DOI: 10.15514/ISPRAS-2026-38(3)-54

Исследование кластеризации эмбеддингов для поиска парафраз в текстах инструкций по медицинскому применению лекарственных средств

Н.В. Кильмишкин, ORCID: 0009-0007-4732-3260 <kilmiskinn@gmail.com>
Д.Д. Кубраков, ORCID: 0000-0003-2986-9343 <kubrakoff.dmitry@yandex.ru>
Ю.П. Титов, ORCID: 0000-0002-9093-6755 <kalengul@mail.ru>
В.И. Пантелеев, ORCID: 0000-0002-1575-1267 <dr.vpantel@gmail.com>
Т.А. Куропаткина, ORCID: 0000-0003-1027-8703 <0sylphide0@gmail.com>
Н.А. Кочина, ORCID: 0000-0001-7748-0071 <natalyakochina@yandex.ru>
П.М. Иванова, ORCID: 0009-0005-4579-4603 <polianna_654@mail.ru>

ФГБОУ ВО «Российский экономический университет имени Г.В. Плеханова», Россия, 115054, г. Москва, Стремянный пер., д. 36.

Аннотация. В данной работе рассматривается комплексный подход для выявления парафраз в текстах медицинских инструкций, объединяющий современные методы обработки естественного языка (NLP), снижения размерности и кластерного анализа эмебеддингов названий вершин семантического графа, который строится для решения задачи поиска взаимодействий лекарственных средств при полифармакотерапии. Наилучшие результаты продемонстрировала комбинация multilingual-модели distiluse_base_multilingual с алгоритмом UMAP и агломеративной кластеризацией. Особенностью методики стало применение стратегии уменьшения размерности с последующим добавлением информации о классе, что позволило сохранить семантические взаимосвязи и улучшить качество кластеризации. Проведенный сравнительный анализ различных языковых моделей (включая Clinical Modern BERT, paraphrase-multilingual и rubert-tiny) выявил преимущества модели distiluse_base_multilingual по показателям точности и вычислительной эффективности. Визуализация результатов подтвердила способность метода к четкому выделению смысловых кластеров, а использование JSON-формата для хранения результатов обеспечило их удобную интеграцию в практические приложения. Разработанный метод позволяет автоматизировать обработку медицинских текстов для унификации терминологии в инструкциях к лекарствам.

Ключевые слова: машинное обучение; обработка естественного языка NLP; кластеризация; инструкции к лекарственным средствам; снижение размерности признакового пространства; поиск парафраз.

Для цитирования: Кильмишкин Н.В., Кубраков Д.Д., Титов Ю.П., Пантелеев В.И., Куропаткина Т.А., Кочина Н.А., Иванова П.М. Исследование кластеризации эмбеддингов для поиска парафраз в текстах инструкций по медицинскому применению лекарственных средств. Труды ИСП РАН, том 38, вып. 3, часть 4, 2026 г., стр. 175–190. DOI: 10.15514/ISPRAS–2026–38(3)–54.

Благодарности: Исследование выполнено при финансовой поддержке Российского научного фонда, проект № 23-75-30012.

A Research on embedding clustering for paraphrase retrieval in texts instructions for medical use of drugs

N.V. Kilmishkin ORCID: 0009-0007-4732-3260 *<kilmiskinn@gmail.com>*
D.D. Kubrakov ORCID: 0000-0003-2986-9343 *<kubrakoff.dmitry@yandex.ru>*
Y.P. Titov, ORCID: 0000-0002-9093-6755 <kalengul@mail.ru>
B.И. Panteleev, ORCID: 0000-0002-1575-1267 <dr.vpantel@gmail.com>
T.A. Kuropatkina, ORCID: 0000-0003-1027-8703 <0sylphide0@gmail.com>
N.A. Kochina, ORCID: 0000-0001-7748-0071 <natalyakochina@yandex.ru>
P.M. Ivanova, ORCID: 0009-0005-4579-4603 <polianna_654@mail.ru>

Plekhanov Russian University of Economics,
36, Stremyanny Lane, Moscow, 115054, Russia.

Abstract. The research developed an integrated approach for paraphrase detection in medical instruction texts, combining modern methods of natural language processing (NLP), dimensionality reduction and cluster analysis. The best results were demonstrated by the combination of the distiluse_base_multilingual model with the UMAP algorithm (parameters: n_components=2, n_neighbors=10, min_dist=0.1, metric=cosine) and agglomerative clustering (n_clusters=200, linkage=ward). A special feature of the methodology was the use of a dimensionality reduction strategy followed by the addition of class information, which preserved semantic relationships and improved the quality of clustering.A comparative analysis of different language models (including Clinical Modern BERT, paraphrase-multilingual and rubert-tiny) revealed the advantages of the distiluse_base_multilingual model in terms of accuracy and computational efficiency. Visualisation of the results confirmed the ability of the method to clearly distinguish semantic clusters, and the use of JSON-format for storing the results ensured their convenient integration into practical applications. The developed approach has the potential to automate the processing of medical texts, especially in the tasks of unifying the terminology of drug instructions.

Keywords: machine learning; NLP natural language processing; clustering; drug instructions; feature space dimensionality reduction; paraphrase search.

For citation: Kilmishkin N.V., Kubrakov D.D., Titov Y.P., Panteleev V.I., Kuropatkina T.A., Kochna N.A., Ivanova P.M. A Research on embedding clustering for paraphrase retrieval in texts instructions for medical use of drugs. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 3, part 4, 2026, pp. 175-190 (in Russian). DOI: 10.15514/ISPRAS-2026-38(3)-54.

Acknowledgements. The study was carried out with the financial support of the Russian Science Foundation, project no. 23-75-30012.

1. *Введение*

В современной обработке естественного языка (NLP) поиск близости слов является ключевой задачей, которая находит применение в машинном переводе, информационном поиске, классификации текстов и других областях. Основные методы определения семантической и синтаксической близости слов включают косинусное сходство, евклидово расстояние, нейросетевые подходы и вероятностные модели [1-2]. Косинусное сходство, измеряющее угол между векторами слов в многомерном пространстве, широко применяется благодаря своей интерпретируемости и устойчивости к различиям в длине векторов. Евклидово расстояние, хотя и менее устойчивое к высокой размерности, остается базовым методом оценки геометрической близости векторов [3]. Более сложные подходы, такие как Word2Vec (включая архитектуры Skip-Gram и CBOW), GloVe и FastText, используют нейронные сети для построения плотных векторных представлений, учитывающих контекстное употребление слов [4-5] Современные контекстуализированные модели, включая BERT (Bidirectional Encoder Representations from Transformers), ELMo и GPT, генерируют динамические числовые векторные представления слов (эмбеддинги), которые адаптируются

к конкретному окружению слова в тексте, что значительно повышает точность определения семантической близости. Вероятностные модели, такие как латентное размещение Дирихле (LDA) и латентный семантический анализ (LSA), опираются на статистические закономерности в корпусах текстов для выявления скрытых тематических структур и оценки близости слов на основе их распределений [3, 6].

