Контекстуально-диффузионный метод обогащения TF-IDF матрицы для целей повышения семантической когерентности тематических моделей корпусов новостных текстов.

Д.Г. Родионов, Е.А. Конников, Г.И. Голиков

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

Аннотация: Статья посвящена исследованию ограничений классического метода векторизации TF-IDF в задачах тематического моделирования, обусловленных его зависимостью от модели «мешка слов», и представляет новый подход — TF-IDF с контекстуальным размытием. Показано, что классический TF-IDF, формируя структурно четкие кластеры документов, приводит к созданию семантически некогерентных тем, что является следствием ортогональности и разреженности векторного представления в узкоспециализированных корпусах. Предложенный метод переосмысливает вычисление весов терминов путем введения итеративной процедуры контекстуального размытия, основанной на графе семантической близости, построенном на асимметричной мере совместной встречаемости. На корпусе текстов информационного поля новостей атомной энергетики путем сравнительного анализа с использованием набора кластерных и семантических метрик (коэффициент силуэта, когерентность и др.) установлено, что новый подход, ценой снижения формальных метрик структурной четкости, радикально повышает тематическую когерентность и разнообразие. Это позволяет переходить от статистической кластеризации к извлечению семантически целостных и интерпретируемых тем, что является критически важным для задач мониторинга и анализа больших текстовых данных в узкоспециализированных областях.
Ключевые слова: тематическое моделирование, латентное размещение Дирихле, TF-IDF, контекстуальное размытие, семантическая близость, совместная встречаемость, векторизация текста, модель «мешка слов», тематическая когерентность, обработка естественного языка, коэффициент силуэта, анализ текстовых данных.

Традиционные методы представления текста, такие как TF-IDF (Term Frequency-Inverse Document Frequency), играют ключевую роль в задачах обработки естественного языка, включая тематическое моделирование и классификацию текстов [1]. Однако, несмотря на свою простоту и вычислительную эффективность, классический подход TF-IDF не учитывает семантический контекст слов и их взаимосвязи в корпусе документов. Он рассматривает текст как «мешок слов» (bag-of-words), игнорируя порядок и смысловые отношения между терминами, что приводит к созданию высокоразмерных и разреженных векторных пространств [2,3]. Эта фундаментальная проблема ограничивает способность TF-IDF эффективно

представлять синонимы и омонимы, а также выявлять скрытые тематические связи [4], что особенно критично для небольших или специализированных наборов данных [5].

Для преодоления этих ограничений в последние годы активно развиваются методы, основанные на контекстных встраиваниях (contextual embeddings) и графовых моделях [6,7]. Эти подходы, включая такие как Word2Vec, GloVe и более сложные модели, как BERT, успешно улавливают семантические нюансы слов [8], однако часто требуют больших объемов данных для обучения и могут страдать от проблемы слабой адаптируемости к новым предметным областям [9]. Исследования показывают, что использование графовых представлений позволяет кодировать сложные отношения в тексте и обогащать языковые модели контекстной информацией для лучшего понимания текстовых данных [10]. В частности, методы на основе графов могут быть особенно эффективны в условиях нехватки данных.

В данной работе мы предлагаем новый метод для создания расширенной TF-IDF-матрицы с контекстным размытием (Contextual TF-IDF), который гибридизирует простоту и интерпретируемость традиционного TF-IDF с семантической глубиной графовых моделей. Наш подход строит граф совместной встречаемости лемм в корпусе документов, что позволяет выявить их статистические связи [11]. На основе этого графа выполняется «контекстное размытие» документального вектора, где вес лемм распределяется не только по исходным словам, но и по их семантическим «соседям». Влияние соседних терминов регулируется глубиной размытия и коэффициентом затухания, что позволяет контролировать степень распространения контекстной информации.

Предлагаемый метод направлен на решение ключевых проблем классического TF-IDF, а именно - на повышение семантической насыщенности текстового представления без необходимости использования

сложных нейронных сетей, что делает его особенно эффективным для задач, где важны как терминологические, так и контекстные связи. Полученная матрица может быть успешно использована для различных задач анализа текста, таких как тематическое моделирование с помощью LDA [12,13], что демонстрирует потенциал метода для более точного и осмысленного извлечения знаний из текстовых данных [14,15].

