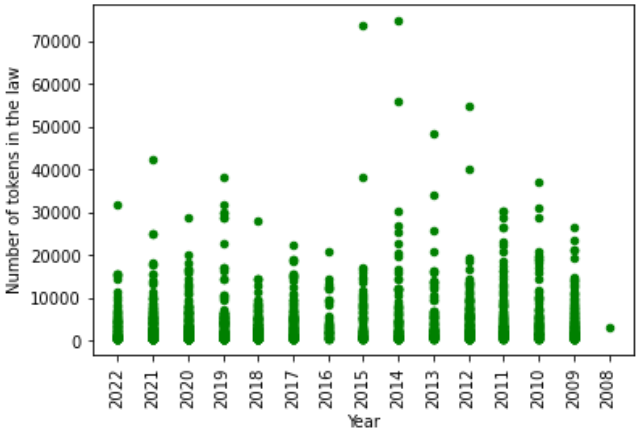

Рис. 2. Распределение документов исследовательского корпуса по годам с учетом объема текстов в токенах
Fig. 2. Distribution of documents in the research corpus by year taking into account text size in tokens

в основном, это поправки в закон и краткие указы Президента РФ или Председателя Правительства. График на рис. 2 показывает распределение документов по годам и их объем в токенах.

Обучение тематических моделей требует предварительной обработки документов корпуса. Для решения обсуждаемых в статье исследовательских задач в текстах корпуса необходимо провести следующие процедуры:

1) упрощенная токенизация с использованием функции .split() в Python;

2) удаление стоп-слов с помощью словаря, составленного на основе списков служебных слов, оборотов, аббревиатур, латинских и числовых обозначений [19, 20, 22];

3) лемматизация с использованием библиотеки для морфологического анализа pymorphy2;[6]

4) приведение дат в формат, соответствующий требованиям к обучению модели: «31 декабря 2008» → «2008-12-31»;

5) сортировка статей по дате (первоначальная сортировка производилась по годам).

Как показали эксперименты, качество предобработки корпуса позволяет повысить качество обучения тематических моделей и содержательно улучшить результат.

Особенности процедур стандартного и динамического тематического моделирования с помощью библиотеки BERTopic

Известные способы тематического моделирования включают в себя группу алгебраических моделей (LSA (латентно-семантический анализ), NMF (неотрицательная матричная факторизация) и т.д.) и вероятностных моделей (pLSA (вероятностный латентно-семантический анализ), LDA (латентное размещение Дирихле), LPA (латентное размещение Патинко), HTMM (скрытая тематическая марковская модель) и т.д.). На практике используются их мультимодальные расширения за счет введения дополнительных параметров корпуса: авторство в ATM (автор-тематической модели), фактор адресата в ARTM (модели автор-получатель), связи между темами в HTM (иерархической тематической модели), наличие заранее заданных ключевых слов-тематизаторов в GuidedLDA (управляемом латентном размещении Дирихле), учет конструкций в составе тем в n-граммных тематических моделях, возможность обобщения состава тем с помощью меток и т.д. [1, 3, 5 и т.д.]. В последние годы появился новый класс тематических моделей, комбинирующих вероятностные процессы и модели распределенных векторов, например, LDA2Vec, Top2Vec, ETM (Embedded topic model), CTM (Contextualized topic model), BERTopic и т.д. [43–47]. Стандартные и комбинированные тематические модели могут использоваться в модификации DTM (динамическая тематическая модель) [39–40].

[6] https://pymorphy2.readthedocs.io/

Преимущество комбинированных тематических моделей заключается в том, что они позволяют улучшить качество семантического представления текста и сократить потери, связанные с использованием технологии представления корпуса в виде мешка слов (bag-of-words). Использование контекстуализированных моделей в комбинированных тематических моделях имеет ряд особенностей: по сравнению с моделями типа word2vec модели BERT (Bidirectional Encoder Representations from Transformers) сохраняют векторные представления слов с учетом контекста и поэтому чувствительны, например, к полисемии.

В комбинированной модели BERTopic, используемой в нашем исследовании, к распределенным векторным вложениям BERT применяются процедуры кластеризации со снижением размерности и ранжированием слов-тематизаторов для формирования тем [47]. Тематическое моделирование в BERTopic проходит в три этапа. На первом этапе каждый документ корпуса преобразуется в векторное представление с использованием предварительно обученной языковой модели BERT. На втором этапе проводится снижение размерности векторов методом UMAP и кластеризация результирующих векторных вложений методом HDBSCAN. На третьем этапе из кластеров документов извлекаются специфичные для них ключевые выражения (n-граммы) с использованием измененного варианта метрики c-TF-IDF. Эти слова и словосочетания ранжируются методом MMR и рассматриваются в качестве кандидатов в тематизаторы. Формула для расчета значений c-TF-IDF представлена ниже:

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right),$$

где $tf_{t,c}$ – частота термина t в классе c. Класс c понимается как набор документов, объединенных в единое целое. Чтобы оценить информативность термина для класса, обратная частота документа заменяется обратной частотой класса, которая рассчитывается как логарифм отношения среднего количества слов в классе A и частоты употребления термина t во всех классах. Чтобы вывести только положительные значения, к логарифмируемому выражению добавляется единица.

Данный подход, позволяющий находить важные слова для кластеров текстов, а не только для отдельных документов, лежит в основе стандартной тематической модели BERTopic, которая описывает темы, характерные для корпуса в целом. Например, документы, связанные с вакцинацией, будут объединены в глобальную тему «вакцинация». Эта тема может со временем развиваться или исчезать: возможная тема для 2011 г. – «вакцинация от гриппа», а для 2021 г. – «вакцинация от ковида». При переходе от стандартного варианта BERTopic к динамическому тематическому моделированию на третьем этапе построения модели к варианту c-TF-IDF добавляются метки времени i, формирующие дополнительную модальность.

$$W_{t,c,i} = tf_{t,c,i} \cdot \log\left(1 + \frac{A}{tf_t}\right).$$

В нашей работе мы использовали архитектуру и модели из библиотеки BERTopic[7] для тематического моделирования корпуса юридических текстов в стандартном и динамическом вариантах, которые рассматриваются в последующих разделах.