Токенизация, как начальный этап обработки текста, играет критическую роль в подготовке данных для последующего анализа. Простейшая словная токенизация, основанная на разделении текста по пробелам и знакам пунктуации, уступает более продвинутым методам в языках со сложной морфологией. Побайтовая токенизация (Byte Pair Encoding, BPE) и её модификация WordPiece, используемые в моделях типа BERT и GPT, эффективно обрабатывают редкие и составные слова за счет разбиения на часто встречающиеся последовательности символов. Для агглютинативных и морфологически богатых языков, таких как русский или турецкий, применяется морфемная токенизация, которая разбивает слова на минимальные значимые единицы. В случаях, когда требуется максимально детализированное представление текста, используется символьная токенизация, хотя она существенно увеличивает вычислительную нагрузку из-за роста длины последовательностей.

Методы получения эмбеддингов эволюционировали от статических представлений к динамическим, учитывающим контекст. Статические эмбеддинги, такие как Word2Vec и GloVe, обучаются на больших корпусах текстов и фиксируют обобщенные семантические отношения между словами. Word2Vec, основанный на предсказании слов в окрестности (Skip-Gram) или предсказании целевого слова по контексту (CBOW), эффективно улавливает синтаксические и семантические закономерности. GloVe комбинирует преимущества глобальной статистики совместной встречаемости слов с локальными контекстными методами, что позволяет более точно отражать семантические связи [7]. FastText расширяет этот подход, учитывая морфемную структуру слов через представление их в виде суммы n-грамм символов, что особенно полезно для языков с богатой словообразовательной системой. Динамические эмбеддинги, генерируемые моделями типа BERT, ELMo и GPT, используют механизмы внимания (Transformer) для создания контекстуализированных представлений, которые варьируются в зависимости от окружения слова в предложении [8-10] BERT, обучаемый на задачах маскированного языка и предсказания следующего предложения, демонстрирует высокую эффективность в различных NLP-задачах благодаря своей способности учитывать глубокие контекстные зависимости. ELMo, основанный на двунаправленных LSTM, также обеспечивает качественное контекстуальное представление слов, в то время как GPT использует авторегрессионные трансформеры для генерации эмбеддингов [11]. Для задач, требующих представления целых предложений или документов, разработаны специализированные методы, такие как Sentence-BERT и Doc2Vec [12]. Sentence-BERT применяет сиамские сети на основе BERT для эффективного сравнения предложений, а Doc2Vec расширяет подход Word2Vec, включая идентификатор документа в модель. Выбор конкретного метода токенизации и генерации эмбеддингов зависит от специфики решаемой задачи, характеристик языка и требуемого уровня детализации. Современные подходы сочетают в себе передовые нейросетевые архитектуры с эффективными алгоритмами предобработки, что позволяет достигать высокой точности в задачах определения семантической близости и других приложениях NLP.

2. *Постановка задачи*

Актуальность исследования обусловлена необходимостью разработки методов анализа взаимодействий лекарственных средств (ЛС) в условиях полифармакотерапии, которая широко применяется при лечении хронических заболеваний. Одной из ключевых задач в этой области является выявление потенциально опасных комбинаций ЛС, что требует глубокого анализа их фармакологических свойств и семантических связей. Для решения этой задачи

используются графовые модели данных, которые позволяют наглядно представить сложные взаимосвязи между сущностями, описанными в инструкциях по медицинскому применению ЛС [13].

В рамках данного исследования графы строятся на основе анализа инструкций каждого ЛС, где узлы соответствуют фармакологическим понятиям (например, действующим веществам, показаниям, побочным эффектам), а рёбра отражают семантические связи между ними. На последующем этапе индивидуальные графы объединяются в единую семантическую сеть, которая служит основой для выявления взаимодействий между различными ЛС.

Однако при построении таких графов возникает проблема дублирования и вариативности представления семантически эквивалентных сущностей. Это приводит к избыточному росту количества уникальных вершин из-за наличия множества лексических вариантов обозначения одних и тех же понятий (например, синонимов, аббревиатур или различных форм записи). Такая избыточность снижает информативность и практическую ценность графовой модели, затрудняя анализ взаимодействий ЛС.

Целью данной работы является разработка метода устранения указанной проблемы путём кластеризации и нормализации семантически эквивалентных сущностей в графовой модели. Это позволит находить больше новых общих вершин при слиянии графов (рис. 1), что в свою очередь может повысить точность и эффективность анализа взаимодействий ЛС в условиях полифармакотерапии, что имеет важное значение для обеспечения безопасности пациентов и оптимизации схем лечения.

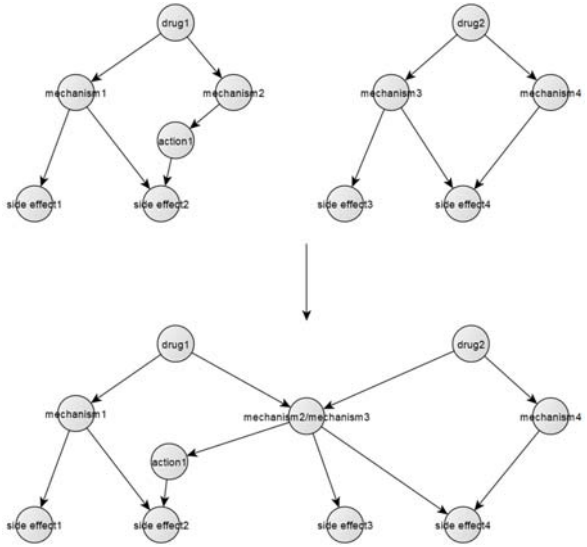

Рис. 1. Выявление парафраз с целью лучшего слияния графов.
Fig. 1. Identifying paraphrases to better merge graphs.

Процедура выявления парафраз включает четыре последовательных этапа. Первый этап предполагает сбор и предварительную обработку текстового корпуса инструкций по медицинскому применению ЛС из надежных источников, включая все значимые описания сущностей, относящихся к механизмам действия, взаимодействию с белками, фармакокинетике и другим аспектам. На втором этапе разрабатывается метод генерации векторных представлений (эмбеддингов) для текстовых описаний, основанный на сравнительном анализе современных языковых моделей, таких как BERT и его

специализированные медицинские модификации, с последующей их дообучением на собранном корпусе. Третий этап заключается в создании алгоритма кластеризации, который группирует семантически близкие сущности на основе их векторных представлений, обеспечивая баланс между объединением дублирующихся формулировок и сохранением смысловых различий между принципиально разными понятиями. Реализация данного подхода оптимизирует интеграцию данных о ЛС из различных источников за счет автоматического обнаружения и слияния схожих сущностей, что повышает эффективность работы с объединенной базой данных, в частности при анализе лекарственных взаимодействий. На заключительном этапе полученный датасет проходит экспертную оценку медицинских специалистов для проверки корректности и практической значимости результатов.