Метод TF-IDF (Term Frequency-Inverse Document Frequency) представляет собой численную статистику, предназначенную для количественной оценки значимости термина в контексте документа, являющегося частью коллекции (корпуса). Данная весовая схема широко применяется в задачах информационного поиска, анализа текстов и машинного обучения для преобразования текстовых данных в векторное представление.

Рассмотрим корпус текстовых документов $D= \{d1, d2,\ldots, dN\}$, где $N=|D|$ — общее число документов. На основе корпуса D после этапа лингвистического препроцессинга (включающего токенизацию, лемматизацию, удаление стоп-слов и фильтрацию по частям речи) формируется словарь уникальных лексических единиц (терминов) $V=\{t1, t2,\ldots,tM\}$, где $M=|V|$ — размер словаря.

Далее в алгоритме TF-IDF компонент TF характеризует локальную значимость термина t в контексте документа d. В классической реализации используется абсолютная частота встречаемости. Для термина $t\in V$ и документа $d\in D$, частота термина определяется как $TF(t,d)=count(t,d)$, где $count(t,d)$ — функция, возвращающая количество вхождений термина t в документ d. Область значений $TF(t,d)$ принимает натуральное множество.

Следующим важнейшим этапом классического алгоритма является компонент IDF, являющийся мерой, отражающей дискриминативную способность термина t в рамках всего корпуса D. Он нивелирует влияние общеупотребительных терминов и повышает значимость терминов,

специфичных для узкого подмножества документов. Данный компонент основывается на т.н. «Документной частоте» (DF), которая определяется как $DF(t)=|\{d \in D | t \in d\}|$ в области значений $1 \leq DF(t) \leq N$. Непосредственно Обратная документная частота (IDF) для термина t вычисляется как логарифм отношения общего числа документов к числу документов, содержащих данный термин:

$$\mathrm{IDF}(t) = \log\left(\frac{N}{\mathrm{DF}(t)}\right)$$

При $DF(t)=N$, что соответствует термину, присутствующему во всех документах, $IDF(t)=\log(1)=0$.

При $DF(t) \to 1$, что соответствует редкому термину, $IDF(t) \to \log(N)$.

Отсюда область определения IDF соответствует $0 \leq IDF(t) \leq \log(N)$.

Итоговый композитный вес $w(t,d)$, присваиваемый термину t в документе d, вычисляется как произведение его локального (TF) и глобального (IDF) компонентов:

$$w(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t)$$

Таким образом, вес термина максимален, когда термин часто встречается в документе (высокое значение TF) и редко в других документах корпуса (высокое значение IDF).

В рамках векторной модели, каждый документ $d_i \in D$ представляется в виде M-мерного вектора $\vec{v}_i$, компонентами которого являются TF-IDF веса всех терминов из словаря V:

$$\vec{v}_i = [w(t_1,d_i), w(t_2,d_i), \ldots, w(t_M,d_i)]$$

Совокупность векторов документов $\vec{v}_1, \ldots, \vec{v}_N$ формирует матрицу W (Document-Term Matrix) размерностью $N \times M$:

$$W = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1M} \\ w_{21} & w_{22} & \ldots & w_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \ldots & w_{NM} \end{pmatrix}$$

где элемент матрицы представляется в виде $w_{ij} = w(t_j, d_i)$. Данная матрица является конечным результатом метода и служит для последующего применения в алгоритмах классификации, кластеризации и тематического моделирования.

