Научная статья

УДК 81'32
DOI 10.25205/1818-7935-2022-20-2-93-109

Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы

Маргарита Александровна Кирина

Национальный исследовательский университет «Высшая школа экономики»
Санкт-Петербург, Россия
mkirina@hse.ru, https://orcid.org/0000-0002-7381-676X

Аннотация

Описываются результаты тематического моделирования малой художественной прозы на основе трех методов – латентного размещения Дирихле (LDA), структурного тематического моделирования (STM) и неотрицательной матричной факторизации (NMF) – в сочетании с разными вариантами предобработки текстов (все части речи vs только существительные). Апробация экспериментального дизайна осуществляется на материале Корпуса русского рассказа 1900–1930 гг. Исследование позволило выявить особенности рассматриваемых алгоритмов и оценить эффективность их применения для качественного анализа художественной прозы.

Ключевые слова

компьютерная лингвистика, автоматическая обработка текста, тематическое моделирование, художественная литература, малая проза, русская литература, русский рассказ, цифровая гуманитаристика

Благодарности

Публикация подготовлена в результате проведения исследования по проекту № 21-04-053 «Методы искусственного интеллекта для филологических исследований» в рамках Программы «Научный фонд Национального исследовательского университета “Высшая школа экономики” (НИУ ВШЭ)» в 2021 г.

Для цитирования

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. 2022. Т. 20, № 2. С. 93–109. DOI 10.25205/1818-7935-2022-202-93-109

A Comparison of Topic Models Based on LDA, STM and NMF for Qualitative Studies of Russian Short Prose

Margarita A. Kirina

National Research University “Higher School of Economics”
St. Petersburg, Russian Federation
mkirina@hse.ru, https://orcid.org/0000-0002-7381-676X

Abstract

The paper describes the results of topic modelling of short prose fiction based on three methods, namely Latent Dirichlet Allocation (LDA), the Structural Topic Model (STM), and the Non-Negative Matrix Factorization (NMF), combined with different text preprocessing options (all parts of speech vs. only nouns). The experimental design is tested on the basis of the Corpus of Russian Short Stories of 1900–1930s. The research made it possible to determine

the specifics of the algorithms under consideration and to assess the effectiveness of their application for the qualitative analysis of fiction texts.

Keywords

computational linguistics, automatic text processing, topic modelling, literary texts, short prose, Russian literature, Russian short story, digital humanities

Acknowledgements

The publication was prepared within the framework of the Academic Fund Program at the National Research University "Higher School of Economics" (HSE) in 2021 (grant no. 21-04-053 "Artificial Intelligence Methods in Literature and Language Studies")

For citation

Kirina, M. A. A Comparison of Topic Models Based on LDA, STM and NMF for Qualitative Studies of Russian Short Prose. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2022, vol. 20, no. 2, pp. 93–109. (in Russ.) DOI 10.25205/1818-7935-2022-20-2-93-109

Введение

Тематическое моделирование – метод машинного обучения, использующийся для категоризации больших неструктурированных текстовых данных. Этот подход к анализу коллекций документов широко применяется для междисциплинарных исследований в таких областях, как компьютерная лингвистика [Mitrofanova, 2015], социология [McFarland et al., 2013], биоинформатика [Liu et al., 2016] и др. Тематические модели способствуют улучшению результатов при решении ряда задач естественной обработки языка – например, автоматического реферирования [Huang et al., 2018], классификации [Moubayed et al., 2016], сентимент-анализа [Rana et al., 2016], а также вносят вклад в обучение систем искусственного интеллекта (в том числе чат-ботов [Guo et al., 2018]).

Цель тематического моделирования – выявление скрытых семантических структур – тем, или *топиков* (*topic*) [1], характеризующих содержание исследуемой текстовой коллекции. Отличительную особенность тематического моделирования составляет то, что в его ходе осуществляется бикластеризация (*biclustering*) – одновременная кластеризация не только слов (термов), но и текстов (документов). В связи с тем что при тематическом моделировании выполняется нечеткая кластеризация, «любое слово или документ с некоторой вероятностью относится к нескольким темам» [Митрофанова, 2014, с. 221]. Количество выделяемых топиков определяется одним из двух способов: самостоятельно исследователем или в соответствии с оптимальным числом, полученным в результате автоматического сравнения формальных характеристик моделей (например, на основе меры когерентности). Важно отметить, что при обработке текстов реализуется подход, который называется «мешком слов» (*bag-of-words*), это означает, что не учитываются ни порядок слов, ни их грамматические и синтаксические характеристики. Слова и документы, оказавшиеся наиболее характерными для топиков, дают представление о тематическом разнообразии корпуса текстов.

Материалом исследований с применением методов тематического моделирования, как правило, становятся специальные тексты, относящиеся к академическому дискурсу, СМИ или социальным сетям [Nikolenko et al., 2017; Jacobs, Tschötschel, 2019]. Тематическое моделирование литературных корпусов, напротив, проводится значительно реже [Jockers, Mimno, 2013; Schöch, 2017; Митрофанова, 2019]. Так, с одной стороны, выделяются проблемы, связанные с оценкой результатов тематического моделирования и их валидностью для качественной интерпретации художественных произведений [Rhody, 2012; Da, 2019]. С другой – подчеркивается, что описать содержание литературного корпуса этими методами можно, однако работа с художественными текстами требует большего внимания на этапе предобработ-

[1] В данной работе при анализе тематических моделей предлагается использовать как термин *топик*, так и термин *тема*. Первый, более технический, отсылает к результатам тематического моделирования; второй – непосредственно к темам, которые можно выявить исходя из топика.

ки, включающей удаление не только стоп-слов, но и значительного количества частотных и редких слов [Uglanova, Gius, 2020]. Стоит отметить, что проводятся эксперименты с применением разных алгоритмов тематического моделирования с целью оценки их потенциала для улучшения интерпретируемости результатов [Navarro-Colorado, 2018; Zamiraylova, Mitrofanova, 2020; Sherstinova et al., 2020].

К настоящему времени разработано значительное количество алгоритмов тематического моделирования (LDA, LSA, pLSA, NMF, STM, CTM и др.). Наиболее известным является метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA), позволяющий выявлять общие темы, встречающиеся в корпусе [Blei et al., 2003]. Для выявления «нишевых» топиков, что ценно при анализе художественных текстов, может использоваться неотрицательная матричная факторизация (NMF) [Lee, Seung, 1999]. Утверждается, что этот вид тематического моделирования позволяет получить более разнообразные и связные топики [O'Callaghan et al., 2015]. Интерес также представляют и структурные тематические модели, в основном применяемые для социологических исследований [Roberts et al., 2013]. Данный метод дает возможность сравнивать словарные составы тем, а также учитывать при анализе метаинформацию, или ковариаты (*covariates*), такие как автор, год написания или жанр текста.

Художественные произведения являются тем типом текстов, в которых находят отражение разнообразные жизненные ситуации, мысли и чувства человека – причем они необязательно имеют реальную природу. Появление фантастического плана действия в значительной мере расширяет поле предметов и явлений, которые могут изображаться в художественной литературе. Поэтому художественные тексты отличаются от других особенно широким тематическим спектром. Так, выделяют как «вечные», так и культурно-исторические, или злободневные, темы [Томашевский, 1996]. Неизбежно и то, что литературные тексты содержат «глубинные», или «подтекстовые», темы, которые могут приводить к неоднозначной интерпретации.