3. *Модели и методы*

3.1 Сбор текстового корпуса

Процесс создания репрезентативного текстового корпуса для последующего семантического анализа включает несколько методологически важных этапов. Исходные данные извлекаются из двух основных источников: официальных PDF-документов инструкций по медицинскому применению ЛС государственного реестра лекарственных средств (ГРЛС), утвержденных регулирующими органами, и структурированных данных с авторитетного фармацевтического портала [14]. Такой комбинированный подход позволяет обеспечить необходимый уровень полноты и достоверности собираемых данных. Для извлечения текста из ГРЛС используется парсинг PDF-документов с применением специализированной библиотек для языка программирования Python Fitz.

Ключевым этапом является селекция релевантных разделов инструкций. В рамках данного исследования основное внимание было уделено трём ключевым разделам:

- Фармакодинамика – содержит описание механизмов действия, рецепторной активности, биохимических эффектов;
- Фармакокинетика – включает данные о всасывании, распределении, метаболизме и выведении;
- Побочные эффекты – описывает нежелательные лекарственные реакции и их частоту.

Текстовый корпус состоит из 8359 предложений.

3.2 Получение эмбеддингов

После формирования репрезентативного текстового корпуса осуществляется ключевой этап адаптации предобученной языковой модели к предметной области фармакологии с помощью. Данный процесс направлен на оптимизацию способности модели генерировать семантически значимые векторные представления для узкоспециализированных медицинских терминов и фармакологических понятий. В качестве базовой архитектуры рассматриваются современные трансформерные модели типа BERT и её специализированные модификации, которые демонстрируют высокую эффективность в задачах обработки естественного языка в медицинской области.

Эмбеддинги – это числовые векторные представления слов, предложений или целых текстов, которые сохраняют их семантические и синтаксические свойства [15]. Они позволяют переводить естественный язык в форму, понятную алгоритмам машинного обучения, и используются для решения множества задач, включая кластеризацию текстовых данных. В

случае кластеризации словосочетаний на русском языке эмбеддинги помогают группировать схожие по смыслу фразы, даже если они выражены разными словами.

Для получения эмбеддингов часто используются предобученные языковые модели на основе BERT, которые способны учитывать контекст слов в предложении и создавать более точные векторные представления.

В данной работе рассматриваются несколько популярных моделей, подходящих для работы с русским языком:

- all-MiniLM-L6-v2 – это компактная модель на основе архитектуры BERT, оптимизированная для эффективного создания эмбеддингов. Она сохраняет высокое качество семантического представления текста, несмотря на уменьшенный размер, и хорошо подходит для задач кластеризации.
- Clinical Modern BERT – специализированная модель, дообученная на медицинских текстах. Если словосочетания относятся к медицинской тематике, эта модель обеспечит более релевантные эмбеддинги за счет учета терминологии и контекста.
- distiluse-base-multilingual – облегченная multilingual-модель, основанная на архитектуре DistilBERT. Она поддерживает несколько языков, включая русский, и эффективна для задач, где требуется обработка текстов на разных языках.
- paraphrase-multilingual – модель, оптимизированная для поиска схожих по смыслу предложений (парафраз). Она хорошо подходит для кластеризации, так как умеет выделять семантически близкие фразы даже при различии в формулировках.
- rubert-tiny и rubert-tiny-2 – это уменьшенные версии BERT для русского языка. Они быстрее работают и требуют меньше вычислительных ресурсов, но при этом сохраняют способность к качественному представлению текста.

Кластеризация текстовых данных требует, чтобы схожие по смыслу фразы находились близко друг к другу в векторном пространстве. Эмбеддинги, полученные с помощью BERT-моделей, кодируют семантику текста, поэтому после их применения алгоритмы кластеризации могут группировать фразы по смыслу, они обеспечивают необходимый уровень семантической чувствительности для выявления и объединения схожих фармакологических сущностей.

Общее количество эмбеддингов для кластеризации: 1050 текстовых строк.

3.3 Описание метода кластеризации эмбеддингов

В данной работе исследуются подходы к улучшению качества кластеризации векторных представлений текстовых данных (эмбеддингов), полученных с использованием современных языковых моделей, таких как BERT и его производные. Основное внимание уделяется двум ключевым аспектам: предварительному снижению размерности эмбеддингов и интеграции дополнительных категориальных признаков в виде индикаторного векторного представления (one-hot encoding) кодирования для учета семантических категорий [16]. Рассматриваются также вариации порядка комбинирования числовых и категориальных признаков с целью оптимизации кластеризации [17-18].

Эмбеддинги, полученные с помощью современных языковых моделей, обладают высокой размерностью (обычно от 384 до 1024 компонент), что может негативно влиять на эффективность алгоритмов кластеризации из-за «проклятия размерности» (curse of dimensionality). Высокомерные данные часто приводят к разреженности пространства признаков, затрудняя выделение компактных кластеров. Для решения этой проблемы применяются методы снижения размерности, среди которых наиболее распространены:

- Метод главных компонент (PCA, Principal Component Analysis) – линейный метод, проецирующий данные на ортогональные направления максимальной дисперсии. PCA позволяет сократить размерность, сохраняя основную информацию, но может терять нелинейные зависимости [19].
- t-SNE (t-Distributed Stochastic Neighbor Embedding) – нелинейный метод, ориентированный на визуализацию, который минимизирует расхождения между распределениями расстояний в исходном и редуцированном пространствах. Однако t-SNE чувствителен к гиперпараметрам и может искажать глобальную структуру данных [20].
- UMAP (Uniform Manifold Approximation and Projection) – современный нелинейный метод, сочетающий преимущества t-SNE в сохранении локальных структур с более устойчивым представлением глобальных зависимостей. UMAP часто демонстрирует лучшую производительность при кластеризации [21].

Применение этих методов перед кластеризацией позволяет уменьшить вычислительную сложность и улучшить качество группировки за счет устранения шумовых компонент и избыточности в данных. Пример использования алгоритма t-SNE для уменьшения размерности эмбеддингов с последующей визуализацией показан на рис. 2.

