DOI 10.37882/2223-2966.2023.12-2.05

ИМЕНОВАНИЕ КЛАСТЕРОВ, ПОСТРОЕННЫХ НА КОРПУСЕ ЕСТЕСТВЕННЫХ ЯЗЫКОВ

NAMING CLUSTERS BUILT ON A CORPUS OF NATURAL LANGUAGES

L. Bilgaeva
Yu. Dmitriev

Summary. The article conducted research on the use of the LDA method to solve the problem of thematic modeling of citizens' appeals to the social fund of Russia and the naming of clusters obtained as a result of modeling using ChatGPT artificial intelligence.

Keywords: topic modeling, Latent Dirichlet allocation, LDA, naming clusters.

Бильгаева Людмила Пурбоевна
Доцент, ФГБОУ ВО «Восточно-Сибирский государственный университет технологий и управления»
bilgaeval@mail.ru
Дмитриев Юрий Александрович
Магистрант, ФГБОУ ВО «Восточно-Сибирский государственный университет технологий и управления»
dya@sibdigital.net

Аннотация. В статье проведены исследования по применению метода LDA для решения задачи тематического моделирования обращений граждан в социальный фонд России и именование полученных в результате моделирования кластеров с использованием искусственного интеллекта ChatGPT.

Ключевые слова: тематическое моделирование, латентное размещение Дирихле, LDA, именование кластеров.

Введение

В современном информационном пространстве, где объем текстовых данных постоянно растет, задача эффективной организации и анализа этой информации становится ключевой. Тематическое моделирование, как мощный инструмент анализа текста, имеет широкий спектр практических применений. Этот процесс позволяет структурировать и классифицировать текстовые данные на основе их содержания, путем присвоения текстам тематических меток [1].

Тематическое моделирование обеспечивает структурирование и классификацию текстовых данных в виде тем или кластеров. Однако для полного понимания и использования этих кластеров необходимо их явное и информативное именование. Именование кластеров придает им смысл и помогает пользователям понимать, о чем эти кластеры, делая их более интерпретируемыми [2].

Решение задачи именования кластеров предлагается на кластерах, построенных на корпусе естественно-языковых текстов. В связи с этим были рассмотрены такие методы, как латентно-семантический анализ, латентное размещение Дирихле, кластеризация суффиксного дерева [2, 3].

Для анализа рассмотренных методов были определены критерии, представляющие интерес при построении кластеров: высокое быстродействие, выявление скрытой семантики, пересекаемость кластеров, наглядность результата (Табл. 1).

Таблица 1.

Методы кластеризации текста и критерии их оценки

N	Методы	Критерии			
		Высокое быстродействие	Выявление скрытой семантики	Пересекаемость кластеров	Наглядность результата
1	Латентно-семантический анализ (LSA)	–	+	–	+ –
2	Латентное размещение Дирихле (LDA)	+ –	+	+	+
3	Кластеризация суффиксного дерева (STC)	+	–	+	+

Из таблицы видно, что наиболее подходящим методом для решения поставленной задачи тематического моделирования является метод латентного размещения Дирихле, имеющий наиболее приемлемые критерии. Этот метод основан на вероятностной модели, которая позволяет выделять темы и их распределение в текстовых данных, при этом обеспечивается надежность и стабильность результатов. Важно отметить, что интерпретируемость LDA — одно из его сильных преимуществ. Этот метод позволяет создавать интерпретируемые темы, представленные ключевыми словами, что делает их более доступными для человеческого понимания. Бо-

лее того, LDA может обнаруживать новые темы, которые ранее не были известны.

В данной статье представлены результаты применения метода LDA для тематического моделирования обращений граждан в социальный фонд России (СФР) и именование полученных в результате моделирования кластеров. Для решения данной задачи использовались данные, представляющие 6399 обращений граждан в социальный фонд России за две недели 2019 года, что составляет 692103 слов.

Метод LDA

Метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) — это статистический метод, который на основе байесовской вероятности позволяет обнаруживать скрытые (латентные) темы из коллекции документов. Основная идея LDA состоит в том, что каждый документ в коллекции представляет собой случайный набор различных тем, которые в свою очередь являются распределением по словам [2]. В модели LDA каждая скрытая тема представляется как вероятностное распределение по словам, а распределение слов в темах имеет распределение Дирихле.