По своему объему и охвату – географическому, культурному и лингвистическому – литературные произведения занимают важное место в информационном пространстве не только современности, но и исторического прошлого. В этом смысле тематическое моделирование художественных текстов может способствовать исследованию процессов, происходивших в литературной системе на протяжении длительного времени. Разработка стратегий тематического моделирования текстов, написанных образным языком, может послужить шагом к улучшению работы интеллектуальных систем, связанных с обработкой и пониманием естественного языка.

В исследовании предпринимается попытка выявить особенности трех методов тематического моделирования – метода латентного размещения Дирихле (LDA), структурного тематического моделирования (STM) и неотрицательной матричной факторизации (NMF), а также оценить эффективность их применения для качественного анализа малой художественной прозы на русском языке. В статье описывается сравнение результатов тематического моделирования литературных текстов в сочетании с разными подходами к формированию выборки – со всеми знаменательными частями речи и после удаления из текстов всех частей речи, кроме существительных. Кроме того, оценивается соответствие тем, выделяемых в ходе работы алгоритма, содержанию отнесенных к ним рассказов. Последнее особенно важно для задачи смысловой компрессии художественного текста, на данный момент наиболее достоверно осуществляемой только с привлечением экспертов [Sherstinova et al., 2021].

Апробация экспериментального дизайна осуществляется на материале Корпуса русского рассказа первой трети XX в. [2] [Мартыненко и др., 2018а; 2018б; Martynenko, Sherstinova, 2020]. Полученные пилотные результаты открывают начальный этап исследования наиболее

[2] Корпус русского рассказа 1900–1930 гг. URL: https://russian-short-stories.ru/.

оптимального сочетания способов предобработки текстов и алгоритмов тематического моделирования для качественного анализа малой художественной прозы.

1. Дизайн эксперимента

Корпус русского рассказа первой трети XX в. разрабатывается с целью сохранения национального литературного наследия, построения модели литературно-художественной системы в рамках одного жанра и проведения различных пилотных экспериментов [Martynenko, Sherstinova, 2020]. Аннотированный подкорпус, на базе которого проводится данное исследование, включает в себя 310 рассказов, написанных 300 авторами, и содержит метаинформацию в соответствие со следующими историческими периодами:

- период I (1900–1913): предреволюционные годы, Русско-японская война;
- период II (1914–1922): Первая мировая война, Февральская и Октябрьская революции, Гражданская война;
- период III (1923–1930): послереволюционные годы с окончания Гражданской войны до 1930-х гг. [Ibid.].

На этапе предварительной обработки тексты подкорпуса были токенизированы, лемматизированы и размечены по частям речи с помощью пакета R ‘udpipe’ (модель Russian-SynTag Rus) [Wijffels, 2020; Straka, Straková, 2019]. Полученные данные проверялись вручную, и по возможности исправлялись ошибки, допущенные при лемматизации. Объем выборки после удаления пунктуации, цифр и прочих нетекстовых символов составил 1 061 785 токенов.

После этого были сформированы две выборки – одна для эксперимента на всех частях речи (выборка-1), вторая – для моделей, построенных только на существительных (выборка-2). Исходя из частотного списка лемм были сформированы два списка стоп-слов для каждой из выборок. Так, были удалены слова, общие для 99 % всех текстов, и редкие слова, встречавшиеся менее 5 раз (для выборки-1) и менее 3 раз (для выборки-2). В соответствии с частеречной разметкой UDPipe из выборок были исключены также имена собственные и названия, служебные части речи (предлоги, междометия, частицы, союзы). Размер выборки-1 составил 268 415 лемм, размер выборки-2 – 106 696 лемм.

В ходе эксперимента к каждой из двух выборок попеременно применялись методы тематического моделирования – латентное размещение Дирихле (LDA), структурное тематическое моделирование (STM) и неотрицательная матричная факторизация (NMF). При построении моделей были использованы имплементации данных алгоритмов для R из соответствующих пакетов: ‘topicmodels’ [Grün, Hornik, 2011], ‘stm’ [Roberts et al., 2019], ‘NMF’ [Gaujoux, Seoighe, 2010]. В качестве общего числа топиков, которое необходимо извлечь, для сравнения тематических моделей на более детальном уровне было выбрано небольшое *k*, равное 10. Для каждой тематической модели рассматривались первые 10 слов и первые 10 документов топика – как наиболее информативные. Кроме того, оценивалась тематическая сочетаемость слов, составивших топики, как предложено в [Jockers, Mimno, 2010], и их соответствие содержанию относящихся к ним рассказов.

Всего было получено 6 моделей [3]: LDA-1_10 (выборка-1, k = 10), LDA-2_10 (выборка-2, k = 10), STM-1_10 (выборка-1, k = 10), STM-2_10 (выборка-2, k = 10), NMF-1_10 (выборка-1, k = 10), NMF-2_10 (выборка-2, k = 10) [4]. В результате было обнаружено, что наилучшей интерпретируемостью для каждой пары моделей обладают следующие: LDA_1-10 (выборка-1,

[3] Дополнительно были построены две модели на основе алгоритма LDA при *k*, соответствующем рекомендации по оптимальному количеству тем: LDA-1_30 и LDA-2_12. Однако это не способствовало улучшению результата с точки зрения интерпретируемости и, напротив, привело к потере связности терм топика (что согласуется с наблюдениями, сделанными ранее в [Uglanova, Guis, 2020; Gryaznova, Kirina, 2021]). По этой причине данные модели были исключены из дальнейшего сравнения.

[4] Здесь и далее вид тематической модели указывается сокращенно по схеме: МЕТОД-НОМЕР ВЫБОРКИ_КОЛИЧЕСТВО ТОПИКОВ.

k = 10); STM_1-10 (выборка-1, *k* = 10); NMF_2-10 (выборка-2, *k* = 10). Поэтому они будут рассмотрены далее.

Для удобства описания каждой из выделенных тем давалось условное название, или метка, которое назначалось вручную. При именовании тем мы старались отразить основную идею, которая могла бы объединять слова, образующие топик. Если тем в одном топике выделялось несколько, то название присваивалось согласно ведущей теме и в соответствии с содержанием наиболее характерных для него документов. Полученный результат, разумеется, не является единственно возможным. Более того, некоторые названия кажутся более удачными, чем другие, что в значительной степени определяется степенью семантической однородности ключевых слов топика [5].

2. Результаты применения LDA- модели

В табл. 1 представлены результаты тематического моделирования на основе алгоритма LDA (LDA-1_10) на материале выборки-1, включающей все знаменательные части речи.