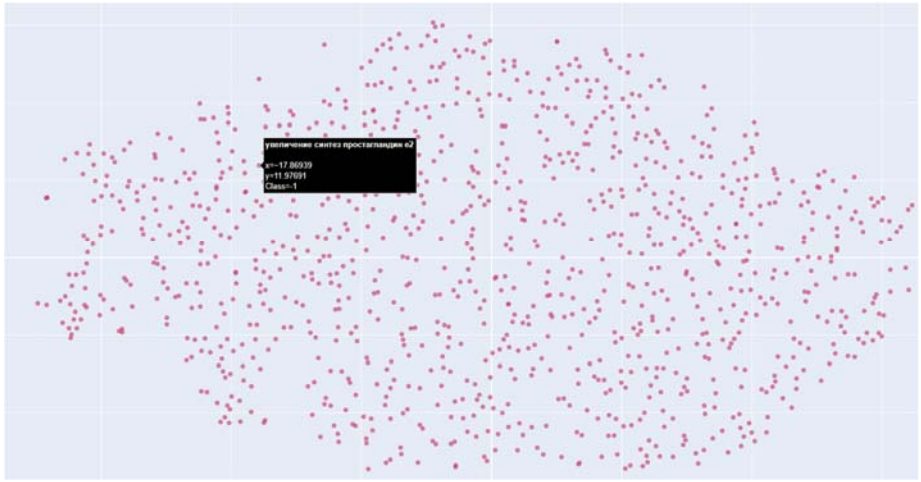

Рис. 2. Визуализация снижения размерности эмбеддигов перед кластеризацией.
Fig. 2. Visualization of embeddig dimension reduction before clustering.

Для повышения информативности эмбеддингов в задаче кластеризации словосочетаний предлагается дополнять их категориальными признаками, закодированными в виде векторов индикаторного представления. В рассматриваемом контексте категории соответствуют семантическим группам, таким как механизм, действие, метаболизм, всасывание, выведение, связь с белками. One-hot кодирование преобразует каждую категорию в вектор, где единица соответствует принадлежности к определенному классу, а остальные элементы равны нулю. Пример использования индикаторного векторного представления показан на рис. 3.

Добавление индикаторного векторного представления признаков решает две задачи:

- Учет предметной области – категории позволяют явно разделять словосочетания по тематикам, что особенно важно в узкоспециализированных доменах (например, в медицине или фармакологии).

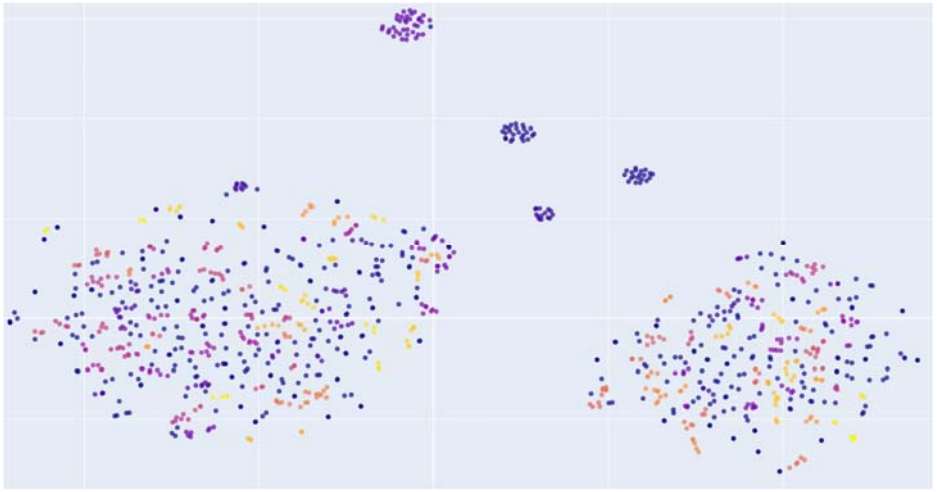

Рис. 3. Использование индикаторного векторного представления кодирования.
Fig. 3. Using one-hot encoding.

- Улучшение интерпретируемости кластеров – интеграция категорий помогает объяснить, по каким признакам образовались группы, что критично для анализа результатов.

Важно учесть, что использование значений, значительно превышающих 1, в индикаторном векторном представлении может привести к нарушению масштабирования признаков, что негативно скажется на работе алгоритмов машинного обучения, особенно чувствительных к дисбалансу весов, таких как линейные модели и методы, основанные на расстояниях. Чрезмерно большие значения искажают процесс регуляризации, искусственно подавляя вклад отдельных признаков, а также усложняют сходимость градиентных методов из-за нестабильности градиентов. Необходимо настраивать значимость категорий через взвешивание классов, чтобы сохранить интерпретируемость.

Важным аспектом является последовательность объединения эмбеддингов и индикаторных векторов. Рассматриваются два варианта:

- Конкатенация перед снижением размерности – индикаторные признаки добавляются к исходным эмбеддингам, после чего применяется PCA, UMAP или другой метод. Такой подход позволяет совместно учитывать числовые и категориальные признаки при редукции, но может приводить к доминированию категорий из-за их дискретного характера.
- Конкатенация после снижения размерности – сначала эмбеддинги редуцируются, а затем комбинируются с векторами индикаторного представления. Это сохраняет структуру числовых данных, но требует тщательного подбора весовых коэффициентов для балансировки вклада признаков.

Экспериментальная оценка обоих подходов позволяет определить оптимальную стратегию для конкретной задачи. Описание стратегий показано на рис. 4.

В рамках данного исследования рассматриваются два алгоритма кластеризации, применяемые для группировки векторных представлений текстовых данных (эмбеддингов):

- K-means [22];
- Агломеративная кластеризация [23].

Рис. 4. Описание предобработки эмбеддигов перед кластеризацией.
Fig. 4. Description of the preprocessing of the seed by forward clustering.

K-means представляет собой наиболее распространенный алгоритм кластеризации, особенно эффективный при работе с эмбеддингами после предварительного снижения размерности. Его главное преимущество заключается в исключительной вычислительной эффективности – алгоритм демонстрирует линейную сложность $O(n)$, что позволяет обрабатывать значительные объемы данных. Важным практическим преимуществом является предсказуемость результатов и простота интерпретации – каждый кластер представлен четким центром, что значительно упрощает анализ.

Агломеративная кластеризация предлагает принципиально иной, иерархический подход к анализу эмбеддингов. Метод особенно ценен при необходимости многоуровневого анализа данных, позволяя исследовать структуры на различных уровнях детализации. Основное преимущество – возможность использования семантически значимых метрик расстояния, в частности косинусной меры, наиболее адекватно отражающей смысловую близость текстовых представлений. Процесс построения дендрограммы предоставляет уникальные возможности для визуального анализа и интерпретации взаимосвязей между кластерами. Хотя алгоритм имеет более высокую вычислительную сложность $O(n^3)$, применение после редукции размерности делает его практичным для большинства реальных задач.

4. *Результаты кластеризации*

Для выбора оптимального метода кластеризации и снижения размерности эмбеддингов была разработана программная реализация на языке Python, которая осуществляла автоматизированное тестирование различных комбинаций алгоритмов. Каждая комбинация алгоритмов оценивалась с использованием двух ключевых метрик: коэффициента силуэта (silhouette score) и индекса Дэвиса-Болдина (Davies-Bouldin index).