Ключевым ограничением классического подхода TF-IDF является его зависимость от модели «мешка слов», в рамках которой предполагается условная независимость терминов друг от друга. Это допущение не позволяет методу улавливать семантическую близость между терминами (например, синонимию) и приводит к созданию излишне разреженных векторов, где вес термина равен нулю, если он явно не присутствует в тексте, даже если его контекст подразумевается. Также еще один из ключевых недостатков классического TF-IDF проявляется именно при работе с узкоспециализированными терминами, хотя компонент IDF предназначен именно для того, чтобы присваивать высокий вес редким и специфичным словам. Проблема возникает, когда в корпусе используется не один, а целый набор синонимичных или семантически близких узкоспециализированных терминов.

Например, в корпусе текстов по атомной энергетике документы могут описывать одну и ту же технологию, используя термины «ВВЭР-1200», «реактор нового поколения» или «ядерная энергетическая установка». Для классического TF-IDF это три совершенно разных, ортогональных признака. В результате, документ, содержащий один из этих терминов, будет иметь нулевой вес по двум другим, что ведет к искусственному занижению тематической близости между текстами, посвященными одной и той же узкой теме.

Именно для решения этих проблем — обогащения векторов документов семантическим контекстом на основе совместной встречаемости терминов и улавливания «семантический ореола» вокруг узкоспециализированных понятий и призван решить новый метод

контекстуального размытия. Он переосмысляет сам подход к векторизации, обогащая представление документа информацией о терминах, которые семантически связаны с присутствующими в тексте, но не упомянуты в нём явно.

Подробно рассмотрим предложенный метод. Этап предобработки полностью идентичен классическому подходу, описанному ранее. Его задача — сформировать нормализованный словарь терминов V и вычислить начальные значения $TF(t,d)$ для всех терминов $t \in V$ в каждом документе $d \in D$. Результатом является набор векторов, представляющих исходные частоты терминов в документах. Вторым этапом является построение графа семантической близости. На этом этапе вводится ключевая структура, отсутствующая в классическом TF-IDF, — взвешенный ориентированный граф $G=(V,E)$, где V — множество вершин, соответствующее словарю терминов, а E — множество ребер, отражающих семантическую связь между терминами. Вес ребра от термина ti к термину tj определяется, как сила ассоциации $A(tj|ti)$, которая количественно характеризует, насколько вероятно встретить термин tj в документе, если в нем уже присутствует термин ti.

Пусть $D_{i,j} = \{d \in D \mid t_i \in d \wedge t_j \in d\}$ — это подмножество документов, содержащих оба термина ti и tj. Тогда сила ассоциации вычисляется как средняя частота термина tj в документах этого подмножества:

$$A(t_j \mid t_i) = \frac{\sum_{d \in D_{i,j}} \mathrm{TF}(t_j, d)}{|D_{i,j}|}$$

Заметим алгебраическое свойство силы ассоциации. Данная мера является асимметричной, то есть, в общем случае, $A(t_j \mid t_i) \neq A(t_i \mid t_j)$. Это позволяет учитывать важную семантическую особенность: асимметричность означает, что семантическая связь между словами имеет направление (присутствие одного слова может быть сильным индикатором присутствия другого, но не наоборот). Также учитывается и иерархические и

гипонимические семантические отношения (частное/общее), что особенно важно для узкоспециализированных терминов.

Вычисления предполагаемые разработанным методом продолжаются определением классических компонентов TF и IDF. Однако здесь TF-веса служат лишь отправной точкой для последующей итеративной процедуры. IDF же вычисляется стандартно и применяется на самом последнем этапе. На основе TF-параметра запускается итеративный процесс: Для каждого документа *d* выполняется *L* шагов расширения (где *L* — глубина размытия). На каждом шаге $k \in \{1, \ldots, L\}$, для каждого термина $t_j \in V$ вычисляется приращение частоты $\Delta\mathrm{TF}^{(k)}(t_j, d)$.

Это приращение формируется за счет преемственности веса от семантически связанных терминов *ti*, которые уже имели ненулевой вес на предыдущем шаге. Формула приращения на шаге k таким образом выглядит, как:

$$\Delta\mathrm{TF}^{(k)}(t_j, d) = \alpha^k \sum_{t_i \in V} \left(\Delta\mathrm{TF}^{(k-1)}(t_i, d) \times A(t_j \mid t_i) \right)$$

Где:

1. $\Delta\mathrm{TF}^{(0)}(t_i, d) = \mathrm{TF}^{(0)}(t_i, d)$.
2. α — коэффициент экспоненциального затухания (0<α<1), который уменьшает вклад более удаленных семантических соседей.
3. Суммирование производится по всем терминам *ti* из словаря, которые получили приращение на предыдущем шаге.