Построение стандартной тематической модели исследовательского корпуса юридических документов

В начале работы с библиотекой BERTopic нужно настроить модель векторизации и инициализировать ее класс. При настройке мы использовали стоп-словарь [19, 20, 22], а также определяли минимальную частоту вхождения слов в корпус (1) и максимальную долю документов корпуса,

[7] https://github.com/MaartenGr/BERTopic

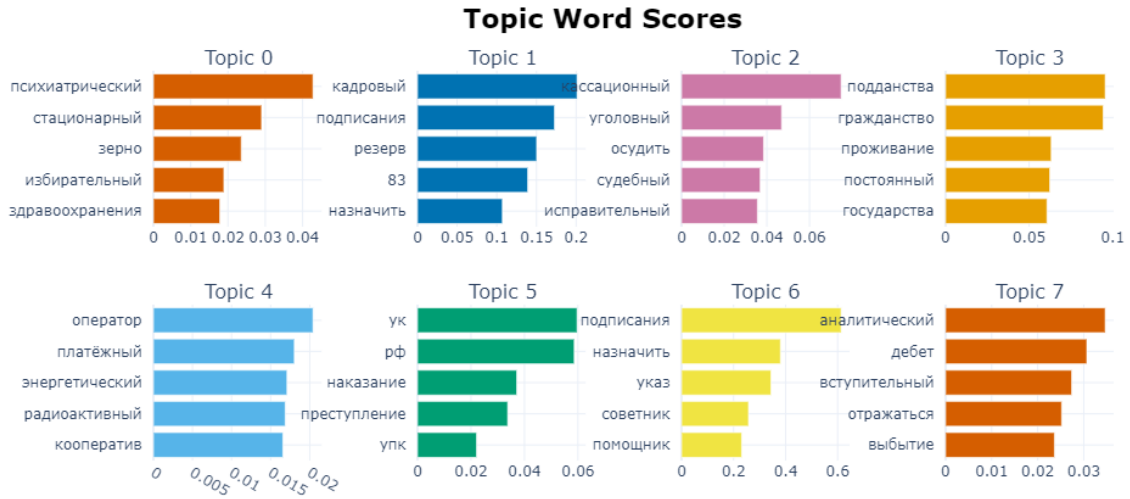

Рис. 3. Первые восемь тем из стандартной модели BERTopic

Fig. 3. The first eight topics from standard BERTopic model

включающих слова (0.6). Если увеличить значение второго параметра (0.8 или 0.95), то в числе слов-тематизаторов будут широко распространенные термины и сокращения, не дающие должного понимания содержательной стороны темы: *гк*, *рф*, *статья* и т.п.

Далее следует задать параметры обучения самой модели BERTopic и инициализировать ее класс. В результате серии экспериментов были подобраны оптимальные параметры обучения:

- минимальный размер темы – 10 слов-тематизаторов,
- первые *n* слов-тематизаторов – 10,
- размер *n*-грамм, выделяемых в корпусе: от 3 до 5.

В результате обучения стандартной тематической модели BERTopic были получены результаты, проиллюстрированные на рис. 3–7. На рис. 1 отображены 8 наиболее значимых тем (для каждой из них при визуализации выводится пять слов-тематизаторов) из 133 тем, предсказанных моделью.

Функционал библиотеки BERTopic позволяет нарисовать «карту расстояний» между темами (по значению), см. рис. 4–7.

На рис. 4–7 темы обозначены серыми кругами. Семантически близкие темы расположены близко друг к другу. Так, например, в правом нижнем углу рисунка расположена группа тем, связанных с налогообложением (рис. 4), внизу слева – группа тем, связанных с уголовными преступлениями (рис. 5), вверху по центру – группа тем, связанных с туристическим обслуживанием (рис. 6), вверху справа – группа тем, связанных с пенсионным законодательством (рис. 7). Выделяются более специфические темы, связанные с правами военнослужащих, пандемией COVID-19, топонимами – названиями регионов, числовыми обозначениями и т.д.

Динамическое тематическое моделирование корпуса русскоязычных юридических документов

Для обучения динамической тематической модели необходимо предварительно задать точки во времени, относительно которых будут оцениваться изменения в темах. В наших экспериментах в качестве данных точек использовались дни публикации текстов законодательных актов. Таким образом, на вход алгоритма подавался список дат вида: [‘2008-12-31’, ‘2009-01-15’, ‘2009-01-16’, ... ‘2022-11-28’]. Такие метки как ‘2009-01-01’ отсутствуют, поскольку в этот момент законы не публиковались. Были проведены эксперименты по динамическому тематическому моделированию с изменением числа тем.

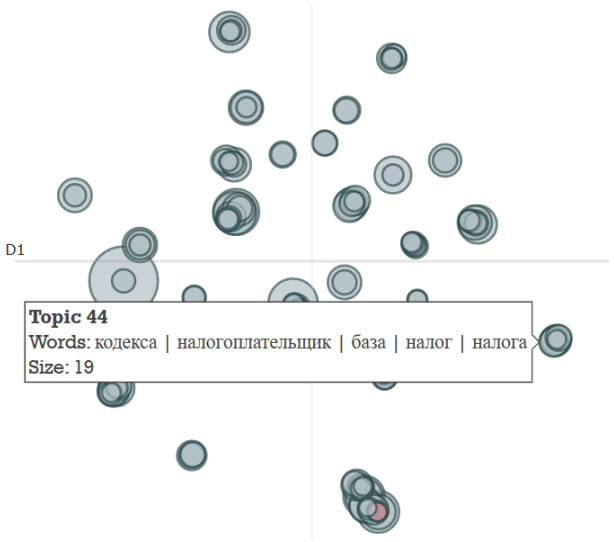

Рис. 4. «Карта расстояний» стандартной модели BERTopic (тема 44)
Fig. 4. The «distance map» for standard BERTopic model (topic 44)