Коэффициент силуэта (silhouette score) является метрикой, позволяющей оценить качество кластеризации на основе компактности и разделимости кластеров. Значение коэффициента варьируется в диапазоне от -1 до 1, где значения, близкие к 1, указывают на четкую структуру кластеров, значения около 0 свидетельствуют о пересекающихся кластерах, а отрицательные значения говорят о некорректной кластеризации. Формально коэффициент силуэта для отдельного объекта вычисляется как:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

где a(i) – среднее расстояние от объекта i до других объектов в том же кластере, а b(i) – среднее расстояние от объекта i до объектов ближайшего соседнего кластера.

Индекс Дэвиса-Болдина (Davies-Bouldin index) – это метрика, оценивающая соотношение внутрикластерной дисперсии и межкластерного расстояния. Чем ниже значение индекса, тем

лучше качество кластеризации. Индекс вычисляется как среднее значение для всех кластеров соотношения их внутренней компактности (например, среднего расстояния между объектами внутри кластера) и разделимости (например, расстояния между центрами кластеров). Формула для вычисления индекса Дэвиса-Болдина имеет вид:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{d(c_i, c_j)} \right) \quad (2)$$

где k – количество кластеров, S_i – среднее расстояние между объектами кластера i и его центром, а $d(c_i,c_j)$ – расстояние между центрами кластеров i и j.

Наилучшие результаты, полученные в ходе экспериментального исследования, представлены на рис. 5-10.

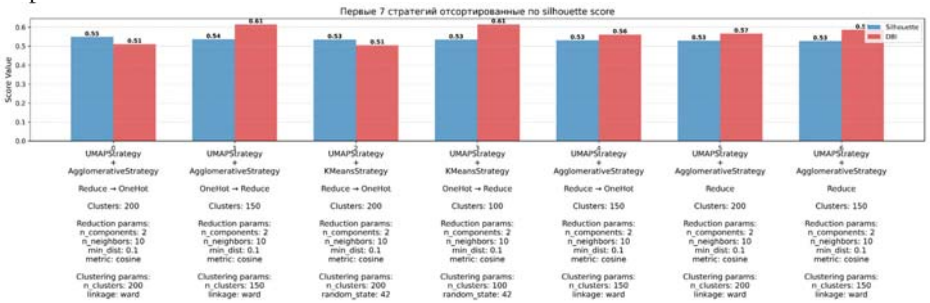

Рис. 5. Гистограмма первых семи лучших результатов на основе модели all_MiniLM_L6_v2.
Fig. 5. Histogram of the first seven best results based on the all_Minilim_L6_v2 model.

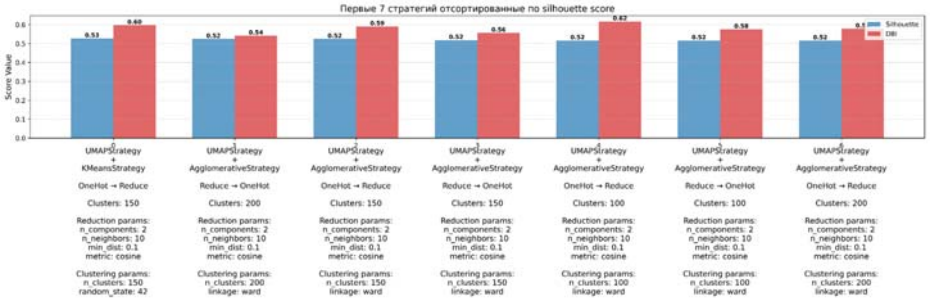

Рис. 6. Гистограмма первых семи лучших результатов на основе модели Clinical_ModernBERT.
Fig. 6. Histogram of the first seven best results based on the Clinical_ModernBERT.

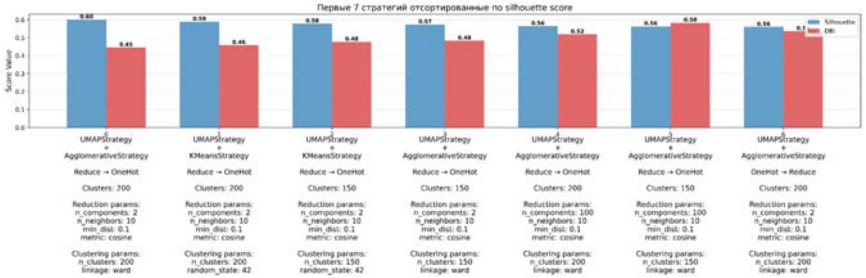

Рис. 7. Гистограмма первых семи лучших результатов на основе модели distiluse_base_multilingual.
Fig. 7. Histogram of the first seven best results based on the distiluse_base_multilingual.

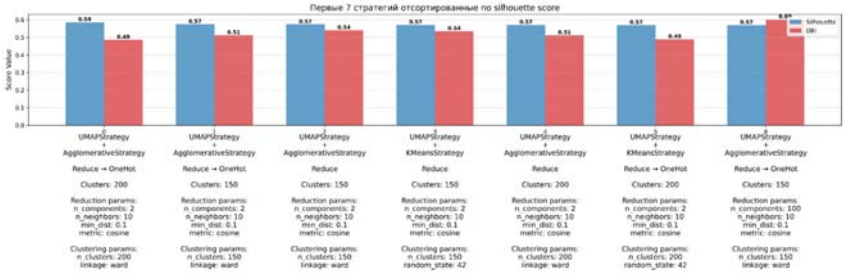

Рис. 8. Гистограмма первых семи лучших результатов на основе модели paraphrase_multilingual.
Fig. 8. Histogram of the first seven best results based on the paraphrase_multilingual.

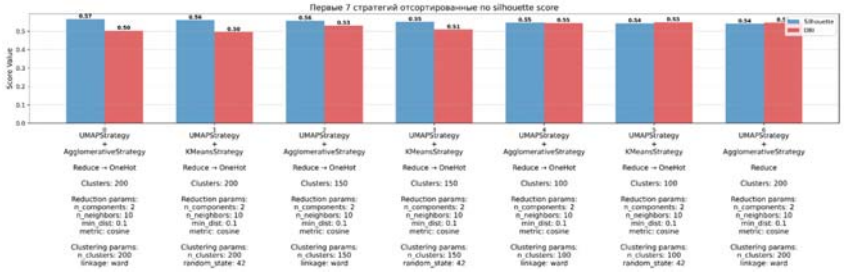

Рис. 9. Гистограмма первых семи лучших результатов на основе модели rubert-tiny.
Fig. 9. Histogram of the first seven best results based on the rubert-tiny.