Итоговая, обогащенная контекстом частота термина *t* в документе *d* формируется как сумма его исходной частоты и всех полученных приращений до глубины *L*:

$$\mathrm{TF}_C(t, d) = \mathrm{TF}^{(0)}(t, d) + \sum_{k=1}^{L} \Delta\mathrm{TF}^{(k)}(t, d)$$

Эта величина, $\mathrm{TF}_C(t,d)$, заменяет собой классическую $\mathrm{TF}(t,d)$, так как теперь она включает в себя информацию не только о явном присутствии термина в документе, но и о присутствии семантически близких ему терминов. Финальный вес $w_{new}(t,d)$ для каждого термина в каждом документе вычисляется путем умножения его контекстуализированной частоты TF_C на стандартную обратную документную частоту IDF:

$$w_{new}(t,d) = \mathrm{TF}_C(t,d) \times \mathrm{IDF}(t)$$

Результатом является новая матрица "Документ-Термин" W' , где значения в ячейках $w'_{ij} = w_{new}(t_j, d_i)$ отражают обогащенную семантическим контекстом важность каждого термина для каждого документа.

Аппробация разработанного алгоритма проводилась на массиве новостей в узкоспециализированной области атомной энергетики. Сперва рассмотрим результаты классического TF-IDF метода, затем метода с контекстуальным размытием и сравним полученные результаты.

Для определения оптимального числа тем (k) был применен метод локтя, основанный на метрике перплексии модели Latent Dirichlet Allocation (LDA) в диапазоне от 2 до 15 тем. Визуальный анализ графика перплексии позволил идентифицировать точку "изгиба", соответствующую оптимальному компромиссу между сложностью модели и качеством аппроксимации данных, при k = 7 (рисунок 1). Данное значение было использовано для построения и последующей оценки финальной тематической модели.

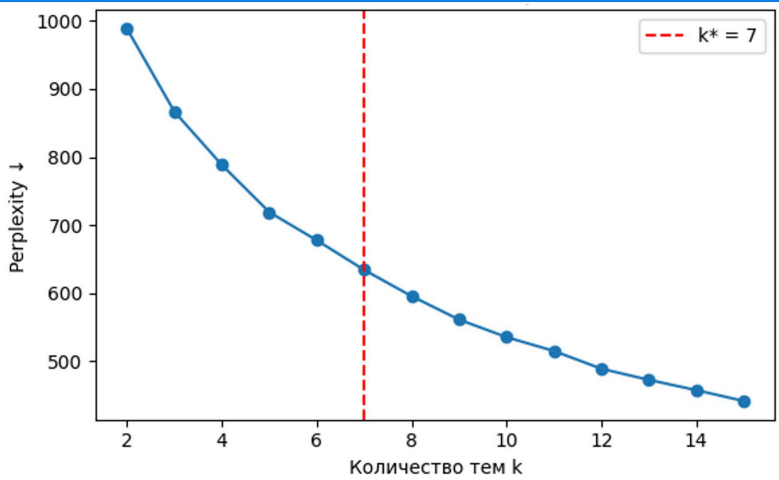

Рис. 1. - Метод локтя для модели LDA на классическом TD-IDF

Гистограммы распределения идентифицированных тем по документам представлены на рисунке 2.