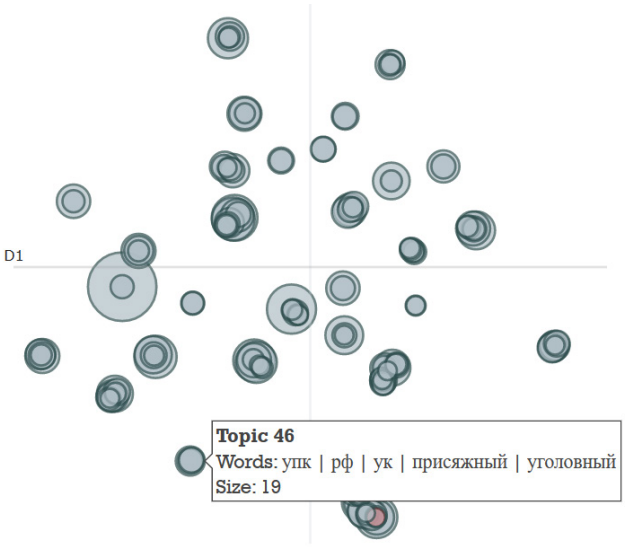

Рис. 5. «Карта расстояний» стандартной модели BERTopic (тема 46)
Fig. 5. The «distance map» for standard BERTopic model (topic 46)

На рис. 8 представлена визуализация изменений 13 наиболее значимых тем. При интерактивной работе с графиком можно отслеживать изменение состава темы: в определенный момент времени слова-тематизаторы могут быть стандартными (глобальными), либо специфическими (локальными). Как видно из графика, в начале 2018 г. резко увеличилось число документов, связанных с темами 1 и 9, которые относятся к кадровым изменениям. В начале 2018 г. прошли Президентские выборы и, в связи с этим, множество назначений на руководящие посты. В юридических документах с 2020 г. фигурирует тема борьбы с новой коронавирусной инфекцией COVID-19. На рис. 8 видно, что глобальный состав темы 2 определяется словами *коронавирусный*, *инфекция*, *covid* и т.д., а локальные темы отличаются друг от друга обозначениями регионов (*Забайкальский край*, *Ростовская область*, *Ямало-Ненецкий автономный округ* и т.д.)

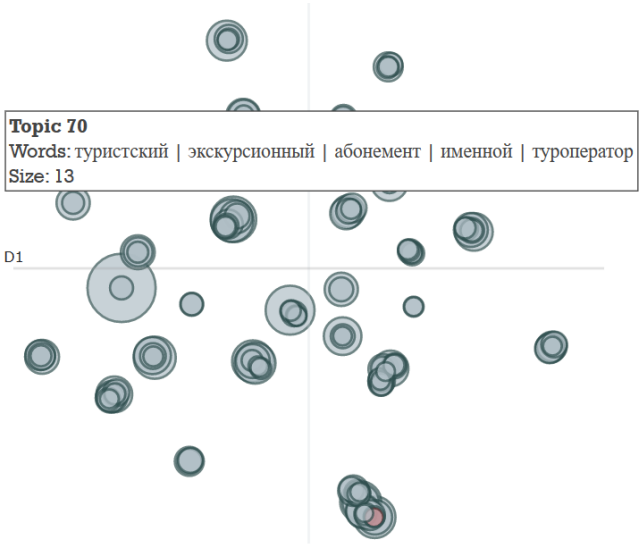

Рис. 6. «Карта расстояний» стандартной модели BERTopic (тема 70)

Fig. 6. The «distance map» for standard BERTopic model (topic 70)

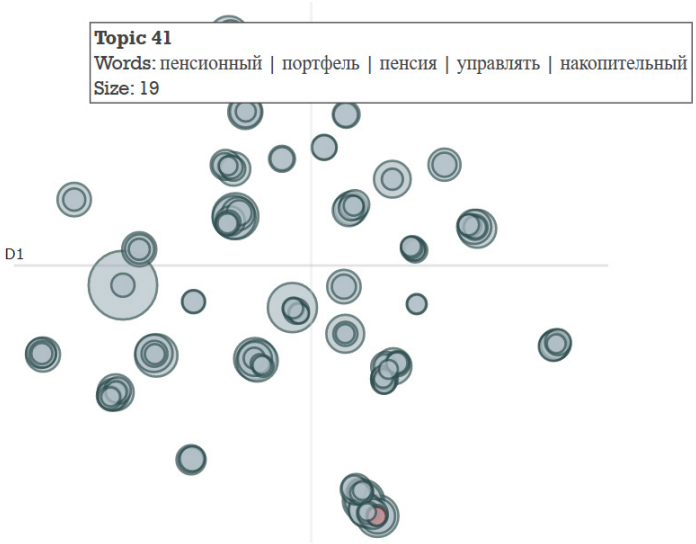

Рис. 7. «Карта расстояний» стандартной модели BERTopic (тема 41)

Fig. 7. The «distance map» for standard BERTopic model (topic 41)

На рис. 9 визуализирована динамика первых 20 значимых тем на отрезке от февраля 2021 г. по ноябрь 2022 г.

Желтым цветом отмечена динамика темы со словами-тематизаторами *подданства*, *гражданство*, *проживание*, *постоянный*, *государства* и т.д., график имеет пик, соответствующий июню 2021 г., и спад в осенне-зимний период. Суть в том, что в весенне-летний период появилось много документов по данной теме, и их число снизилось к концу 2021 г. Для рассматриваемой темы на пике ее популярности в июне 2021 г. характерны глобальные слова-тематизаторы, а вот 20 декабря 2021 г. ее состав изменился, в ней появились слова, обозначающие холодное оружие: *клинковый*, *холодное*, *кортики*, *кортиков*, *оружие* и т.д.