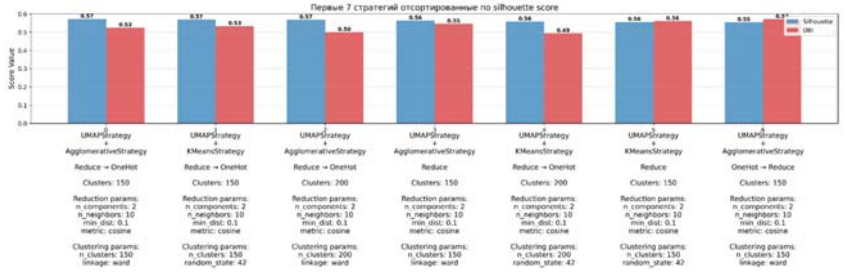

Рис. 10. Гистограмма первых семи лучших результатов на основе модели rubert-tiny2.
Fig. 10. Histogram of the first seven best results based on the rubert-tiny2.

В ходе исследования был достигнут наилучший результат кластеризации и уменьшения размерности данных при использовании multilingual-модели distiluse_base_multilingual в сочетании с агломеративной кластеризацией и алгоритмом уменьшения размерности UMAP с применением стратегии Reduce – One-hot (уменьшение размерности эмбеддинга, добавление информативного вектора с последующей кластеризацией). Данный подход позволил эффективно снизить размерность пространства признаков с сохранением ключевых структур данных, что в дальнейшем обеспечило высокое качество кластеризации.

Для алгоритма UMAP были выбраны следующие гиперпараметры: количество компонент (n_components = 2), что позволило визуализировать данные в двумерном пространстве, число соседей (n_neighbors = 10), обеспечивающее баланс между локальной и глобальной структурой данных, минимальное расстояние между точками (min_dist = 0.1), способствующее оптимальному разделению кластеров, а также метрика косинусного

сходства (metric = cosine), учитывающая угловое расстояние между векторами и тем самым повышающая устойчивость к различиям в длине текстовых эмбеддингов.

На этапе кластеризации применялся агломеративный алгоритм с параметрами: количество кластеров (n_clusters = 200), что соответствовало предполагаемому уровню детализации категоризации, и метод связи Ward (linkage = ward), минимизирующий дисперсию внутри кластеров и обеспечивающий компактные и хорошо разделенные группы. Использование стратегии Reduce – One-hot позволило дополнительно обогатить данные за счет добавления информативного вектора, что улучшило разделимость кластеров.

Полученные результаты демонстрируют эффективность комбинации современных методов NLP, нелинейного уменьшения размерности и иерархической кластеризации для задач анализа многомерных текстовых данных. Применение UMAP в сочетании с агломеративной кластеризацией показало свою устойчивость к шуму и способность выявлять сложные зависимости в данных, что делает данный подход перспективным для обработки многоязычных текстовых корпусов.

На рис. 11 представлена визуализация результатов кластеризации, полученных с применением комбинации модели distiluse_base_multilingual, алгоритма уменьшения размерности UMAP и агломеративной кластеризации. Визуальное представление демонстрирует разделение данных на кластеры.

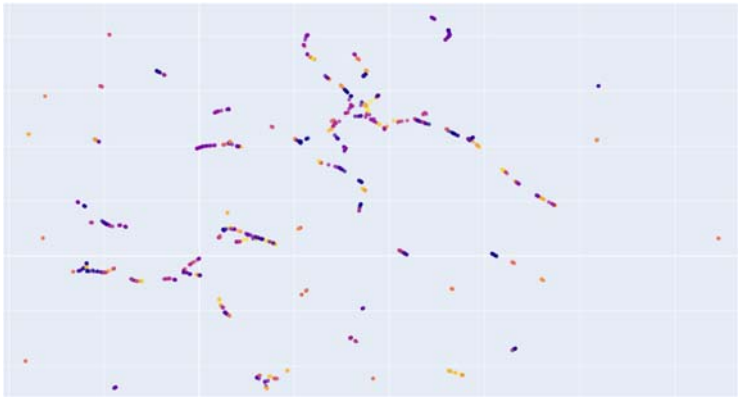

Рис. 11. Визуализация кластеризации эмбеддингов с помощью модели distiluse_base_multilingual с алгоритмом Агломеративной кластеризацией, алгоритмом уменьшения размерности UMAP со стратегией Reduce – One-hot.
Fig. 11. Visualization of embedding clustering using the distiluse_base_multilingual model with Agglomerative clustering algorithm, UMAP dimensionality reduction algorithm with Reduce – One-hot strategy.

Результаты поиска парафраз сохранены в формате JSON (рис. 12), что обеспечивает структурированное хранение данных и удобство их дальнейшего анализа. В файле содержатся сгруппированные текстовые фрагменты, объединенные на основе семантической близости.

5. *Заключение*

Проведенное исследование продемонстрировало эффективность комбинированного подхода, объединяющего методы обработки естественного языка (NLP), нелинейного снижения размерности и иерархической кластеризации, для решения задачи поиска парафраз в текстах инструкций по медицинскому применению ЛС.

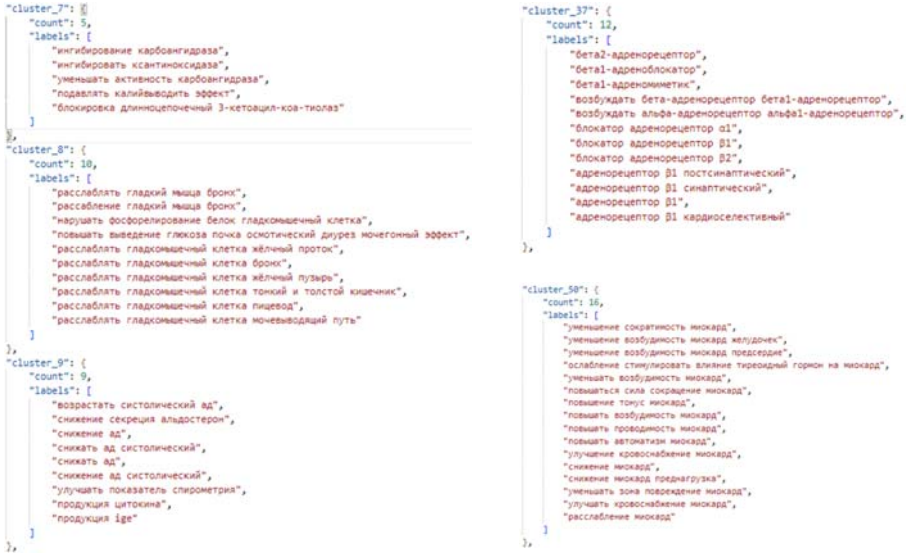

Рис. 12. Пример результата поиска парафраз с помощью модели distiluse_base_multilingual с алгоритмом Агломеративной кластеризацией, алгоритмом уменьшения размерности UMAP со стратегией Reduce - One-hot.
Fig. 12. An example of a paraphrase search result using the distiluse_base_multilingual model with the Agglomerative clustering algorithm, the UMAP dimensionality reduction algorithm with the Reduce - One-hot strategy.