Таблица 1

Результаты тематического моделирования LDA-1_10

Table 1

Results of topic modeling with LDA-1_10

№	Название темы	Содержание
t_1	ВСТРЕЧИ С ДАМОЙ	квартира, кабинет, номер, дама, лестница, счастливый, художник, звонок, зеркало, нравиться
t_2	ЗАСТОЛЬЕ	выпить, водка, фабрика, праздник, лавка, кухня, жалко, власть, курица, плохой
t_3	ЗАКЛЮЧЕНИЕ В ТЮРЬМЕ И КАЗНЬ	камера, тюрьма, лагерь, надзиратель, человеческий, мгновение, двигаться, яркий, непонятный, казнь
t_4	ВЛАСТЬ ИМУЩИЕ	губернатор, вещь, степь, обезьяна, дацан, власть, дагестанец, лето, усадьба, голый
t_5	УБИЙСТВО ЗВЕРЯ	винтовка, телега, шапка, икона, зверь, овраг, монастырь, веревка, тянуть, рубаха
t_6	ПЛАВАНИЕ НА КОРАБЛЕ	волна, чистый, цветы, капитан, кучка, степь, стекло, матрос, вещь, бегать
t_7	ПОЖАР РЕВОЛЮЦИИ	игра, революция, рыжий, доска, дым, дружба, машина, стекло, сосна, кофейня
t_8	НА ФРОНТАХ ВОЙНЫ	немец, война, полковник, окоп, учитель, фронт, матрос, отряд, начальник, команда
t_9	ОХОТА НА ВОЛКА	дядя, волк, спинка, тетка, охота, постель, кровать, подниматься, бабушка, мужчина
t_10	ОБЩЕЕ СОБРАНИЕ	библиотека, лодка, утопленник, газета, детский, мастерская, собрание, завод, речь, портрет

Тема t_1 «ВСТРЕЧИ С ДАМОЙ» характеризует пространство, в котором разворачивается действие (*квартира*, *кабинет*, *номер*). Это также и места романтических встреч героев – в рассказах описывается любовь к *даме* и эпизоды из семейной жизни (в том числе супруже-

[5] Надо сказать, что проблема именования тем – отдельная задача тематического моделирования [Lau et al., 2010; Ерофеева, Митрофанова, 2019], которая в данной статье не рассматривается, так как именование топиков для художественного текста – задача малоизученная.

ская измена) (А. Вербицкая «Поздно», П. Невежин «Обломки семьи», А. Лазарев-Грузинский «Незабудки»). Слово *художник*, затрудняющее интерпретацию, встречается в текстах как в прямом значении (рассказ Б. Бентовина «Завещание», герой которого когда-то проводил время «в безалаберной среде художников»), так и в переносном – в рассказе А. Вербицкой «Поздно»:

(1) *– Какое сравнение! – горячо говорил он. – Здесь* ***художник-жизнь*** *только набросал эскиз...* [6]

Тема t_2 «ЗАСТОЛЬЕ» связана с описанием деревенской жизни: слова *выпить* и *водка* характеризуют быт людей, населяющих эту местность (например, рассказы «Как Иван “провел время”» С. Подъячева, «Спектакль в селе Огрызове» В. Шишкова, «Как гуляет Тихоныч» Г. Гребенщикова).

Тема t_3 «ЗАКЛЮЧЕНИЕ В ТЮРЬМЕ И КАЗНЬ» объединяет тематически связанные «Рассказ о семи повешенных» Л. Андреева и рассказы «Тюрьма» М. Горького и «Баррикада» Г. Яблочкова. Стоит также отметить, что слово *лагерь*, несмотря значительную связность других терм топика, здесь не указывает на лагерь, например, военнопленных, а используется для обозначения больших групп людей, в том числе массовых сборищ:

(2) *«Произвольно и несправедливо всё это... Разве можно делить людей только на два* ***лагеря****?.. А например – я? Ведь, в сущности, я – не господин и не раб!»* (М. Горький «Тюрьма»)

(3) *...идут толпами, – что там уже толпа, – целый* ***лагерь****, с ночлегами и чуть ли даже не с палатками* (Ф. Сологуб «В толпе»)

Тема t_6 «ПУТЕШЕСТВИЕ НА КОРАБЛЕ» описывает путешествие по морю, нахождение на корабле и, возможно, службу на флоте. Действительно, среди наиболее вероятных документов этого топика рассказы, в которых морская тематика играет ключевую роль – «Морской ветер» И. Соколова-Микитова и «Юнга» В. Билля-Белоцерковского. Остальные связываются с ними общими элементами пейзажа.

Топик 8, главная тема которого сформулирована здесь как t_8 «НА ФРОНТАХ ВОЙНЫ», оказался самым связным и явно сопоставляемым с включенными в него текстами (А. Фадеев «Рождение Амгуньского полка», А. Далматов «Гильза», Д. Фурманов «На Черном Ереке»). В этих рассказах повествуется о военных действиях, жизни на фронте как солдат, так и гражданских лиц, и потому выделяются соответствующие тематические слова: *полковник*, *матрос*, *начальник*.

Остальные топики вызывают сложности для тематической интерпретации.

Так, общую тему t_4 «ВЛАСТЬ ИМУЩИЕ» понять непосредственно из термов топика, без прочтения текстов, не представляется возможным.

В топике 5 выделяются две темы: *охота на зверя* (*винтовка*, *телега*, *шапка*, *зверь*) и *монастырская жизнь* (*монастырь*, *икона*), связь между которыми затемнена. Раздвоение темы может быть объяснено, по-видимому, общим контекстуальным фоном, изобилующим описаниями природы, леса и зверей. В качестве ведущей выбрана первая, t_5 «УБИЙСТВО ЗВЕРЯ», как более ясно выраженная в словарном составе топика. Аналогичная тематическая неоднородность отмечается и в топике 9, условно названном t_9 «ОХОТА НА ВОЛКА».

Топик 7 (t_7 «ПОЖАР РЕВОЛЮЦИИ») позволяет описать тему только одного рассказа – «Шахматы» Я. Брауна. Слова *игра* и *доска* связаны с игрой в шахматы, характерной только для этого произведения и происходящей в *кофейне*. Тем не менее, в рассказе присутствуют развернутые рассуждения о революции и ее влиянии на общественное устройство:

(4) *Мировая социалистическая* ***революция*** *– это стенка, огромная китайская стена, к которой поставят полтора миллиарда людей… Мировая* ***революция*** *– это подвал, в кото-*

[6] Здесь и далее примеры приводятся по Корпусу русского рассказа первой трети XX в.

рый провалятся под вашими, хе-хе, революционными руками сто тысяч дураков и пара десятков героев… Мировая ***революция*** *– это мировой погром под красным флагом…* (Я. Браун «Шахматы»)

Связь с другими рассказами может быть прослежена только на уровне общих деталей – например, *дым* (табачный, цигарки и др.).

В топике 10 (t_10 «ОБЩЕЕ СОБРАНИЕ») различается сразу три тематических подгруппы: первая – *лодка*, *утопленник*; вторая – *мастерская*, *портрет*; третья – *собрание*, *завод*, *речь*. Вероятно, еще можно объединить слова *библиотека* и *газета*. Однако, как видно, подгруппы мало связаны друг с другом и, скорее всего, либо описывают отдельные рассказы, либо связывают схожие в них эпизоды.

3. Результаты применения STM-модели

В табл. 2 представлены результаты структурного тематического моделирования (STM-1_10). В качестве ковариатов были выбраны ‘год’ и ‘период’ написания текстов (по разметке Корпуса).