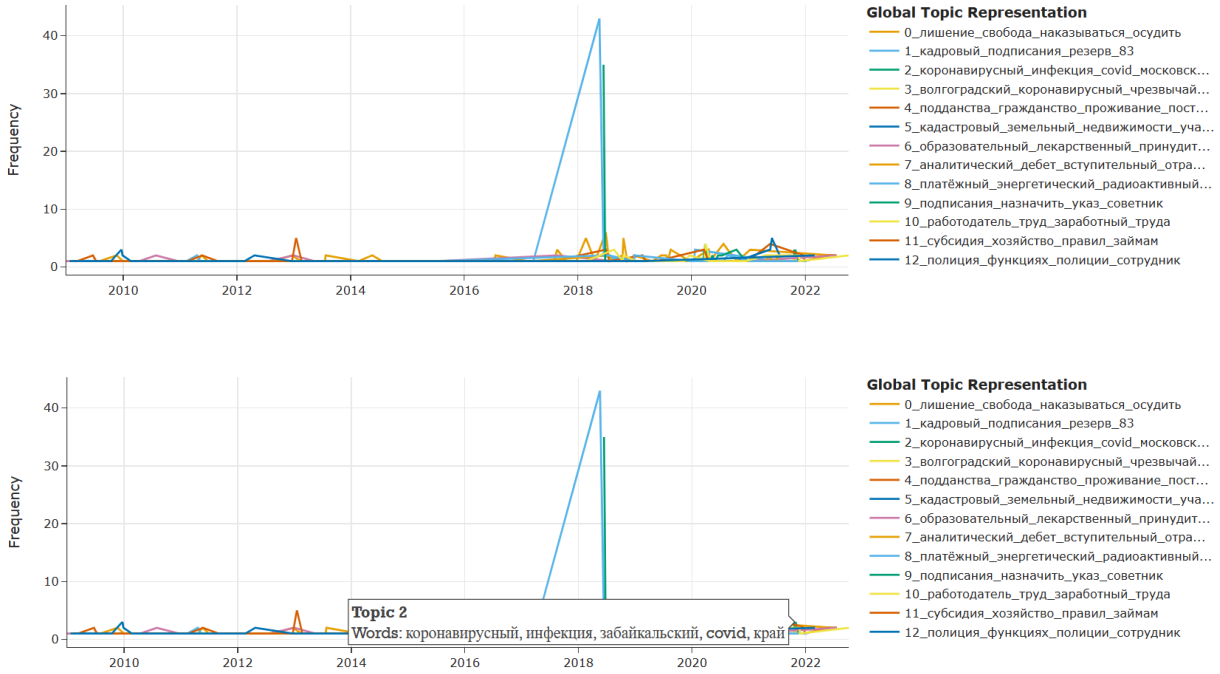

Рис. 8. Значимые темы динамической модели BERTopic

Fig. 8. Significant topics of dynamic BERTopic model

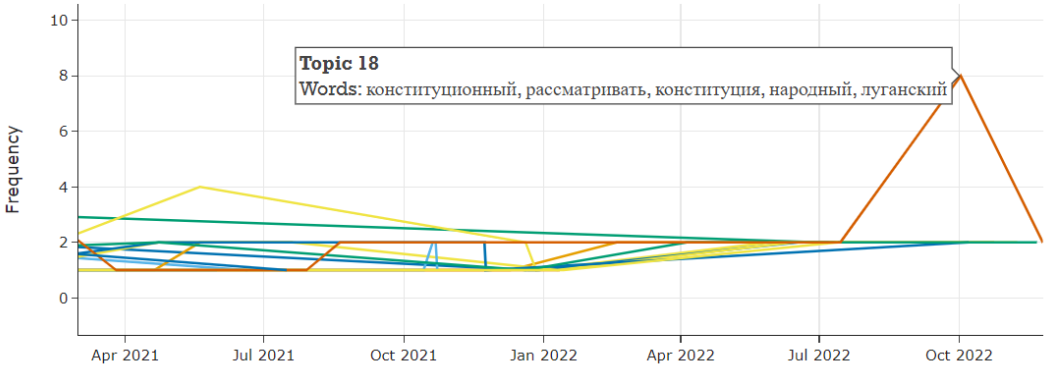

Рис. 9. Значимые темы динамической модели BERTopic (тема 18)

Fig. 9. Significant topics of dynamic BERTopic model (topic 18)

К концу выбранного нами периода, в октябре 2022 г., популярной темой в документах стала тема, связанная с конституционными изменениями в РФ и принятием в состав РФ новых регионов (Донецкая Народная Республика (ДНР), Луганская Народная Республика (ЛНР), Запорожская и Херсонская области). Данная тема на графике обозначена номером 18 и выделена красным цветом (рис. 9). В тот же период становится актуальной тема, связанная с указом об объявлении частичной мобилизации в РФ: 5 октября 2022 г. локальными ключевыми словами этой темы были: *задач*, *работодатель*, *мобилизация*, *контракт*, *добровольный*.

Рассмотрим картину динамики тем в исследовательском корпусе юридических документов за период от 31 декабря 2008 г. до конца 2011 г., см. график на рис. 10.

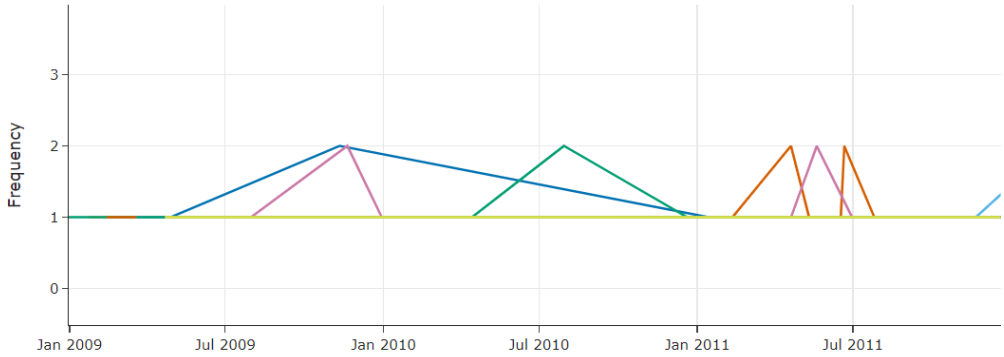

Рис. 10. Значимые темы динамической модели BERTopic (2008–2011 г.)
Fig. 10. Significant topics of dynamic BERTopic model (2008–2011)

Зеленым отмечен график темы со словами-тематизаторами, соотносимыми с правами на интеллектуальную собственность: *патентный*, *поверить*, *поверенного*, *интеллектуальный*, *квалификационный* и т.д. Синим помечен график глобальной темы правосудия ук, рф, наказание, преступление и т.д. Розовым показан график темы взяточничества: в ноябре 2009 г. она имела следующие локальные слова-тематизаторы: *лишение*, *крупный*, *такового*, *наказываться*, 1854 и т.д., а в мае 2011 г. – *взятка*, *лишение*, *штраф*, *взятки*, *наказываться* и т.д.

Интерфейс библиотеки BERTopic позволяет пользователю выбрать отдельную глобальную тему и проследить ее развитие за весь период времени, охваченный корпусом. На рис. 11 представлены изменения темы правосудия со словами-тематизаторами *ук*, *рф*, *наказание*, *преступление* и т.д. Как можно заметить, пики популярности уголовной тематики приходятся на конец 2013 г. и начало 2018 г.