Использование multilingual-модели distiluse_base_multilingual в сочетании с алгоритмом UMAP и агломеративной кластеризацией позволили достичь высокой точности группировки семантически близких текстовых фрагментов. Ключевым фактором успеха стало применение стратегии Reduce – One-hot, которая обеспечила сохранение структурной целостности данных после снижения размерности и улучшила разделимость кластеров за счет интеграции категориальных признаков.

Оптимальные результаты были получены при следующих параметрах:

- Для UMAP: метрика косинусного расстояния (metric = cosine), количество компонент (n_components = 2), число соседей (n_neighbors = 10) и минимальное расстояние (min_dist = 0.1). Эти настройки позволили сохранить локальные и глобальные зависимости в данных, что критически важно для последующей кластеризации.
- Для агломеративной кластеризации: метод связи Ward (linkage = ward) и количество кластеров (n_clusters = 200), что обеспечило минимизацию внутрикластерной дисперсии и формирование компактных, хорошо разделенных групп.

Визуализация результатов (рис. 11) подтвердила с достаточной точностью четкое разделение данных на кластеры, а сохранение результатов в формате JSON (рис. 12) обеспечило удобство их дальнейшего использования в прикладных задачах, таких как анализ взаимодействий лекарственных средств или автоматическое реферирование. Предложенный метод демонстрирует потенциал для автоматизации обработки медицинских текстов. Предложен универсальный pipeline для обработки медицинских текстов, сочетающий современные NLP-модели, методы снижения размерности и кластеризации. Этот подход может быть адаптирован для других узкоспециализированных доменов. Решение проблемы

дублирования и вариативности терминологии в инструкциях ЛС повышает точность интеграции данных, что важно для систем поддержки врачебных решений и анализа полифармакотерапии. Сравнительный анализ моделей (включая Clinical Modern BERT, paraphrase-multilingual и rubert-tiny и т.д.) показал, что distiluse_base_multilingual обеспечивает наилучший баланс между качеством кластеризации и вычислительной эффективностью. Данное исследование направлено на разработку методов анализа взаимодействий лекарственных средств при полифармакотерапии с использованием графовых моделей данных, позволяющих выявлять потенциально опасные комбинации ЛС путем семантического анализа инструкций и устранения лексической вариативности описаний. Исследование выполнено при финансовой поддержке Российского научного фонда (проект № 23-75-30012).

Список литературы / References

[1]. Лях А.П. Классификация и основные алгоритмы эмбендинга в контексте больших языковых моделей. Электронное научное издание «Ученые заметки ТОГУ», 2024, т. 15, № 3, стр. 79-83, ISSN 2079-8490.

[2]. Жаксыбаев Д.О., Мизамова Г.Н. Алгоритмы обработки естественного языка для понимания семантики текста. Труды ИСП РАН, 2022, том 34, вып. 1, стр. 141-150. DOI: 10.15514/ISPRAS-2022-34(1)-10. / Zhaxybayev D.O., Mizamova G.N. Natural Language Processing Algorithms for Understanding the Semantics of Text. Trudy ISP RAN/Proc. ISP RAS, 2022, vol. 34, issue 1, pp. 141-150 (in Russian). DOI: 10.15514/ISPRAS-2022-34(1)-10.

[3]. Лыченко Н.М., Сороковая А.В. Сравнение эффективности методов векторного представления слов для определения тональности текстов. Институт машиноведения и автоматики Национальной академии наук Кыргызской республики, Бишкек, Кыргызстан. Математические структуры и моделирование, 2019, № 4(52), стр. 97-110. DOI 10.24147/2222-8772.2019.4.97-110.

[4]. Нгуен Нгок Зиеп, Ле Мань Ха. Нейросетевой метод снятия омонимии. Московский физико-технический институт (государственный университет). Информатика, вычисл. техника и управление. ТРУДЫ МФТИ, 2015, т. 7, № 4.

[5]. Частикова В.А., Козачёк К.В., Гуляй В.Г. Методы обработки естественного языка в решении задач обнаружения атак социальной инженерии. Кубанский государственный технологический университет, Краснодар, Россия, ISSN 2410-3225. Ежеквартальный рецензируемый, реферируемый научный журнал «Вестник АГУ», 2021, вып. 4 (291).

[6]. Blei D.M., Ng A.Y., Jordan M.I. Latent dirichlet allocation. Journal of machine Learning research, 2003, vol. 3, Jan., pp. 993-1022.

[7]. Марков А.К., Семёночкин Д.О., Кравец А.Г., Яновский Т.А. Сравнительный анализ применяемых технологий обработки естественного языка для улучшения качества классификации цифровых документов. International Journal of Open Information Technologies ISSN: 2307-8162, 2024, vol. 12, no. 3.

[8]. Салып Б.Ю., Смирнов А.А., Ничушкина Т.Н. Анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка. МГТУ им. Н.Э. Баумана, Москва. Научно-образовательный журнал для студентов и преподавателей «StudNet», 2022, №5.

[9]. Пикалёв Я.С., Ермоленко Т. В. Адаптация нейросетевой модели albert для задачи языкового моделирования. Проблемы искусственного интеллекта, 2020, № 3(18). Доступно по ссылке: https://cyberleninka.ru/article/n/adaptatsiya-neyrosetevoy-modeli-albert-dlya-zadachi-yazykovogo-modelirovaniya, дата обращения: 29.03.2026.

[10]. Спивак А.И., Лапшин С.В., Лебедев И.С. Классификация коротких сообщений с использованием векторизации на основе ELMo, Известия ТулГУ. Технические науки, 2019, вып. 10.

[11]. Швенк М.В., Бручес Е.П., Леман А.Я. Сравнение методов машинного обучения для решения задачи анализа тональности. Вестник НГУ. Серия: Информационные технологии, 2024, № 3. Доступно по ссылке: https://cyberleninka.ru/article/n/sravnenie-metodov-mashinnogo-obucheniya-dlya-resheniya-zadachi-analiza-tonalnosti, дата обращения: 29.03.2026. DOI 10.25205/1818-7900-2024-22-3-49-61.

[12]. Ковалев А. Д., Никифоров И. В., Дробинцев П. Д. Автоматизированный подход к семантическому поиску по программной документации на основе алгоритма Doc2Vec. Санкт-Петербургский политехнический университет Петра Великого, DOI: 10.31799/1684-8853-2021-1-17-27.

[13]. Муромцев Д.И., Шилин И.А., Плюхин Д.А., Баймуратов И.Р., Хайдарова Р.Р., Дементьева Ю.Ю., Ожигин Д.А., Малышева Т.А. Построение графов знаний нормативной документации на основе семантического моделирования и автоматического извлечения терминов. Научнотехнический вестник информационных технологий, механики и оптики. 2021, т. 21, № 2, стр. 256-266. DOI: 10.17586/2226-1494-2021-21-2- 256-266.