Таблица 2

Результаты тематического моделирования STM-1_10

Table 2

Results of topic modeling with STM-1_10

№	Название темы	Содержание
t_1	ОХОТА В ЛЕСУ	лесной, сосна, дядя, зверь, цветы, туман, озеро, шапка, охота, лодка
t_2	ОБЕД	спинка, мужчина, матрос, вещь, обедать, песок, обед, бульвар, пробовать, счастливый
t_3	БОРЬБА С ВОЛКОМ	волк, рыжий, запах, звезда, борьба, стекло, тянуть, хвост, лезть, платок
t_4	РЕВОЛЮЦИЯ	икона, фабрика, тетка, монастырь, доска, игра, революция, завод, машина, война
t_5	ВОЙНА И ПЛЕН	винтовка, немец, война, окоп, лагерь, полковник, отряд, фронт, начальник, выстрел, пленный
t_6	НА СЦЕНЕ	кабинет, собственный, сцена, настоящий, известный, лестница, выражение, квартира, внимание, служащий
t_7	НОВОЕ ВРЕМЯ	библиотека, утопленник, капитан, лодка, завод, машина, полковник, агент, мастерская, шейка
t_8	В ТЮРЬМЕ	губернатор, тюрьма, власть, мрак, камера, телега, надзиратель, болото, крестьянин, начальник
t_9	ЗАСТОЛЬЕ	учитель, водка, выпить, добрый, праздник, видать, плохой, известный, вышка, жалко
t_10	ПЕРЕД КАЗНЬЮ	камера, нежный, номер, постель, вчера, квартира, кровать, закрыть, чистый, казнь

Как и моделью на основе LDA, выделяются темы, связанные с *охотой* и *природой* (t_1 «ОХОТА В ЛЕСУ» и t_3 «БОРЬБА С ВОЛКОМ»), а также две «гастрономические» темы – t_2 «ОБЕД» и t_9 «ЗАСТОЛЬЕ». Слова, составляющие топик 4, связаны с описанием нового, революционного времени и, вероятно, Гражданской войны, что позволяет определить тему t_4 «РЕВОЛЮЦИЯ». Помимо этого, выделяются темы *войны* (t_5 «ВОЙНА И ПЛЕН») и *казни* (t_10 «ПЕРЕД КАЗНЬЮ»).

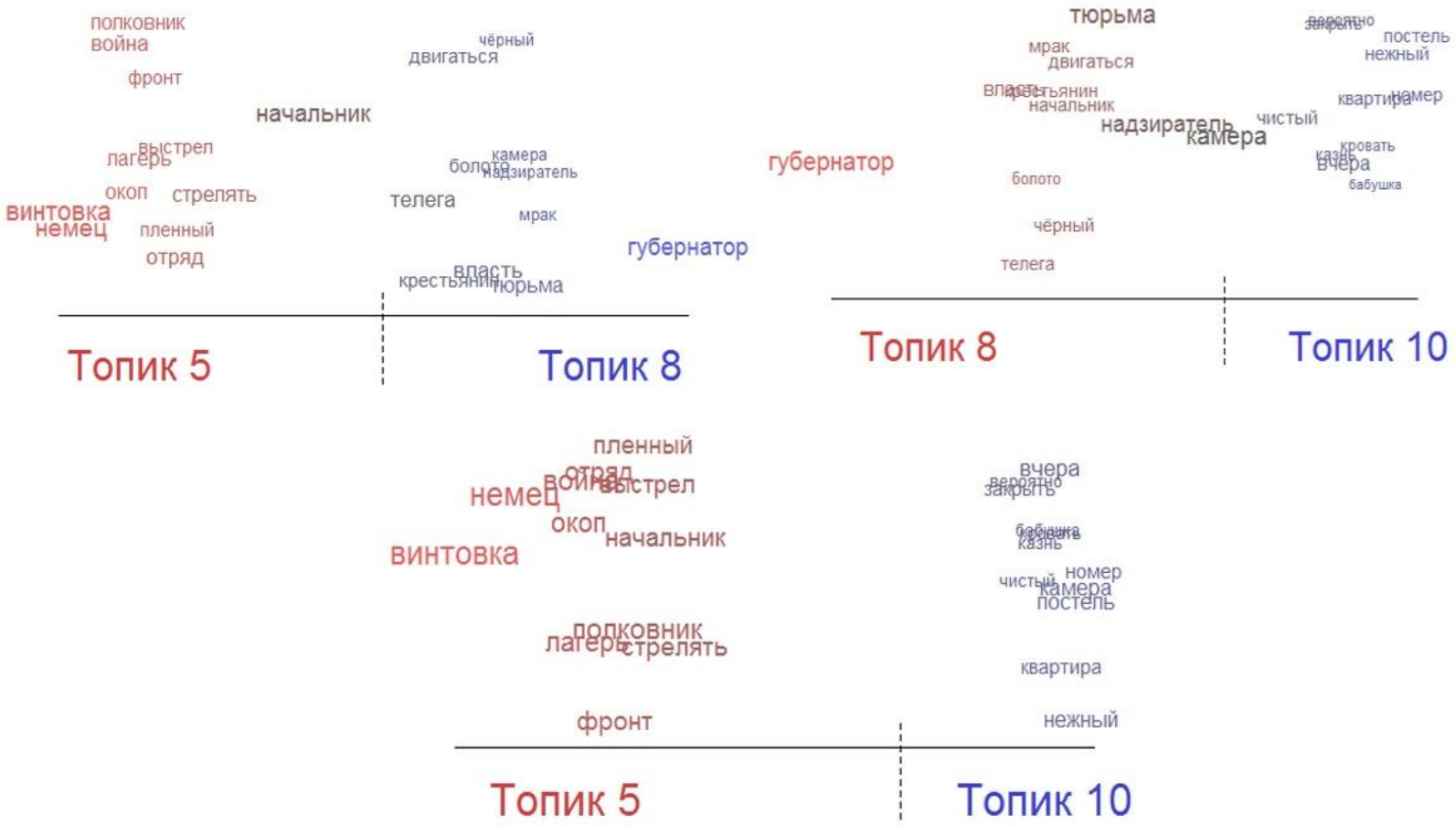

Рис. 1. Сравнение словарного состава топиков 5, 8 и 10 (STM_1-10)
Fig. 1. Comparison of word distribution of topics 5, 8 and 10 (STM_1-10)