В библиотеке BERTopic есть возможность проводить сравнительный анализ нескольких тем одновременно, см. рис. 12. На графике синим и красным цветами обозначены изменения в упоминании двух смежных тем, связанных с Конституцией РФ. До 2016 г. конституционная тематика не встречается в корпусе вообще. Пик популярности приходится на конец 2022 г., что может быть связано с присоединением к РФ новых территорий. Характерно, что летом 2021 г., когда вносились поправки в Конституцию РФ, популярность тем была минимальной.

Более подробно ознакомиться с результатами экспериментов можно источниках Jupyter Notebook[8] и GitHub[9].

Заключение

В данной статье представлены результаты экспериментального исследования динамики основных тем в корпусе русскоязычных юридических документов за 2008–2022 гг. с использованием методов тематического моделирования. Были построены стандартная и динамическая тематические модели корпуса с помощью библиотеки BERTopic, интегрирующей использование нейросетевых моделей распределенных векторных вложений типа трансформер (BERT), алгоритмы кластеризации и снижения размерности. В результате экспериментов были оценены возможности инструмента BERTopic в исследовании юридических текстов, осуществлена интерпретация полученных данных. Созданные тематические модели позволили выявить наиболее значимые события в жизни государства и общества, отраженные в юридических документах, помогли найти соответствующие им слова-тематизаторы, и, тем самыми, представить эти события в развитии с течением времени.

[8] https://drive.google.com/file/d/1MHGglLDtv4-HMSAgLHRFkHb0Y_Pey2Cn/view
[9] legal_dtm/BERTopic_DTM_legal_docs.ipynb at main · Athugodage/legal_dtm (github.com)

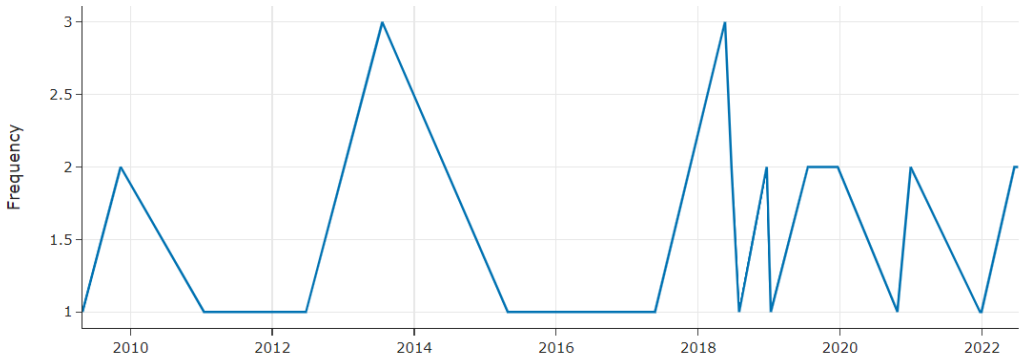

Рис. 11. Динамика темы *ук*, *рф*, *наказание*, *преступление* и т.д.

Fig. 11. Dynamic changes in the topic *Criminal code*, *Russian Federation*, *punishment*, *crime* etc.

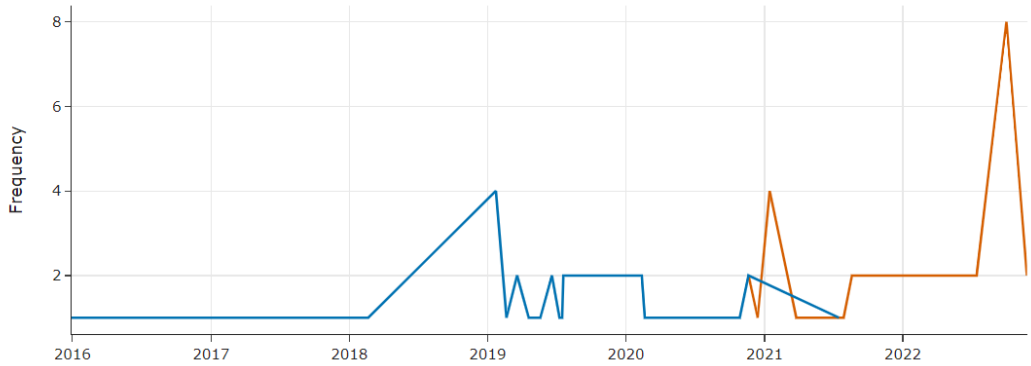

Рис. 12. Динамика двух смежных тем, связанных с Конституцией РФ

Fig. 12. Dynamic changes of two topics related to the Constitution of the Russian Federation

Данное исследование может быть полезно как компьютерным лингвистам, так и юристам, политологам, социологам и специалистам, исследующие историю развития российского законодательства в первой четверти XXI в. Проведенный анализ материала расширяет наши знания о созданном корпусе русскоязычных юридических документов, который предназначен для обучения алгоритмов упрощения текстов.

Результаты нашей работы позволят облегчить поиск релевантных документов и повысят доступность юридической информации для неспециалистов.

СПИСОК ИСТОЧНИКОВ

1. **Daud A., Li J., Zhou L., Muhammad F.** Knowledge discovery through directed probabilistic topic models: a survey // Proceedings of Frontiers of Computer Science in China, 2010. P. 280–301. URL: https://www.researchgate.net/publication/215904200_Latent_Dirichlet_allocation_LDA_and_topic_modeling_models_applications_future_challenges_a_survey

2. **Blei D.M., Ng A.Y., Jordan M.I.** Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

3. **Милкова М.А.** Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. № 1(5). 2019. С. 57–70. URL: http://digital-economy.ru/images/easyblog_articles/520/DE-2019-01-06.pdf?ysclid=lefaetif1z601986150

4. **Николенко С., Кадурин Е., Архангельская Е.** Глубокое обучение. Погружение в мир нейронных сетей. СПб., 2018. URL: https://b-ok.cc/book/4987601/95075c

5. **Кирина М.А.** Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. Вып. 20(2). 2022. С. 93–109. URL: https://lingngu.elpub.ru/jour/article/view/384

6. **Воронцов К.В.** Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM 2023. URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

7. **Карпович С.Н.** Русскоязычный корпус текстов СКТМ-ру для построения тематических моделей // Международная конференция «Корпусная лингвистика-2015». СПб., 2015.