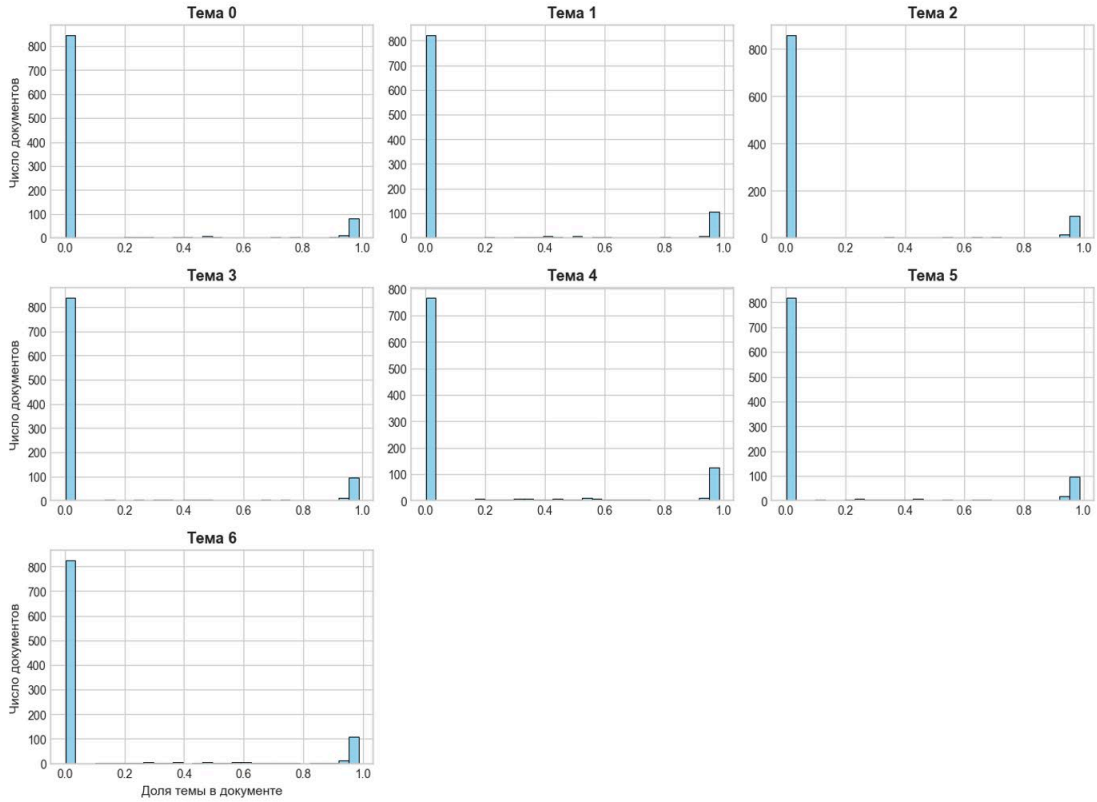

Рис. 2. - Гистограммы распределения идентифицированных тем по документам классическим TD-IDF методом

Метрики качества кластеризации классическим TF-IDF методом представлены на рисунке 3.

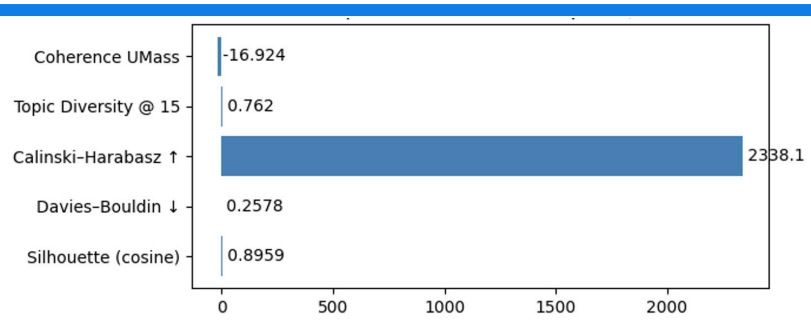

Рис. 3. – Метрики качества кластеризации классическим TD-IDF методом

Оценка структурного качества полученной кластеризации производилась с использованием набора стандартных метрик. Коэффициент силуэта, рассчитанный с использованием косинусной метрики, достиг исключительно высокого значения 0.8959, что свидетельствует о высокой плотности внутрикластерных распределений и значительном расстоянии между центроидами кластеров. Этот вывод подтверждается низким значением индекса Дэвиса-Болдина (0.2578) и высоким значением индекса Калински-Харабас (2338.1), которые также указывают на формирование четко определенных и хорошо разделенных тематических групп документов.