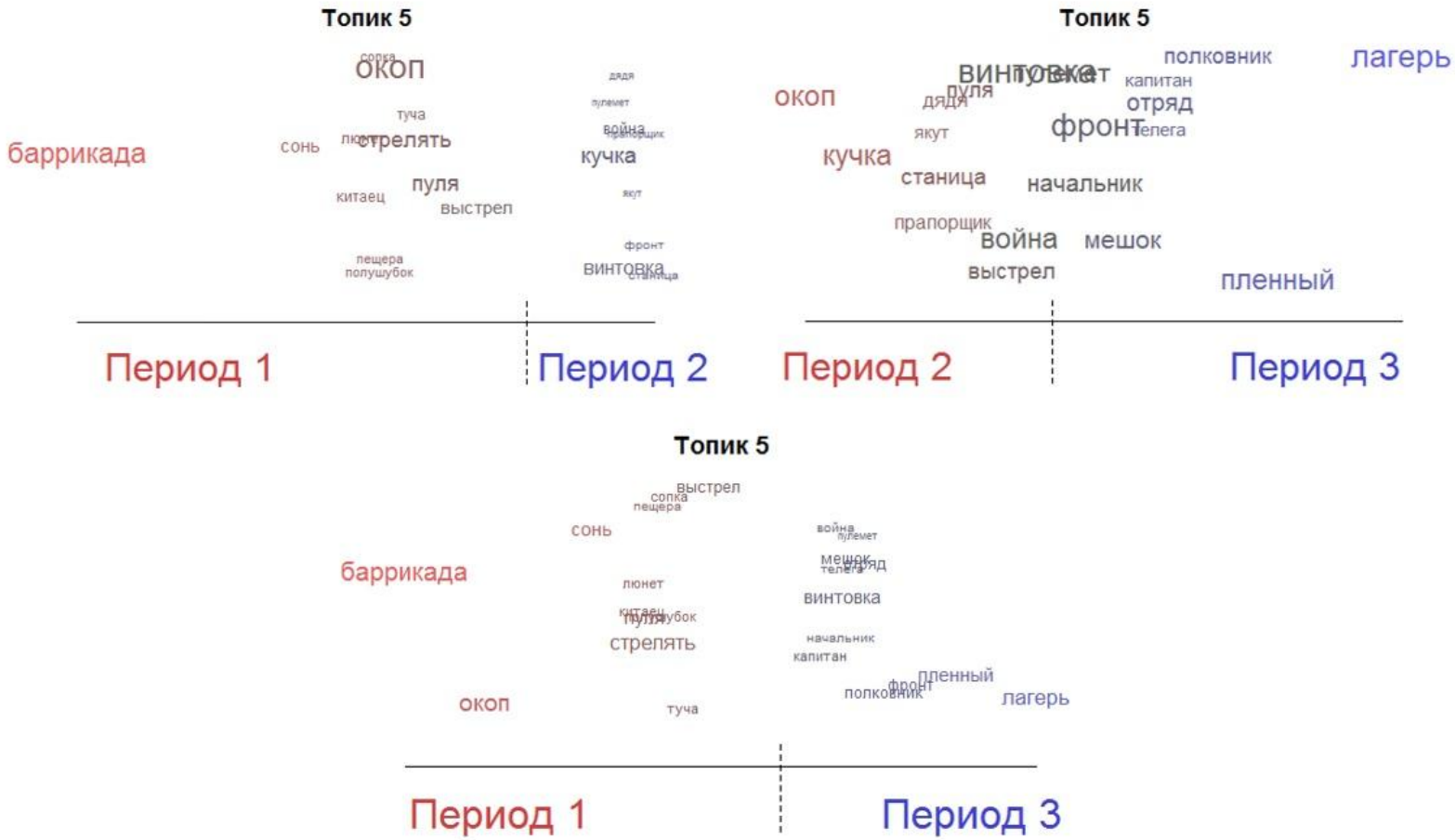

Рис. 2. Изменение словарного состава топика 5 по периодам (STM_2-10)
Fig. 2. Dynamics of word distribution of topic 5 within periods (STM_2-10)

Однако преимуществом принципа структурно-тематического моделирования является то, что оно позволяет учитывать влияние ковариатов на содержание топиков. Так, например, примечательно, что тема t_5 «ВОЙНА И ПЛЕН» лексически не связана с темой t_10 «ПЕРЕД КАЗНЬЮ» и мало сопоставима с темой t_8 «В ТЮРЬМЕ» (рис. 1). Темы t_10 «ПЕРЕД КАЗНЬЮ» и t_8 «В ТЮРЬМЕ», в свою очередь, объединяются друг с другом двумя общими терминами – *надзиратель* и *камера*, что позволяет говорить о схожих сюжетных элементах, характеризующих данные нарративы.

При этом при оценке влияния ковариата ‘год’ на пропорциональное распределение топиков по текстам тема t_5 «ВОЙНА И ПЛЕН» оказалась статистически значимой для 1920-х гг. Прослеживается изменение словарного состава темы в соответствии с историческим периодом, на который приходится время написания рассказа: с периода I (1900–1913) по период II (1914–1922) была популярна революционная и военная тематика (ср. *баррикада*, *окоп*, *стрелять*, *война*, *пуля*, *выстрел*), в то время как в период III (1923–1930) появляются также и рассказы, в которых описывается жизнь в лагере военнопленных (*фронт*, *пленный*, *лагерь*) (рис. 2).

4. Результаты применения NMF-модели

Построение тематической модели на существительных способствовало улучшению интерпретируемости результатов только в случае применения алгоритма неотрицательной матричной факторизации (NMF-2_10, табл. 3).

Таблица 3

Результаты тематического моделирования NMF-2_10

Table 3

Results of topic modeling with NMF-2_10

№	Название темы	Содержание
t_1	ПРЕДЧУВСТВИЕ СМЕРТИ	спинка, сознание, мозг, борьба, камера, пространство, казнь, ощущение, дыхание, зеркало
t_2	СВИДАНИЕ В НОМЕРЕ	дама, номер, звонок, приятель, диван, бутылка, сцена, гимназист, ресторан, класс
t_3	В ПЛЕНУ	библиотека, лагерь, утопленник, лодка, пленный, машина, кладбище, завод, золото, крестьянин
t_4	ПОСЛЕ АРЕСТА	губернатор, художник, бабушка, камера, надзиратель, баррикада, монах, шепот, картина, заимок
t_5	ОХОТА НА ВОЛКА	дядя, тетка, волк, сосна, охота, овраг, рыба, дружба, крыло, печь,
t_6	НА СЦЕНЕ ТЕАТРА И РЕВОЛЮЦИИ	сцена, игра, доска, степь, отряд, матрос, дацан, кофейня, публика, революция
t_7	ВОЙНА	немец, окоп, полковник, выстрел, мешок, команда, отряд, пуля, матрос, капитан
t_8	В УЕДИНЕНИИ	степь, монастырь, икона, топор, капитан, купец, озеро, казначей, колесо, монахиня
t_9	НА СЛУЖБЕ	фабрика, служащий, контора, болото, сторож, насыпь, грязь, бабушка, пари, пруд
t_10	ТЯЖЕЛЫЙ ТРУД	завод, станок, кучка, вышка, машина, факел, мрак, мастерская, шахта, конторщик

Кроме того, лишь для этой модели отмечается формирование тематических групп рассказов не только по принципу наличия в них общих сюжетных элементов и деталей, но также и схожих мотивов и приемов, использованных авторами. Рассмотрим несколько характерных случаев.