8. **Shavrina T., Shapovalova O.** To the Methodology of Corpus Construction for Machine Learning: «TAIGA» Syntax Tree Corpus and Parser // Proceedings of the International Conference «Corpus Linguistics – 2017». Saint-Petersburg, 2019.

9. **Mitrofanova O., Kriukova A., Shulginov V., Shulginov V.** E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science, vol. 1357. Springer, 2021. P. 102–114. URL: https://doi.org/10.1007/978-3-030-71214-3_9

10. **Кольцов С.Н., Кольцова О.Ю., Митрофанова О.А., Шиморина А.С.** Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS–2014, СанктПетербург, 19–20 ноября 2014 г. СПб., 2014. С. 135–142.

11. **Bodrunova S., Blekanov I.S., Kukarkin M.** Topics in the Russian Twitter and Relations between their Interpretability and Sentiment // 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019. P. 549–554.

12. **Mamaev I.D, Mitrofanova, O.A.** Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus // A. Filchenkov, J. Kauttonen, L. Pivovarova (Eds.). Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings. Communications in Computer and Information Science. Vol. 1292. Springer, 2020. P. 17–33. URL: https://doi.org/10.1007/978-3-030-59082-6_2

13. **Mitrofanova O., Sampetova V., Mamaev I., Moskvina A., Sukharev K.** Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling // CEUR Workshop Proceedings, 2813. 2021. P. 101–116.

14. **Nikolenko S.I., Koltcov S., Koltsova O.** Topic modelling for qualitative studies // Journal of Information Science. Vol. 43. 2017.

15. **Khawaji K., Almubark I., Almalki A., Taylor B.** Similarity Matching for Workflows in Medical Domain Using Topic Modeling // 2018 IEEE World Congress on Services (SERVICES). San Francisco, CA, USA, 2018.

16. **Шишкина В.С.** Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным. М., 2019.

17. **Dowling M., Piepenbrink A., Saqib A., Helmi H.** Machine learning in finance: A topic modeling approach. 2019. URL: https://arxiv.org/ftp/arxiv/papers/1911/1911.12637.pdf

18. **Vorontsov K.V., Voronov S.O.** Automatic Filtering of Russian Scientific Content using Machine Learning and Topic Modeling // International Conference on Computational Linguistics and Intellectual Technologies «Dialogue–2015». Moscow, 2015. URL: http://www.dialog-21.ru/media/2135/vorontsov.pdf

19. **Митрофанова О.А.** Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика–2015». СПб., 2015.

20. **Митрофанова О.А.** Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Труды международной конференции «Корпусная лингвистика–2019». СПб: Издательство Санкт-Петербургского университета, 2019.

21. **Скоринкин Д.А.** Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа «Война и мир» Л.Н. Толстого). Дис. ... канд. филол наук. М., 2019.

22. **Mitrofanova O.A., Sedova A.G.** Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // Information Technology and Computational Linguistics (ITCL 2017). Association for Computing Machinery, 2017.

23. **Rhody L.M.** Topic Modeling and Figurative Language // Journal of Digital Humanities. Vol. 2(1). Winter 2012. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

24. **Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina T.** Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction // L. Martínez-Villaseñor, H. Ponce, O. Herrera-Alcántara, F.A. Castro-Espinoza (Eds.). Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Proceedings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 12469. Springer, 2020. P. 134–151. URL: https://doi.org/10.1007/978-3-030-60887-3_13

25. **Zamiraylova E., Mitrofanova O.** Dynamic topic modeling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization // R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL-2019: Proceedings of the III International Conference RWTH Aahen University. CEUR Workshop Proceedings. Vol. 2552. 2019. P. 321–339.

26. **Badenes-Olmedo C., Redondo-Garcia J.-L., Corcho O.** Legal document retrieval across languages: topic hierarchies based on synsets // arXiv. 2019. URL: https://arxiv.org/abs/1911.12637v1

27. **Дмитриева А.В.** «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации. Сравнительное конституционное обозрение. 2017. 118 (3). С. 125–133.

28. **Mattila H.E.S.** Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas. Routledge, 2013.

29. **Tiersma P.M.** Legal Language. Chicago, London: University of Chicago Press, 1999.

30. **Туранин В.Ю.** Юридическая терминология в современном российском законодательстве (теоретико-правовое исследование). Дис ... д-ра юрид. наук. Белгород: БГУ, 2017.

31. **Виландеберк А.А.** Принципы и методы гармонизации терминологии на основе корпуса специальных параллельных текстов: на материале документов ООН. Автореф. дис. ... канд филол. наук. СПб, РГПУ им. А.И.Герцена, 2005. URL: https://new-disser.ru/_avtoreferats/01002771726.pdf/f?ysclid=leciciqi30122006496

32. **Виландеберк А.А.** Корпус параллельных правовых документов как составная часть АРМ юриста-переводчика // Труды Международной научной конференции «Корпусная лингвистика 2004». СПб., 2004.

33. **Mirzagitova A.** Realisation of statistical machine translation based on a parallel Tatar-Russian corpus of legal texts Proceedings of the International Conference «Turkic Languages Processing: TurkLang – 2015». Kazan, 2015. P. 39–49.

34. **Блинова О.В., Белов С.А.** Языковая неоднозначность и неопределённость в русских правовых текстах. Вестник Санкт-Петербургского университета. Право. 2020. № 11(4). С. 774–812.

35. **Блинова О.В.** Оценка сложности русских правовых текстов: архитектура модели // Мир русского слова. 2022. № 2. С. 4–13.

36. **Блинова О.В., Тарасов Н.А.** Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21. М.: Изд-во РГГУ, 2022. С. 1017–1028.

37. **Кучаков Р.К., Савельев Д.А.** Сложность правовых актов в России: Лексическое и синтаксическое качество текстов / Под ред. Д. Скугаревского. СПб.: ИПП ЕУСПб, 2018.

38. **Кнутов А.В., Плаксин С.М., Григорьева Н.Л., Синятуллин Р.Х., Чаплинский А.В., Успенская А.М.** Сложность российских законов. Опыт синтаксического анализа. М.: Изд. дом Высшей школы экономики, 2020. 311 с.