В корпусе D из M документов, где каждый документ d состоит из N_d термов ($d \in 1, \ldots, M$), а Q –количество различных термов во всем корпусе документов, предполагает следующий процесс генерации [4].

1. Выбирается полиномиальное распределение φ_i топика t_i ($t_i \in \{1, \ldots, T\}$) с гиперпараметром β

$$\varphi_i \sim Dir(\beta)\ (\varphi_i \in \Delta_Q)$$

2. Выбирается полиномиальное распределение θ_j для топика d_j ($d_j \in \{1, \ldots, M\}$) с гиперпараметром α

$$\theta_j \sim Dir(\alpha)\ (\varphi_i \in \Delta M)$$

3. Для каждого терма ω_n ($n \in \{1, \ldots, N_d\}$) в документе d_j:

— выбрать топик z_n из θ_j

$$z_n \sim Mult(\theta_j)\ (z_n \in T)$$

— выбрать слово ω_n из φ_{z_n}

$$\omega_n \sim Mult(\varphi_{z_n})\ (\omega_n \in W)$$

В процессе генерации, слова в документе являются наблюдаемыми переменными, в то время как $\alpha \in \mathbb{R}^M$ и $\beta \in \mathbb{R}^Q$ являются гиперпараметрами, отвечающими за выраженность топиков в документах и определяющими разреженность векторов, описывающих распределение слов в топике соответственно, а φ и θ — скрытыми переменными.

Архитектура программного обеспечения

Для решения задачи тематического моделирования была разработана программа с использованием библиотеки Gensim, которая представляет собой микросервисную архитектуру (рис. 1).

В состав программы входит база данных PostgreSQL, Java-сервис, Python-сервис, являющийся модулем тематического моделирования, брокер сообщений RabbitMQ.

База данных содержит обращения пользователей, список стоп-слов, параметры тематического моделирования и результаты выполнения тематического моделирования.

Java-сервис включает следующие функции: авторизация; валидация входных данных; выбор периода обращений для выполнения тематического моделирования; настройка параметров для генерации модели LDA; настройка параметров запроса ChatGPT; настройка предварительной обработки слов; обновление словаря стоп-слов; историческое отображение предыдущих результатов тематического моделирования.

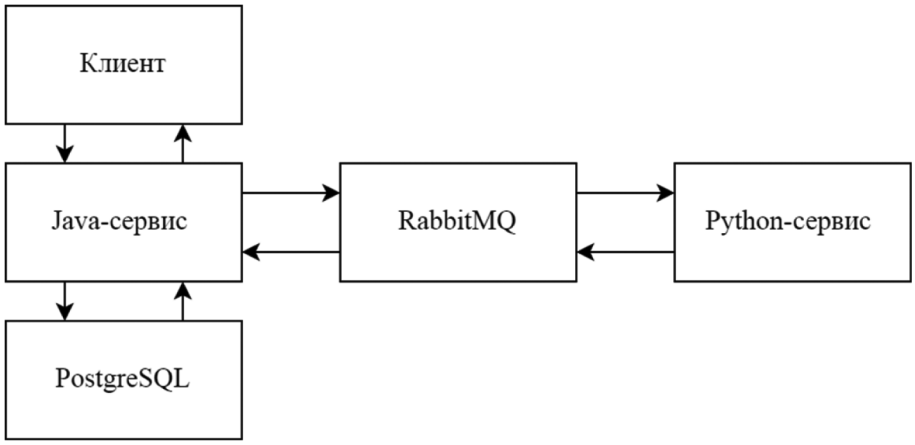

Рис. 1. Архитектура программного обеспечения

Python-сервис состоит из функций тематического моделирования и запросов к ChatGPT.

С помощью брокера сообщений RabbitMQ осуществляется связь между Java-сервисом и Python-сервисом.