Топик 1 (t_1 «ПРЕДЧУВСТВИЕ СМЕРТИ») четко выделяет тему *казни* и нахождения в *камере*. Однако такие слова, как *сознание*, *мозг*, *ощущение*, *дыхание*, позволяют судить о том, что здесь также выражена и тема *ожидания смерти*, ее предвосхищения. Действительно, на примере приведенных ниже фрагментов из рассказов можно заметить, что NMF удалось распознать эту тему даже на уровне развернутой метафоры «гаснущее сознание – смерть», хотя выделение темы смерти часто представляет трудности ввиду ее эпизодичности [Sherstinova et al., 2020]:

(5) ***Сознание*** *погасло, как потухающий разбросанный костер, холодело, как труп только что скончавшегося человека, у которого тепло еще в сердце, а ноги и руки уже окоченели.* (Л. Андреев «Рассказ о семи повешенных»)

(6) *Потом все исчезло: и мысль, и* ***сознание****, и боль, и тоска. И это случилось так же просто и быстро, как если бы кто дунул на свечу, горевшую в темной комнате, и погасил ее...* (А. Куприн «В цирке»)

Примечательна вербализация темы t_6 «НА СЦЕНЕ ТЕАТРА И РЕВОЛЮЦИИ», связывающая игру на сцене с игрой в шахматы, чего не было сделано другими моделями. Причем *игра* здесь представляет пример не полисемии, а скорее мотивной структуры, характеризуя сразу несколько рассказов, включенных в этот топик, и являясь значимой для их понимания (Я. Браун «Шахматы», В. Шишков «Спектакль в селе Огрызове», Скиталец «Любовь декоратора»).

Подобное замечается и в отношении топика 10 (t_10 «ТЯЖЕЛЫЙ ТРУД»), относящегося к теме *труда*. Метод NMF позволил выделить в нем тему тяжелого, изнурительного труда, монотонной работы, причем и на заводе (*завод*, *станок*, *машина*, *мастерская*), и в шахте (*факел*, *мрак*, *шахта*, *вышка*), и в конторе (*конторщик*).

5. Выводы

В анализируемой текстовой коллекции все сравниваемые тематические модели выделяют темы *революции*, *войны*, *плена*, *казни*, *тюрьмы*, *природы* и *охоты*. Темы *романтических отношений* и *застолья* были выделены LDA-моделью; STM-модель смогла обнаружить только тему *застолья*, а NMF – тему *романтических отношений*. Вне зависимости от алгоритма в явном виде определяются конкретно-исторические топики, связанные с войной и пребыванием на фронте, а также темы, составляющие общий предметный фон произведения, например, связанные с описанием природы.

Установлено, что выделение моделями тем войны и природы объясняется не столько значительным числом рассказов этой тематики в выборке [7], сколько тематической однородностью и лексической специфичностью этих текстов. К такому выводу подталкивает сравнение с результатами работы моделей, полученными для тем взаимоотношений, смерти и любви. На каждую из них, по экспертному заключению, приходится более 100 рассказов, однако определение этих тем в ходе автоматического моделирования путем анализа только слов, составляющих топики, практически невозможно. Причин может быть несколько: разнообразие инструментов, используемых авторами, для ввода «общих» тем в повествование; политематичность произведений, в которых они встречаются; второстепенная роль / эпизодичность

[7] Согласно экспертной разметке, в подкорпусе насчитывается 60 рассказов военной тематики и 52 рассказа – природной [Sherstinova et al., 2020].

этих тем. Возможно, увеличение объема выборки позволит генерализировать различные приемы конструирования подобных тем.

В отношении особенностей функционала рассмотренных алгоритмов можно сделать следующие выводы. Метод LDA наиболее успешно справляется с выделением общих тем и идентификацией мест, где происходит действие (тема *природы* и тема *застолья*). NMF позволил объединять рассказы по слабо выраженным темам и мотивным структурам (тема *предвосхищения смерти* и тема *игры*), а также различать «схожие», т. е. ассоциативно связанные (например, тема *войны* и тема *плена*, тема *казни* и тема *тюрьмы*).

Стоит также отметить, что NMF оказался единственным методом, показавшим значительное улучшение интерпретируемости модели, построенной на выборке, содержащей только существительные.

И, наконец, алгоритм STM (структурно-тематическое моделирование) позволил сформировать достаточно дистинктивный словарный состав тем, в частности тема *казни* и тема *тюрьмы*.

Заключение

В статье рассмотрен эксперимент, направленный на сравнение результатов тематического моделирования на основе метода латентного размещения Дирихле (LDA), структурно-тематического моделирования (STM) и неотрицательной матричной факторизации (NMF) на материале художественных текстов малой формы. Помимо этого, было оценено и влияние на интерпретируемость топиков двух лингвистических видов выборки по частеречному параметру: 1) содержащей все знаменательные части речи и 2) содержащей только существительные.

Сравнение трех алгоритмов тематического моделирования привело к некоторым интересным наблюдениям с точки зрения особенностей их функционала на материале художественных текстов. Несмотря на то что моделирование на основе LDA смогло эффективно выявить основные темы корпуса, оно оказалось менее продуктивным, по сравнению с STM и NMF, в задаче различения нескольких тем внутри большей категории. Так, и STM, и NMF отнесли лексику, характеризующую темы *заключения в тюрьме* и *казни*, к разным топикам, а LDA – к одному. В связи с этим STM- и NMF-модели в задачах анализа художественных текстов обладают, на наш взгляд, бóльшим потенциалом.

Что касается варьирования подходов к формированию выборки, то улучшение интерпретируемости модели было замечено в случае NMF-модели. Топики NMF-модели, построенной на существительных, оказались более детализированными для соответствующих групп рассказов. Стоит также сказать, что именно при таком варианте предобработки NMF-моделью были выявлены так называемые «нишевые» топики. Модель «только на существительных» связала рассказы скорее на уровне мотивов, а не на основании общих сюжетных элементов, что наблюдалось для модели «на всех частях речи» (после удаления служебных и стоп-слов). Достичь подобного результата для других пар моделей не удалось, по нашему предположению, ввиду ошибок, допущенных при автоматической частеречной разметке.

Как правило, интерпретация лексического наполнения топика сводится к выводу из него тематических элементов на некотором «усредненном» уровне абстракции. В отношении тематического моделирования художественных текстов исследователь сталкивается со следующей проблемой: те абстрактные категории, на которые указывают топики, не обязательно отражают тематику произведения. Как демонстрируют эксперименты, проведенные на небольших выборках, причина кроется не только в том, что абстрактный образный язык затрудняет понимание содержания топиков, но и в том, что выявляемые паттерны в принципе более разнообразны. Тематические модели, построенные на литературном материале, могут включать информацию и о других семантических структурах, объединять тексты не только

тематически, но и на мотивном и сюжетном уровнях (место / время действия, сходные художественные детали и стилистические приемы).

В дальнейшем представляется целесообразным построение моделей на базе алгоритмов STM и NMF с большим количеством тем и на расширенной выборке. Выдвигается предположение, что достижению лучших результатов может способствовать сегментация рассказов значительного размера [8], что должно предотвратить их выделение в индивидуальные топики (как в случае рассказа «Шахматы» Я. Брауна и «Рассказа о семи повешенных» Л. Андреева). Интерес также представляют способы нейтрализации влияния лингвистического вида выборки – «все части речи» vs «только существительные» – на результаты тематического моделирования. Обсуждается, что для выявления причин этого влияния необходимо сопоставить полученные на данном этапе тематические модели с моделями на текстах, которые пройдут предварительную частеречную обработку, но с использованием другого морфоанализатора.