Анализ распределения долей тем по документам выявил высокую степень тематической специализации. Для каждой из семи тем наблюдается сильно асимметричное распределение, где подавляющее большинство документов имеет долю, близкую к нулю, в то время как небольшое, четко выраженное подмножество документов практически полностью относится к данной теме. Это указывает на то, что модель успешно идентифицировала узкоспециализированные группы текстов.

Несмотря на превосходные показатели структурной четкости, оценка семантического качества тем выявила существенные недостатки. Показатель тематического разнообразия на уровне 0.762 свидетельствует о приемлемой, но не идеальной уникальности наборов ключевых слов для каждой темы. Ключевым индикатором является тематическая когерентность (UMass Coherence), которая продемонстрировала крайне низкое значение -16.924. Столь сильно отрицательный результат указывает на то, что наиболее

вероятные термины в рамках одной и той же темы семантически не связаны и редко встречаются совместно в документах корпуса.

Таким образом, результаты анализа тематической модели, построенной на базе классической TF-IDF матрицы, демонстрируют парадоксальную картину. С одной стороны, метрики структурного качества указывают на превосходную разделимость кластеров, что обусловлено способностью TF-IDF выделять документы по уникальным, высокочастотным терминам. С другой стороны, низкая тематическая когерентность свидетельствует о семантической несвязности этих тем, что является прямым следствием модели "мешка слов", неспособной уловить контекстуальные связи между узкоспециализированными, но семантически близкими терминами.

Теперь рассмотрим результаты разработанного метода на том же массиве данных. Аналогичным образом, для матрицы, построенной с использованием контекстуального размытия, оптимальное число тем было определено методом локтя и составило k = 6 (рисунок 4).

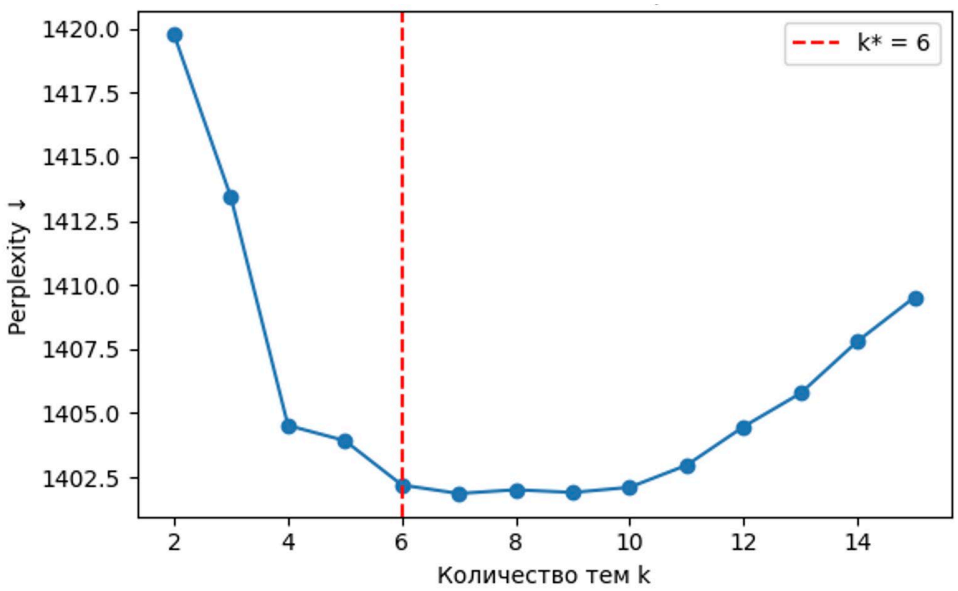

Рис. 4. - Метод локтя для модели LDA на контекстуальном TD-IDF

Гистограммы распределения идентифицированных тем по документам представлены на рисунке 5.