Предварительная обработка текста

При работе с данными из СФР, транскрибированные записи обращений граждан и ответы операторов хранятся в разных столбцах базы данных. Первоначальная попытка объединения этих текстов для создания датасета привела к увеличению его объема почти вдвое, достигнув 1433140 слов. Однако увеличение размера датасета повлекло за собой низкую оценку согласованности модели LDA. Эта оценка позволяет оценить, насколько хорошо модель LDA разделяет тексты на различные темы и насколько интерпретируемыми и согласованными оказываются эти темы.

При визуальном анализе ключевых слов в сформированных кластерах было замечено, что появляются часто повторяющиеся шаблонные слова оператора, которые лишены информационной ценности и представляют собой информационный шум. Для решения этой проблемы можно рассмотреть возможность расширения списка стоп-слов, чтобы включить в него эти часто встречающиеся слова. Однако такой список может стать слишком объемным. Кроме того, при анализе самих обращений становится ясно, что по ответам операторов довольно трудно определить, с какой конкретной проблемой обратились граждане. Поэтому было решено исключить тексты операторов из датасета, что привело к увеличению оценки согласованности, существенному снижению нагрузки на вычислительные ресурсы и улучшению качества формирования кластеров и ключевых слов.

Тем не менее, даже после этого сокращения датасет остается подверженным шумам, и поэтому требуется тщательная предобработка текстов для эффективного построения модели LDA.

Предварительная обработка служит для улучшения качества анализа текста и важна по следующим причинам.

1. Очистка и структурирование текста позволяет создать более качественную тематическую модель. Это связано с тем, что модель LDA опирается на статистику слов и их взаимосвязей, поэтому чистые и информативные тексты дают более точные результаты.
2. Предварительная обработка уменьшает размерность текстовых данных. Удаление стоп-слов, лемматизация и фильтрация редких слов сокращают количество уникальных слов, с которыми модель должна работать. Это ускоряет процесс и снижает потребление вычислительных ресурсов.
3. Очищенные тексты исключают ненужные шумы и делают тематические модели более интерпретируемыми. Темы и кластеры становятся более четкими и понятными для людей.
4. Построение биграмм и триграмм позволяет учесть контекст и связи между словами. Это особенно важно для анализа текстов, где слова часто встречаются в определенных фразах или сочетаниях.
5. Модель LDA может обрабатывать качественные и чистые данные более эффективно, что позволяет быстрее обучать модель и получать результаты анализа.

Таким образом, предварительная обработка текстовых данных играет решающую роль в тематическом моделировании. Она помогает создать качественные и интерпретируемые модели, делает результаты более точными и понятными, а также повышает эффективность анализа текстовых данных. К основным этапам предварительной обработки данных относятся такие этапы, как очистка слов, лемматизация, токенизация, фильтрация.

Очистка слов

Этап очистки слов включает в себя два важных шага:
— удаление знаков препинания и цифр;
— приведение к нижнему регистру.

Следует отметить, что для транскрибированных текстов этап очистки является излишним, однако этап очистки важен для текстовых данных из электронных писем, чат-ботов и других источников, которые могут содержать знаки препинания и цифры. Поэтому включение этого функционала делает алгоритм более универсальным и готовым к разнообразным источникам данных.

Лемматизация

Этап лемматизации текста предполагает приведение слов к начальной форме. Лемматизация сохраняет семантическую точность слов и обеспечивает более читаемые и понятные ключевые слова кластеров, что важно при анализе и интерпретации данных.

Для лемматизации текста были рассмотрены две библиотеки: PyMorphy 2 и PyMystem 3. Был выполнен сравнительный анализ использования этих библиотек по быстродействию, который представлен на рисунке 2.

Как видно из диаграммы, несмотря на заявленную точность, библиотека PyMorphy 2 работает медленнее, чем PyMystem 3 почти в 12 раз на одном и том же датасете. Учитывая, что объемы данных могут быть значительно больше, предпочтение отдается PyMystem 3 как более эффективной и быстрой библиотеке.