39. **Greene D., Cross J.** Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach // Political Analysis. 2016. Vol. 25. P. 77–94. // arXiv. URL: https://arxiv.org/pdf/1607.03055.pdf

40. **Blei D.M., Lafferty J.D.** Dynamic topic models // Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, 2006. URL: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf

41. **Усманова Е.Ф.** Правовая культура российского общества в современных условиях // Мир науки и образования. 2016. № 3(7). URL: https://cyberleninka.ru/article/n/pravovaya-kultura-rossiysa kogo-obschestva-v-sovremennyh-usloviyah?ysclid=lefhwff9kf730175397

42. Федеральный закон от 14.06.1994 N 5-ФЗ (ред. от 01.05.2019) «О порядке опубликования и вступления в силу федеральных конституционных законов, федеральных законов, актов палат Федерального Собрания». URL: http://www.kremlin.ru/acts/bank/6332

43. **Dieng A.B., Ruiz F.J.R., Blei D.M.** Topic Modeling in Embedding Spaces // arXiv. 2019. URL: https://arxiv.org/abs/1907.04907

44. **Moody C.E.** Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2Vec // arXiv. 2016. URL: https://arxiv.org/abs/1605.02019

45. **Angelov D.** Top2Vec: Distributed Representations of Topics // arXiv. 2020. URL: https://arxiv.org/abs/2008.09470

46. **Bianchi F., Terragni S., Hovy D.** Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence // Proceedings 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. ACL, 2021. P. 759–766. URL: https://aclanthology.org/2021.acl-short.96/

47. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv. 2022. URL: https://arxiv.org/pdf/2203.05794.pdf

REFERENCES

[1] **A. Daud, J. Li, L. Zhou, F. Muhammad,** Knowledge discovery through directed probabilistic topic models: a survey, Proceedings of Frontiers of Computer Science in China, 2010. Pp. 280–301. Available at: https://www.researchgate.net/publication/215904200_Latent_Dirichlet_allocation_LDA_and_topic_modeling_models_applications_future_challenges_a_survey

[2] **D.M. Blei, A.Y. Ng, M.I. Jordan,** Latent Dirichlet Allocation, Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. Available at: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

[3] **M.A. Milkova,** Topic models as a tool for "distant reading", Digital Economy. 1 (5) 2019 57–70. Available at: http://digital-economy.ru/images/easyblog_articles/520/DE-2019-01-06.pdf?ysclid=lefaetif1z601986150

[4] **S. Nikolenko, E. Kadurin, E. Arkhangelskaya,** Deep learning. Dive into the world of neural networks. SPb., 2018. Available at: https://b-ok.cc/book/4987601/95075c

[5] **M.A. Kirina,** Comparison of thematic models based on LDA, STM and NMF for a qualitative analysis of Russian short fiction. Vestnik NGU. Series: Linguistics and intercultural communication. 20 (2) 2022 93–109. Available at: https://lingngu.elpub.ru/jour/article/view/384

[6] **K.V. Vorontsov,** Probabilistic topic modeling: ARTM regularization theory and BigARTM open source library. 2023. Available at: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

[7.] **S.N. Karpovich,** Russian-language corpus of SKTM-ru texts for building thematic models, International Conference "Corpus Linguistics – 2015". SPb., 2015.

[8] **T. Shavrina, O. Shapovalova,** To the Methodology of Corpus Construction for Machine Learning: "TAIGA" Syntax Tree Corpus and Parser, Proceedings of the International Conference "Corpus Linguistics – 2017". Saint-Petersburg, 2019.

[9] **O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov,** E-hypertext Media Topic Model with Automatic Label Assignment, Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science, vol. 1357. Springer, 2021. Pp. 102–114. Available at: https://doi.org/10.1007/978-3-030-71214-3_9

[10] **S.N. Koltsov, O.Yu. Koltsova, O.A. Mitrofanova, A.S. Shimorina,** Interpretation of semantic links in the texts of the Russian-language segment of LiveJournal based on the LDA thematic model, Technologies of the information society in science, education and culture: a collection of scientific articles. Proceedings of the XVII All-Russian Joint Conference "Internet and Modern Society" IMS-2014, St. Petersburg, November 19-20, 2014. St. Petersburg, 2014. Pp. 135–142.

[11] **S. Bodrunova, I.S. Blekanov, M. Kukarkin,** Topics in the Russian Twitter and Relations between their Interpretability and Sentiment, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019. Pp. 549–554.

[12] **I.D. Mamaev, O.A. Mitrofanova,** Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus, A. Filchenkov, J. Kauttonen, L. Pivovarova (Eds.). Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings. Communications in Computer and Information Science. Vol. 1292. Springer, 2020. Pp. 17–33. Available at: https://doi.org/10.1007/978-3-030-59082-6_2

[13] **O. Mitrofanova, V. Sampetova, I. Mamaev, A. Moskvina, K. Sukharev,** Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling, CEUR Workshop Proceedings, 2813. 2021. Pp. 101–116.

[14] **S.I. Nikolenko, S. Koltcov, O. Koltsova,** Topic modelling for qualitative studies, Journal of Information Science, 43 (2017).

[15] **K. Khawaji, I. Almubark, A. Almalki, B. Taylor,** Similarity Matching for Workflows in Medical Domain Using Topic Modeling, 2018 IEEE World Congress on Services (SERVICES). San Francisco, CA, USA, 2018.

[16] **V.S. Shishkina,** Tematicheskoye modelirovaniye finansovykh potokov korporativnykh kliyentov banka po tranzaktsionnym dannym [Thematic modeling of financial flows of corporate clients of the bank based on transactional data]. M., 2019.

[17] **M. Dowling, A. Piepenbrink, A. Saqib, H. Helmi,** Machine learning in finance: A topic modeling approach, 2019. Available at: https://arxiv.org/ftp/arxiv/papers/1911/1911.12637.pdf

[18] **K.V. Vorontsov, S.O. Voronov,** Automatic Filtering of Russian Scientific Content using Machine Learning and Topic Modeling, International Conference on Computational Linguistics and Intellectual Technologies "Dialogue–2015". Moscow, 2015. Available at: http://www.dialog-21.ru/media/2135/vorontsov.pdf

[19] **O.A. Mitrofanova,** Probabilistic modeling of the topics of Russian text corpora using the GenSim computer tool, Proceedings of the international conference "Corpus Linguistics – 2015". SPb., 2015.

[20] **O.A. Mitrofanova,** The study of the structural organization of a work of art using thematic modeling: experience with the text of the novel "The Master and Margarita" M.A. Bulgakova, Proceedings of the International Conference "Corpus Linguistics – 2019". St. Petersburg: St. Petersburg University Press, 2019.