[14]. Регистр лекарственных средств России. Сетевое издание. ООО «РЛС-Патент». Москва, 2000-2026. Доступно по ссылке: https://www.rlsnet.ru, дата обращения: 30.03.2026.

[15]. Le Q.V., Mikolov T. Distributed Representations of Sentences and Documents. ICML, 2014, vol. 14, pp. 1188-1196.

[16]. Двойникова А. А., Карпов А. А. Аналитический обзор подходов к распознаванию тональности русскоязычных текстовых данных. Информационно-управляющие системы, 2020, № 4, стр. 20–30. DOI: 10.31799/1684-8853-2020-4-20-30.

[17]. Пархоменко П.А., Григорьев А.А., Астраханцев Н.А. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, 2017, т. 29, вып. 2, стр. 161-200. DOI: 10.15514/ISPRAS-2017-29(2)-6. / Parhomenko P.A., Grigorev A.A., Astrakhantsev N.A. A survey and an experimental comparison of methods for text clustering: application to scientific articles. Trudy ISP RAN/Proc. ISP RAS, 2017, vol. 29, issue 2, pp. 161-200 (in Russian). DOI: 10.15514/ISPRAS-2017-29(2)-6.

[18]. Xie P., Xing E. P. Integrating document clustering and topic modeling. arXiv preprint, arXiv: 1309.6874, 2013.

[19]. Шарамет А.В. Снижение размерности данных в системах многоканальной пространственно-временной обработки информации. Вестник ВГТУ, 2023, т. 19, № 1. Доступно по ссылке: https://cyberleninka.ru/article/n/snizhenie-razmernosti-dannyh-v-sistemah-mnogokanalnoy-prostranstvenno-vremennoy-obrabotki-informatsii, дата обращения: 29.03.2026. DOI 10.36622/VSTU.2023.19.1.016.

[20]. Попова И.А., Попова А.А., Соболева Е.Д. Визуализация многомерных наборов данных при помощи алгоритмов снижения пространства признаков PCA и t-SNE. Научно-образовательный журнал для студентов и преподавателей «StudNet», 2020, №11.

[21]. Проневич О.Б., Клокова А.П. Анализ UMAP – метода снижения размерности исходных данных в машинном обучении для прогнозирования отказов в локомотивном комплексе. Надежность, 2022, № 4, стр. 53-62. DOI: 10.21683/1729-2646-2022-22-4-53-62.

[22]. Булыга Ф.С., Курейчик В.М. Кластеризация корпуса текстовых документов при помощи алгоритма k-means. Изв. вузов. Сев.-Кавк. регион. Техн. науки, 2022, № 3, стр. 33-40. DOI: 10.17213/1560-3644-2022-3-33-40.

[23]. Булыга Ф.С., Курейчик В.М. Алгоритмы агломеративной кластеризации применительно к задачам анализа лингвистической экспертной информации. Известия ЮФУ. Технические науки. Раздел II. Методы, модели и алгоритмы обработки информации. DOI 10.18522/2311-3103-2021-6-73-88.

Информация об авторах / Information about authors

Никита Владимирович КИЛЬМИШКИН является сотрудником лаборатории «Прикладное моделирование» Российского Экономического Университета имени Г.В. Плеханова. Его научные интересы включают машинное обучение.

Nikita Vladimirovich KILMISHKIN is an employee of the laboratory "Applied Modeling" of the Plekhanov Russian University of Economics. His research interests include machine learning.

Дмитрий Дмитриевич КУБРАКОВ – является сотрудником лаборатории «Прикладное моделирование» Российского Экономического Университета имени Г.В. Плеханова. Его научные интересы включают машинное обучение, машинная лингвистика, обработка больших данных.

Dmitry Dmitrievich KUBRAKOV is an employee of the laboratory of "Applied Modeling" of the Plekhanov Russian University of Economics. His research interests include machine learning, machine linguistics, and big data processing.

Юрий Павлович ТИТОВ – кандидат технических наук, доцент, ведущий научный сотрудник научной лаборатории «Прикладное моделирование» Российского Экономического Университета имени Г.В. Плеханова. Сфера научных интересов: метаэвристическая оптимизация, графовые модели, машинное обучение, нечеткая логика и имитационные модели.

Yuri Pavlovich TITOV – Cand. Sci. (Tech.), Assoc. Prof., Leading Researcher of the Scientific Laboratory of “Applied Modeling” of the Plekhanov Russian University of Economics. Research interests: metaheuristic optimization, graph models, machine learning, fuzzy logic and simulation models.

Владимир Игоревич ПАНТЕЛЕЕВ – кандидат медицинских наук, старший научный сотрудник научной лаборатории «Медицинская информатика и экономика здравоохранения» Российского Экономического Университета имени Г.В. Плеханова. Области исследований: медицина, медицинские технологии, искусственный интеллект

Vladimir Igorevich PANTELEEV – Cand. Sci. (Medicine), Senior Researcher at the Scientific Laboratory “Medical Informatics and Healthcare Economics,” Federal State Budgetary Educational Institution of Higher Education “Plekhanov Russian University of Economics”. Research interests: medicine, medical technologies, artificial intelligence.

Татьяна Анатольевна КУРОПАТКИНА – кандидат биологических наук, старший научный сотрудник научной лаборатории «Медицинская информатика и экономика здравоохранения» Российского Экономического Университета имени Г.В. Плеханова. Области исследований: медицина, экспериментальная фармакология, медицинские технологии, искусственный интеллект.

Tatiana Anatolyevna KUROPATKINA, Cand. Sci. (Biology), Senior Researcher at the Scientific Laboratory “Medical Informatics and Healthcare Economics”, Federal State Budgetary Educational Institution of Higher Education “Plekhanov Russian University of Economics.” Research areas: medicine, experimental pharmacology, medical technologies, artificial intelligence.

Наталия Андреевна КОЧИНА – кандидат медицинских наук, старший научный сотрудник научной лаборатории «Медицинская информатика и экономика здравоохранения» Российского Экономического Университета имени Г.В. Плеханова. Области исследований: фармакология, медицинские технологии, искусственный интеллект, пролиферативные процессы эндометрия, стероидный транскриптом клеток.

Natalia Andreevna KOCHINA – Cand. Sci. (Medicine), Senior Researcher at the Scientific Laboratory “Medical Informatics and Healthcare Economics,” Federal State Budgetary Educational Institution of Higher Education “Plekhanov Russian University of Economics”. Research areas: pharmacology, medical technologies, artificial intelligence, endometrial proliferative processes, steroid transcriptome of cells.

Полина Михайловна ИВАНОВА – Младший научный сотрудник научной лаборатории «Прикладное моделирование» Российского Экономического Университета имени Г.В. Плеханова. Области исследований: обработка больших данных, машинное обучение.

Polina Mikhailovna IVANOVA – Junior Researcher at the Scientific Laboratory “Applied Modeling”, Plekhanov Russian University of Economics. Research interests: big data processing, machine learning.