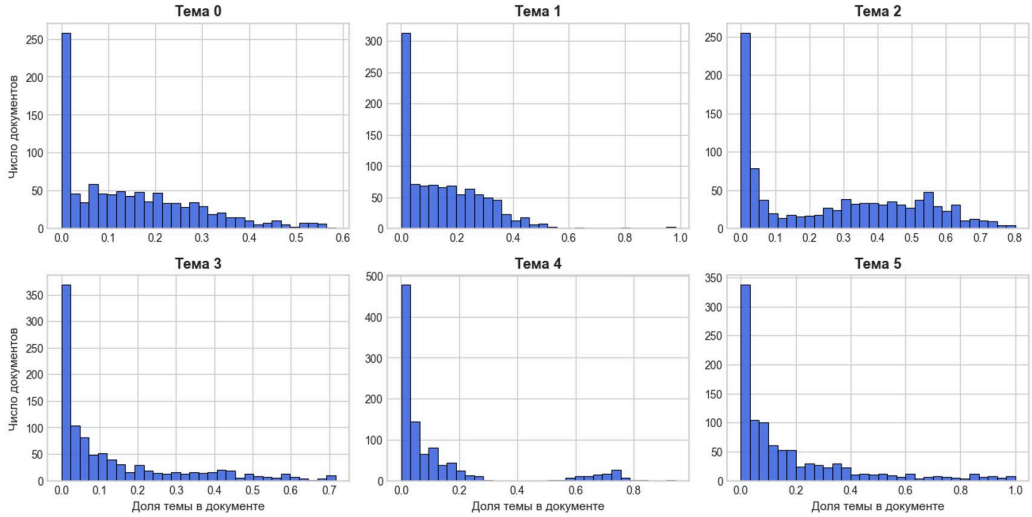

Рис. 5. - Гистограммы распределения идентифицированных тем по документам контекстуальным TD-IDF методом

Метрики качества кластеризации контекстуальным TF-IDF методом представлены на рисунке 6.

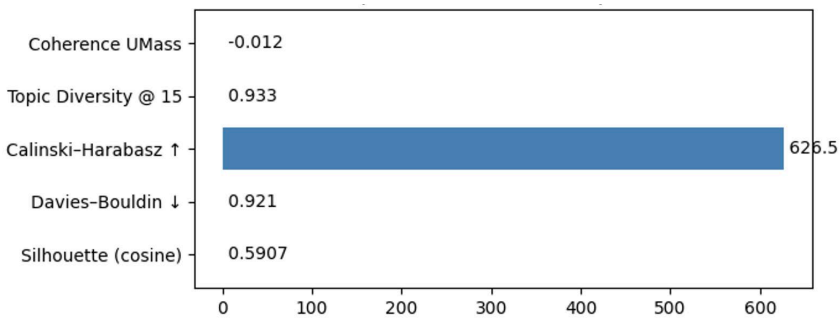

Рис. 6. – Метрики качества кластеризации контекстуальным TD-IDF методом

Оценка структурного качества кластеризации выявила ключевое отличие нового подхода. Формально, все метрики оказались ниже, чем у классического метода: Коэффициент силуэта составил 0.5907, индекс Дэвиса-Болдина — 0.921, а индекс Калински-Харабас — 626.5. Однако данное снижение является ожидаемым и позитивным следствием контекстуального размытия: метод формирует более "мягкие", семантически пересекающиеся тематические группы, что ближе к реальному распределению тем в текстах, но приводит к менее четкому структурному разделению.

Это различие наглядно подтверждается анализом распределения долей тем (Рисунок 5). В отличие от бимодальной структуры классического метода, здесь наблюдается унимодальное распределение с центром в области низких значений, плавно затухающее к единице. Это означает, что документы теперь имеют небольшую, но ненулевую принадлежность ко многим темам, что отражает их естественное тематическое разнообразие.