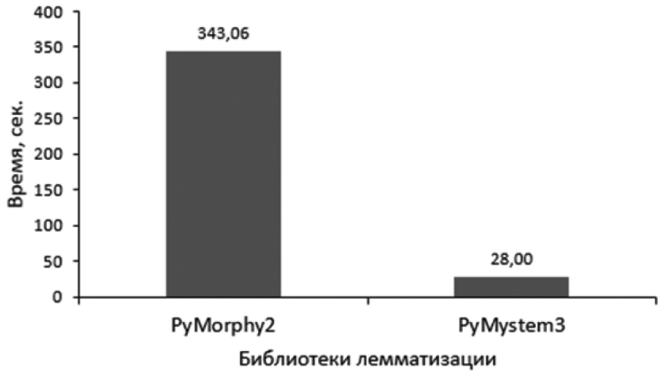

Рис. 2. Лемматизация текста

Фильтрация

Процесс фильтрации включает удаление слов и текстов. Фильтрация начинается с вычисления частоты встречаемости слов в текстах и общего количества слов. Это позволяет определить, какие слова более распространены в текстовом корпусе, а какие менее. Затем применяются условия фильтрации, которые позволяют определить, какие слова следует оставить, а какие удалить.

В процессе проведенных экспериментов было определено, какие критерии фильтрации важно учитывать.

1. Редко встречаемые слова. В процессе проведенных экспериментов был установлен порог встречаемости слова, который составляет 0,0001. Это означает, что для рассматриваемого датасета из 692103 слов, слово должно встретиться более 70 раз.
2. Короткие слова. В результате анализа текстов выявлено, что слова, имеющие менее 2 букв, чаще всего являются неинформативными и подлежат удалению.
3. Стоп-слова. Удаление стоп-слов не только общепринятых, но и относящихся к предметной области, позволяет упростить текст, сосредоточив внимание на существенных и информативных частях.
4. Короткие тексты. Было решено оставлять только те тексты, которые содержат не менее четырех слов.

В процессе фильтрации текстовых данных количество обращений сократилось с 6399 до 5849. Одновременно сократилось и количество токенов (слов) в корпусе с 692103 до 90199.

Биграммы, триграммы

Биграммы и триграммы представляют собой последовательности из двух или трех слов, которые часто встречаются в текстах и тесно связаны друг с другом. Например, такие выражения как «материнский капитал» или «почта банк» являются примерами биграмм и триграмм. Подобные комбинации слов могут содержать важную семантическую информацию и отражать тематические связи между словами.

Тематическое моделирование на основе библиотеки Gensim

Метод LDA является статистическим методом тематического моделирования текстов. Он позволяет выявлять скрытые темы в наборе текстов. Однако одним из ограничений LDA является необходимость предварительного задания количества тем, на которые будет разделено множество текстов. В реальности сложно точно определить, сколько тем существует в текстах. Поэтому в данной работе предлагается методика выбора оптимального числа тем из заданного диапазона. Для этого проводится анализ нескольких вариантов LDA-моделей с различным количеством тем в заданном диапазоне. Для каждой из этих моделей вычисляются две важные метрики: согласованность, перплексия.