Список литературы

Ерофеева А. Р., Митрофанова О. А. Автоматическое назначение меток тем в тематических моделях русскоязычных корпусов текстов // Структурная и прикладная лингвистика. СПб.: Изд-во СПбГУ, 2019. С. 122–147.

Мартыненко Г. Я., Шерстинова Т. Ю., Мельник А. Г., Попова Т. И. Методологические проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) // Компьютерная лингвистика и вычислительные онтологии. 2018а. № 2. С. 97–102.

Мартыненко Г. Я., Шерстинова Т. Ю., Попова Т. И., Мельник А. Г., Замирайлова Е. В. О принципах создания корпуса русского рассказа первой трети XX века // Тр. XV Междунар. конф. по компьютерной и когнитивной лингвистике «TEL-2018». Казань, 2018б. С. 180–197.

Митрофанова О. А. Моделирование тематики специальных текстов на основе алгоритма LDA // XLII Междунар. филол. конф. СПб., 2014. С. 220–233.

Митрофанова О. А. Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М. А. Булгакова // Корпусная лингвистика – 2019. СПб., 2019. С. 387–394.

Томашевский Б. В. Теория литературы. Поэтика: Учеб. пособие. М.: Аспект Пресс, 1996. С. 176–192.

Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 2003, vol. 3, pp. 993–1022.

Da, N. Z. The Computational Case against Computational Literary Studies. *Critical Inquiry*, 2019, vol. 45, no. 3, pp. 601–639.

Gaujoux, R., Seoighe, C. A Flexible R package for Nonnegative Matrix Factorization. *BMC Bioinformatics*, 2010, vol. 11, no. 1, pp. 1–9.

Grün, B., Hornik, K. Topicmodels: An R package for Fitting Topic Models. *Journal of Statistical Software*, 2011, vol. 40, no. 13, pp. 1–30.

Gryaznova, E., Kirina, M. Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900–1930. In: Proc. of International Conference “Internet and Modern Society”, 2021, pp. 281–290.

[8] Влияние размера текста на интерпретируемость тематической модели представляет собой отдельную и на данный момент не до конца разработанную проблему тематического моделирования литературных текстов, в особенности малых жанров (например, см. [Navarro-Colorado, 2018]).

Guo, F., Metallinou, A., Khatri, C., Raju, A., Venkatesh, A., Ram, A. Topic-based Evaluation for Conversational Bots. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, 2018, arXiv preprint arXiv:1801.03622.

Huang, T. C., Hsieh, C. H., Wang, H. C. Automatic Meeting Summarization and Topic Detection System. In: Data Technologies and Applications, 2018, pp. 351–365.

Jacobs, T., Tschötschel, R. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 2019, vol. 22, no. 5, pp. 469–485.

Jockers, M. L., Mimno, D. Significant themes in 19th-century literature. *Poetics*, 2013, vol. 41, no. 6, pp. 750–769.

Lau, J. H., Newman, D., Karimi, S., Baldwin, T. Best Topic Word Selection for Topic Labelling. In: Proc. of the 23rd Int. Conf. on Computational Linguistics, Association for Computational Linguistics. Stroudsburg, PA, 2010, pp. 605–613.

Lee, D., Seung, H. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 1999, vol. 401, pp. 788–791.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W. An Overview of Topic Modeling and Its Current Applications in Bioinformatics. *SpringerPlus*, 2016, vol. 5, no. 1, pp. 1–22.

Martynenko, G., Sherstinova, T. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). St. Petersburg, 2020, vol. 2552, pp. 105–120.

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., Jurafsky, D. Differentiating language usage through topic models. *Poetics*, 2013, vol. 41, no. 6, pp. 607–625.

Mitrofanova, O. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In: International Workshop on Language, Music, and Computing. Springer, Cham, 2015, pp. 69–76.

Moubayed, N. A., Breckon, T., Matthews, P., McGough, A. S. SMS Spam Filtering Using Probabilistic Topic Modelling and Stacked Denoising Autoencoder. In: International Conference on Artificial Neural Networks. Springer, Cham, 2016, pp. 423–430.

Navarro-Colorado, B. On Poetic Topic Modeling: Extracting Themes and Motifs from a Corpus of Spanish Poetry. *Frontiers in Digital Humanities*, 2018, vol. 5, pp. 5–15.

Nikolenko, S. I., Koltsov, S., Koltsova, O. Topic Modelling for Qualitative Studies. *Journal of Information Science*, 2017, vol. 43, no. 1, pp. 88–102.

O'Callaghan, D., Greene, D, Carthy, J., Cunningham, P. An Analysis of the Coherence of Descriptors in Topic Modeling. *Expert Systems with Applications (ESWA)*, 2015, vol. 42, no. 13, pp. 5645–5657.

Rana, T. A., Cheah, Y. N., Letchmunan, S. Topic Modeling in Sentiment Analysis: A Systematic Review. *Journal of ICT Research & Applications*, 2016, vol. 10, no. 1, pp. 76–93.

Rhody, L. M. Topic Modelling and Figurative Language. *Journal of Digital Humanities*, 2012, pp. 19–35.

Roberts, M., Stewart, B., Tingley, D. Airoldi, E. The Structural Topic Model and Applied Social Science. *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*, 2013, pp. 1–20.

Roberts, M., Stewart, B., Tingley, D. STM: An R package for structural topic models. *Journal of Statistical Software*, 2019, no. 91.1, pp. 1–40.

Schöch, C. Topic modeling genre: an exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*, 2017, vol. 11, no. 2. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Sherstinova, T., Mitrofanova, O., Skrebtsova, T., Zamiraylova, E., Kirina, M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction. *Advances in Com-*

putational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, 2020, vol. 12469, pt. 2, pp. 134–152.

Sherstinova, T., Moskvina, A., Kirina, M. Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian. *Proc. of the 29th Conference of Open Innovations Association FRUCT*, 2021, pp. 313–323.

Straka, M., Straková, J. Universal Dependencies 2.5 Models for UDPipe (2019-12-06). In: LINDAT / CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University, 2019. URL: http:// hdl. handle.net/11234/1-3131

Uglanova, I., Gius, E. The Order of Things. A Study on Topic Modelling of Literary Texts. *Proc. of the CHR 2020: Workshop on Computational Humanities Research, CEUR Workshop Proceedings*, 2020, pp. 57–76.

Wijffels, J. UDPipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.4-1. 2020.

Zamiraylova, E., Mitrofanova, O. Dynamic topic modeling of Russian fiction prose of the first third of the 20th century by means of non-negative matrix factorization. *Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019)*, 2020, vol. 2552, pp. 321–339.

Список источников

Корпус русского рассказа 1900–1930 гг. URL: https://russian-short-stories.ru/.

References

Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 2003, vol. 3, pp. 993–1022.

Da, N. Z. The Computational Case against Computational Literary Studies. *Critical Inquiry*, 2019, vol. 45, no. 3, pp. 601–639.

Erofeeva, A., Mitrofanova, O. Automatic assignment of topic labels in topic models for Russian text corpora. In: Structural and Applied Linguistics. St. Petersburg Uni. Press, 2019, pp. 122–147. (in Russ.)