[21] **D.A. Skorinkin,** Semantic markup of literary texts for quantitative research in philology (on the example of the novel "War and Peace" by L.N. Tolstoy). PhD Thesis. M., 2019.

[22] **O.A. Mitrofanova, A.G. Sedova,** Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose), Information Technology and Computational Linguistics (ITCL 2017). Association for Computing Machinery, 2017.

[23] **L.M. Rhody,** Topic Modeling and Figurative Language, Journal of Digital Humanities. Vol. 2(1). Winter 2012. Available at: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

[24] **T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, T. Kirina,** Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction, L. Martínez-Villaseñor, H. Ponce, O. Herrera-Alcántara, F.A. Castro-Espinoza (Eds.). Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Proceedings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 12469. Springer, 2020. Pp. 134–151. Available at: https://doi.org/10.1007/978-3-030-60887-3_13

[25] **E. Zamiraylova, O. Mitrofanova,** Dynamic topic modeling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization, R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL–2019: Proceedings of the III International Conference RWTH Aahen University. CEUR Workshop Proceedings. Vol. 2552. 2019. Pp. 321–339.

[26] **C. Badenes-Olmedo, J.-L. Redondo-Garcia, O. Corcho,** Legal document retrieval across languages: topic hierarchies based on synsets, arXiv. 2019. Available at: https://arxiv.org/abs/1911.12637v1

[27] **A.V. Dmitrieva,** "The art of legal writing": a quantitative analysis of the decisions of the Constitutional Court of the Russian Federation. Comparative constitutional review. 118 (3) (2017) 125–133.

[28] **H.E.S. Mattila,** Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas. Routledge, 2013.

[29] **P.M. Tiersma,** Legal Language. Chicago, London: University of Chicago Press, 1999.

[30] **V.Yu. Turanin,** Yuridicheskaya terminologiya v sovremennom rossiyskom zakonodatelstve (teoretiko-pravovoye issledovaniye). Dis ... d-ra yurid. nauk. Belgorod: BGU, 2017.

[31] **A.A. Vilandeberk,** Principles and methods of terminology harmonization based on the corpus of special parallel texts based on UN documents. PhD Thesis Abstract. St. Petersburg, Russian State Pedagogical University, 2005. Available at: https://new-disser.ru/_avtoreferats/01002771726.pdf?ysclid=leciciqi30122006496

[32] **A.A. Vilandeberk,** Corpus of Parallel Legal Documents as a Part of AWP of a Lawyer-Translator, Proceedings of the International Scientific Conference "Corpus Linguistics 2004". SPb., 2004.

[33] **A. Mirzagitova,** Realisation of statistical machine translation based on a parallel Tatar-Russian corpus of legal texts Proceedings of the International Conference "Turkic Languages Processing: TurkLang – 2015". Kazan, 2015. Pp. 39–49.

[34] **O.V. Blinova, S.A. Belov,** Linguistic ambiguity and uncertainty in Russian legal texts. Bulletin of St. Petersburg University. Right. 11 (4) (2020) 774–812.

[35] **O.V. Blinova,** Assessing the complexity of Russian legal texts: the architecture of the model, The World of the Russian Word, 2 (2022) 4–13.

[36] **O.V. Blinova, N.A. Tarasov,** Metrics of complexity of Russian legal texts: selection, use, primary evaluation of effectiveness, Computational Linguistics and Intelligent Technologies: Based on the materials of the annual international conference "Dialogue". Issue. 21. Moscow: RGGU Press, 2022. Pp. 1017–1028.

[37] **R.K. Kuchakov, D.A. Saveliev,** Complexity of Legal Acts in Russia: Lexical and Syntactic Quality of Texts / Ed. D. Skugarevsky. St. Petersburg: IPP EUSPb, 2018.

[38] **A.V. Knutov, S.M. Plaksin, N.L. Grigor'yeva, R.H. Sinyatullin, A.V. Chaplinskiy, A.M. Uspenskaya,** Complexity of Russian Laws. Parsing experience. M.: Ed. house of the Higher School of Economics, 2020.

[39] **D. Greene, J. Cross,** Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach, Political Analysis. 25 (2016) 77–94. arXiv. Available at: https://arxiv.org/pdf/1607.03055.pdf

[40] **D.M. Blei, J.D. Lafferty,** Dynamic topic models, Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, 2006. Available at: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf

[41] **E.F. Usmanova,** Legal culture of the Russian society in modern conditions, World of Science and Education. 3 (7) (2016). Available at: https://cyberleninka.ru/article/n/pravovaya-kultura-rossiyskogo-obschestva-v-sovremennyh-usloviyah?ysclid=lefhwff9kf730175397

[42] Federal Law № 5-FZ of June 14, 1994 (as amended on May 1, 2019) "On the Procedure for Publication and Entry into Force of Federal Constitutional Laws, Federal Laws, and Acts of the Chambers of the Federal Assembly." Available at: http://www.kremlin.ru/acts/bank/6332

[43] **A.B. Dieng, F.J.R. Ruiz, D.M. Blei,** Topic Modeling in Embedding Spaces, arXiv. 2019. Available at: https://arxiv.org/abs/1907.04907

[44] **C.E. Moody,** Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2Vec, arXiv. 2016. Available at: https://arxiv.org/abs/1605.02019

[45] **D. Angelov,** Top2Vec: Distributed Representations of Topics, arXiv. 2020. Available at: https://arxiv.org/abs/2008.09470

[46] **F. Bianchi, S. Terragni, D. Hovy,** Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 2: Short Papers. ACL, 2021. Pp. 759–766. Available at: https://aclanthology.org/2021.acl-short.96/

[47] **M. Grootendorst,** BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv. 2022. Available at: https://arxiv.org/pdf/2203.05794.pdf

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Митрофанова Ольга Александровна
Olga A. Mitrofanova
E-mail: o.mitrofanova@spbu.ru
ORCID: https://orcid.org/0000-0002-3008-5514

Атугодаге Марк Махешевич
Mark M. Athugodage
E-mail: m.athugodage@yahoo.com

Поступила: 22.01.2023; Одобрена: 16.03.2023; Принята: 17.03.2023.
Submitted: 22.01.2023; Approved: 16.03.2023; Accepted: 17.03.2023.