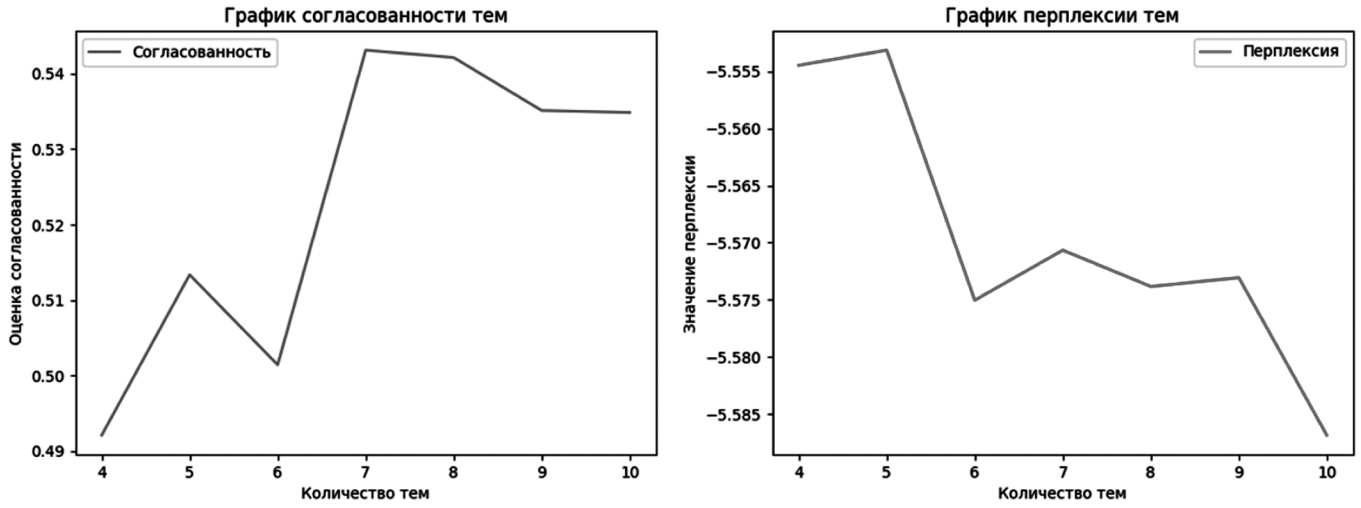

Рис. 3. Графики оценок LDA-моделей

Согласованность — это метрика, которая оценивает, насколько слова в одной и той же теме связаны между собой. Более согласованные темы имеют более логичный и понятный набор ключевых слов.

Перплексия представляет собой меру того, насколько хорошо модель предсказывает наблюдаемые данные. Более низкое значение перплексии свидетельствует о более качественной модели.

Выбор оптимальной модели LDA осуществляется на основе взвешенной оценки этих двух метрик. В данном случае, взвешенная оценка согласованности и перплексии вычисляется с коэффициентами, где согласованность имеет более высокий вес, чем перплексия. Такой подход помогает выбрать модель, которая обеспечивает наилучший баланс между пониманием тем и качеством прогнозирования. Для данного датасета оптимальным выбором является использование модели с 7-ю темами, как показано на графиках оценок LDA-моделей (рис. 3).

Анализ графиков показывает, согласованность имеет максимальное значение для модели, имеющей семь тем. А перплексия имеет минимальное значение для модели с 10-ю темами. Поскольку используется взвешенная оценка с коэффициентом для оценки согласованности 0,95, а для перплексии — 0,05, то оптимальное количество тем равно семи. Таким образом, оптимальная LDA-модель сформировала семь кластеров, как показано на рисунке 4.

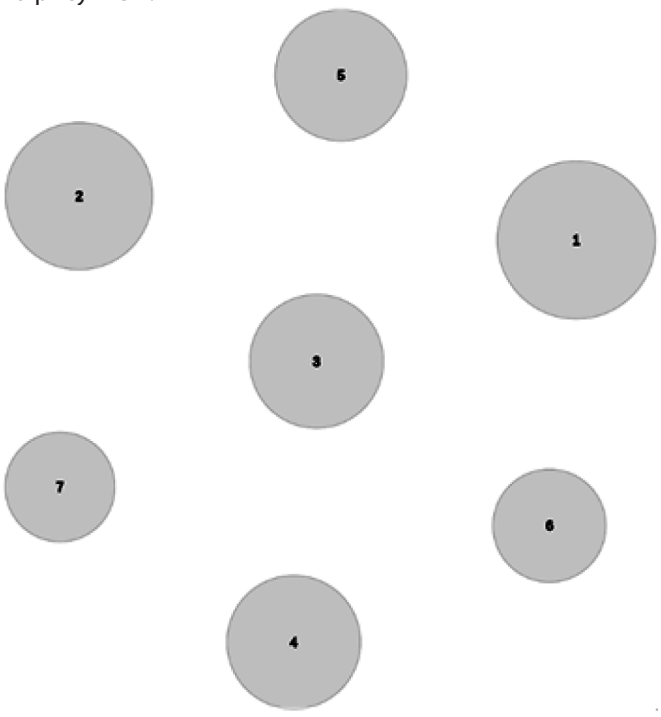

Рис. 4. Кластеры оптимальной LDA-модели

После выбора оптимальной LDA-модели, формируются темы (кластеры), каждая из которых связана с определенным набором ключевых слов. Эти ключевые слова представляют собой самые значимые и характерные термины для каждой темы. На основе этих ключевых слов можно проводить анализ содержания текстовых данных и понимать, о чем конкретно каждая тема. Например, кластер 3 содержит ключевые слова, представленные на рисунке 5.