Gaujoux, R., Seoighe, C. A Flexible R package for Nonnegative Matrix Factorization. *BMC Bioinformatics*, 2010, vol. 11, no. 1, pp. 1–9.

Gryaznova, E., Kirina, M. Defining Kinds of Violence: A Comparison of Topic Modelling with Latent Dirichlet Allocation and Principal Component Analysis for Russian Short Stories of 1900–1930. In: Proc. of International Conference "Internet and Modern Society", 2021, pp. 281–290.

Grün, B., Hornik, K. Topicmodels: An R package for Fitting Topic Models. *Journal of Statistical Software*, 2011, vol. 40, no. 13, pp. 1–30.

Guo, F., Metallinou, A., Khatri, C., Raju, A., Venkatesh, A., Ram, A. Topic-based Evaluation for Conversational Bots. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, 2018, arXiv preprint arXiv:1801.03622.

Huang, T. C., Hsieh, C. H., Wang, H. C. Automatic Meeting Summarization and Topic Detection System. In: Data Technologies and Applications, 2018, pp. 351–365.

Jacobs, T., Tschötschel, R. Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, 2019, vol. 22, no. 5, pp. 469–485.

Jockers, M. L., Mimno, D. Significant themes in 19th-century literature. *Poetics*, 2013, vol. 41, no. 6, pp. 750–769.

Lau, J. H., Newman, D., Karimi, S., Baldwin, T. Best Topic Word Selection for Topic Labelling. In: Proc. of the 23rd Int. Conf. on Computational Linguistics, Association for Computational Linguistics. Stroudsburg, PA, 2010, pp. 605–613.

Lee, D., Seung, H. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 1999, vol. 401, pp. 788–791.

Liu, L., Tang, L., Dong, W., Yao, S., Zhou, W. An Overview of Topic Modeling and Its Current Applications in Bioinformatics. *SpringerPlus*, 2016, vol. 5, no. 1, pp. 1–22.

Martynenko, G. Ya., Sherstinova, T. Yu., Melnik, A. G., Popova, T. I. Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century)]. In: Computational Linguistics and Computational Ontologies. ITMO University. St. Petersburg, 2018a, iss. 2, pp. 97–102. (in Russ.)

Martynenko, G. Ya., Sherstinova, T. Yu., Popova, T. I., Melnik, A. G., Zamirajlova, E. V. On the principles of the Creation of the Russian Short Story Corpus of the First Third of the 20th Century]. In: Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics 'TEL 2018'. Kazan, 2018b, pp. 180–197. (in Russ.)

Martynenko, G., Sherstinova, T. Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. In: R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019). St. Petersburg, 2020, vol. 2552, pp. 105–120. (in Russ.)

McFarland, D. A., Ramage, D., Chuang, J., Heer, J., Manning, C. D., Jurafsky, D. Differentiating language usage through topic models. *Poetics*, 2013, vol. 41, no. 6, pp. 607–625.

Mitrofanova, O. A. Analysis of Fiction Text Structure by Means of Topic Modelling: A Case Study of "Master and Margarita" Novel by M. A. Bulgakov]. In: Corpus Linguistics – 2019. St. Petersburg, 2019, pp. 387–394. (in Russ.)

Mitrofanova, O. A. Topic modelling of special texts based on LDA algorithm]. In: Proceedings of XLII International Philological Conference. Selected works. St. Petersburg, 2014, pp. 220–233. (in Russ.)

Mitrofanova, O. Probabilistic Topic Modeling of the Russian Text Corpus on Musicology. In: International Workshop on Language, Music, and Computing. Springer, Cham, 2015, pp. 69–76.

Moubayed, N. A., Breckon, T., Matthews, P., McGough, A. S. SMS Spam Filtering Using Probabilistic Topic Modelling and Stacked Denoising Autoencoder. In: International Conference on Artificial Neural Networks. Springer, Cham, 2016, pp. 423–430.

Navarro-Colorado, B. On Poetic Topic Modeling: Extracting Themes and Motifs from a Corpus of Spanish Poetry. *Frontiers in Digital Humanities*, 2018, vol. 5, pp. 5–15.

Nikolenko, S. I., Koltsov, S., Koltsova, O. Topic Modelling for Qualitative Studies. *Journal of Information Science*, 2017, vol. 43, no. 1, pp. 88–102.

O'Callaghan, D., Greene, D, Carthy, J., Cunningham, P. An Analysis of the Coherence of Descriptors in Topic Modeling. *Expert Systems with Applications (ESWA)*, 2015, vol. 42, no. 13, pp. 5645–5657.

Rana, T. A., Cheah, Y. N., Letchmunan, S. Topic Modeling in Sentiment Analysis: A Systematic Review. *Journal of ICT Research & Applications*, 2016, vol. 10, no. 1, pp. 76–93.

Rhody, L. M. Topic Modelling and Figurative Language. *Journal of Digital Humanities*, 2012, pp. 19–35.

Roberts, M., Stewart, B., Tingley, D. Airoldi, E. The Structural Topic Model and Applied Social Science. *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*, 2013, pp. 1–20.

Roberts, M., Stewart, B., Tingley, D. STM: An R package for structural topic models. *Journal of Statistical Software*, 2019, no. 91.1, pp. 1–40.

Schöch, C. Topic modeling genre: an exploration of French classical and enlightenment drama. *Digital Humanities Quarterly*, 2017, vol. 11, no. 2. URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

Sherstinova, T., Mitrofanova, O., Skrebtsova, T., Zamiraylova, E., Kirina, M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction. *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020*, 2020, vol. 12469, pt. 2, pp. 134–152.

Sherstinova, T., Moskvina, A., Kirina, M. Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian. *Proc. of the 29th Conference of Open Innovations Association FRUCT*, 2021, pp. 313–323.

Straka, M., Straková, J. Universal Dependencies 2.5 Models for UDPipe (2019-12-06). In: LINDAT / CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL). Faculty of Mathematics and Physics, Charles University, 2019. URL: http:// hdl. handle.net/11234/1-3131

Tomashevsky, B. The Theory of Literature. Moscow, Aspect Press, 1996, pp. 176–192. (in Russ.)

Uglanova, I., Gius, E. The Order of Things. A Study on Topic Modelling of Literary Texts. *Proc. of the CHR 2020: Workshop on Computational Humanities Research, CEUR Workshop Proceedings*, 2020, pp. 57–76.

Wijffels, J. UDPipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.4-1. 2020.

Zamiraylova, E., Mitrofanova, O. Dynamic topic modeling of Russian fiction prose of the first third of the 20th century by means of non-negative matrix factorization. *Proc. of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL-2019)*, 2020, vol. 2552, pp. 321–339.

List of Sources

Corpus of Russian Short Stories of 1900–1930s. URL: https://russian-short-stories.ru/.

Информация об авторе

Маргарита Александровна Кирина, магистрант

Information about the Author

Margarita A. Kirina, Master's Student

Статья поступила в редакцию 05.12.2021;
одобрена после рецензирования 10.04.2022; *принята к публикации 20.04.2022*
The article was submitted 05.12.2022;
approved after reviewing 10.04.2022; *accepted for publication 20.04.2022*