Анализ ключевых слов кластера 3 позволяет определить то, что тексты этого кластера относятся к теме «Материнский капитал». Из всех возможных ключевых слов было выбрано первые десять ключевых слов, которые позволяют определить тему кластера, представляющие наибольшую вероятность принадлежности к данному кластеру.

Именование кластеров

Однако задача не ограничивается только выявлением ключевых слов и тем. Конечной целью является поиск названия темы. Для этого сначала была предпринята попытка использования гиперонимов, т.е. более общих понятий для ключевых слов, чтобы найти общий термин, описывающий тему [1]. Однако такой подход оказался неэффективным, так как гиперонимы для ключевых слов часто сильно различались, и не удавалось найти общий и однозначный термин.

В дальнейшем для анализа ключевых слов и именования тем было решено использовать искусственный интеллект ChatGPT, который показал хорошие результаты, но не всегда стабильные. ChatGPT часто предоставляет правильные и информативные названия для тем, основанные на ключевых словах. Результаты вычислительного эксперимента, которые позволили выделить семь кластеров и дать им имена, представлены в таблице 2.

Таким образом, совмещение ключевых слов и анализа ChatGPT позволяет более глубоко и точно интерпретировать темы, выявленные в данных, что важно для дальнейшего анализа и принятия решений на основе результатов тематического моделирования.

Заключение

Результаты исследований показывают, что тематическое моделирование можно использовать в качестве инструмента для анализа и классификации разговоров граждан по интересующим вопросам при их обращении в социальный фонд России. После транскрибации и записи этих разговоров в базу данных тематическое моделирование помогает не только выявить новые темы обращений или отнести их к уже имеющимся, но и определить ключевые темы и тем самым упростить анализ данных. Более того результаты моделирования позволяют классифицировать обращения граждан посредством

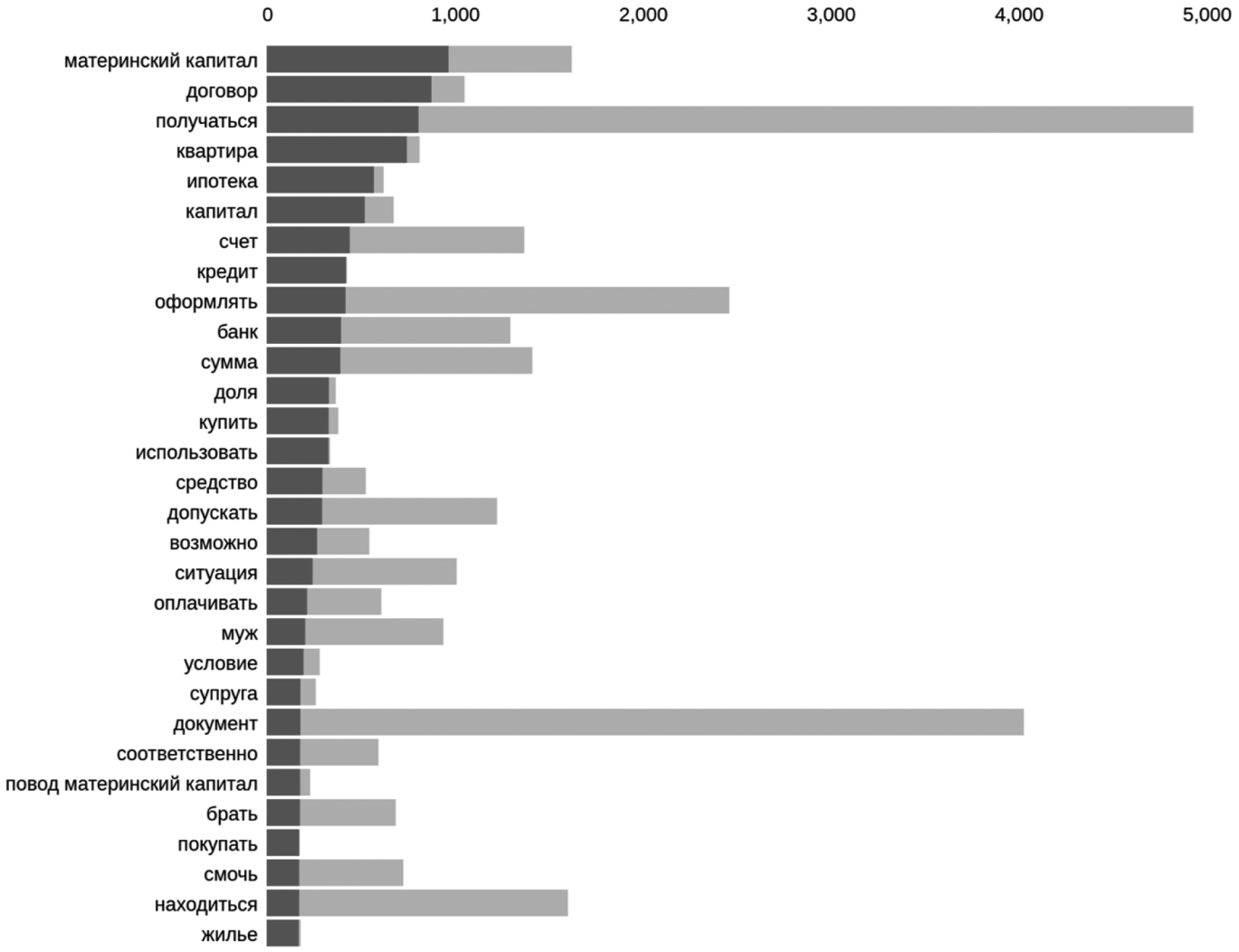

Рис. 5. Ключевые слова кластера 3

Таблица 2.

Результаты именования кластеров методом LDA

Номер кластера	Процент от общего числа тем	Количество текстов	Ключевые слова	Название
1	18,24	1067	пенсия, получать, получаться, выплата, пенсионер, рубль, сумма, оформлять, социальный, доплата	Социальные доплаты
2	8,24	482	стаж, входить, пенсия, получаться, справка, работодатель, общий, страховой, трудовой книжка, организация	Стаж и трудовая деятельность
3	8,39	491	материнский капитал, договор, получаться, квартира, ипотека, капитал, счет, кредит, оформлять, банк	Материнский капитал и ипотека
4	16,96	992	пенсия, получать, переводить, карта, сбербанк, карточка, банк, почта банк, перечислять, почта	Пенсионные выплаты
5	13,92	814	получать, документ, справка, паспорт, оформлять, прописка, прописывать, регистрация, материнский капитал, предоставлять	Оформление документов
6	14,89	871	пенсия, получать, получаться, инвалидность, информация, выходить, пенсионер, уходить, отвечать, возраст	Информация о пенсии
7	19,35	1132	заявление, сайт, подавать, документ, написать, госуслуга, личный кабинет, находить, информация, данные	Подача заявлений и получение информации

именования получаемых кластеров. Это позволяет аналитикам более эффективно и систематически анализировать данные обращений граждан за определенный период времени.

ЛИТЕРАТУРА

1. Tammishetti Vishnu, Konda Himakireeti. Automated Text Clustering and Labeling using Hypernyms. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 2 (2019) pp. 447–451. [электронный ресурс]. — Режим доступа: — https://www.ripublication.com/ijaer19/ijaerv14n2_16.pdf
2. David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. // Journal of Machine Learning Research. 2003. Vol. 3. pp. 993–1022 [электронный ресурс]. — Режим доступа: — https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
3. Pooja Kherwa, Poonam Bansal. Topic Modeling: A Comprehensive Review [электронный ресурс]. — Режим доступа: — https://www.researchgate.net/publication/334667298_Topic_Modeling_A_Comprehensive_Review
4. Named Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, Liang Zhao Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. — 2019. Vol.78. Pp 15169–15211. [электронный ресурс]. — Режим доступа: — https://www.semanticscholar.org/reader/b22de434b462558a127f327f29e2b0c673c0d7ab

Журнал «Современная наука: актуальные проблемы теории и практики»