Ключевые преимущества нового метода проявляются в оценке семантического качества тем. Показатель тематического разнообразия (Topic Diversity) достиг высокого значения 0.933, свидетельствуя о большей уникальности и различимости сформированных тем. Наиболее значимым результатом является показатель тематической когерентности (UMass Coherence), который продемонстрировал значение -0.012. Этот результат, близкий к нулю, указывает на радикальное улучшение семантической связности: наиболее вероятные термины в рамках одной и той же темы теперь часто встречаются совместно в документах корпуса.

Таким образом, в отличие от классического подхода, который формировал структурно идеальные, но семантически несвязные кластеры, новый метод успешно решает эту проблему. Он сознательно "жертвует" формальной структурной четкостью в пользу создания семантически когерентных и легко интерпретируемых тем, что является основной целью тематического моделирования в узкоспециализированных корпусах.

Проведенный сравнительный анализ двух подходов к векторизации текстовых данных — классического TF-IDF и расширенного метода с контекстуальным размытием — убедительно демонстрирует преимущества последнего для задач тематического моделирования в узкоспециализированных корпусах. Анализ, основанный на классической TF-IDF матрице, продемонстрировал парадоксальные результаты. С одной стороны, модель успешно сформировала структурно четкие и хорошо разделенные кластеры документов. С другой стороны, была выявлена крайне

низкая семантическая когерентность тем, что свидетельствует о том, что ключевые слова внутри одной темы оказались логически не связаны. Это подтвердило фундаментальное ограничение модели «мешка слов», неспособной уловить контекст.

Предложенный метод контекстуального размытия, напротив, показал свою состоятельность именно в решении этой семантической проблемы. Несмотря на формально менее выраженную структурную четкость кластеров, он позволил сформировать темы с высокой семантической связностью и разнообразием. Слова в рамках каждой темы оказались логически взаимосвязаны, что делает их легко интерпретируемыми для экспертного анализа.

Таким образом, проведенное исследование доказывает, что обогащение векторов документов семантическим контекстом является ключевым фактором для успешного тематического моделирования. Метод контекстуального размытия, в отличие от классического TF-IDF, позволяет перейти от простого статистического группирования документов к выявлению подлинных, семантически целостных тематических структур. Он является значительно более адекватным и эффективным инструментом для глубокого анализа текстов, особенно в областях со сложной и узкоспециализированной терминологией.

Литература

1. Сайгин А.А., Федосин С.А. Влияние уменьшения размерности словоформенных эмбеддингов на качество классификации текста // Инженерный вестник Дона, 2025, № 1. URL: ivdon.ru/ru/magazine/archive/n1y2025/9741/.

2. Корчагин С.А., Сердечный Д.В., Окунев А.И., Андриянов Н.А. Методы машинного обучения для автоматической обработки документов //

Инженерный вестник Дона, 2025, № 3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9914/.

3. Atoum I., Otoom A.A. Enhancing Software Effort Estimation with Pre-Trained Word Embeddings: A Small-Dataset Solution for Accurate Story Point Prediction // Electronics, 2024, Vol. 13, № 23, P. 4843. DOI: 10.3390/electronics13234843.

4. Сайгин А.А., Федосин С.А. Использование метода определения схожести слов для оценки алгоритмов векторизации текста // Инженерный вестник Дона, 2024, № 7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9381/.

5. Mimno D., Wallach H.M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Proceedings of the 2011 conference on empirical methods in natural language processing. 2011. pp. 262–272.

6. Alangari H., Algethami N. Exploring the Effects of Pre-Processing Techniques on Topic Modeling of an Arabic News Article Data Set // Applied Sciences, 2024, Vol. 14, № 23, P. 11350, DOI: 10.3390/app142311350.

7. Mitrov G., Stanoev B., Gievska S., Mirceva G., Zdravevski E. Combining Semantic Matching, Word Embeddings, Transformers, and LLMs for Enhanced Document Ranking: Application in Systematic Reviews // Information, 2022, Vol. 8, № 9, P. 110. DOI: 10.3390/info8090110.

8. Носко В.И., Свечкарев В.П., Розин М.Д. Методика и фреймворк конструирования лингвистических моделей для сетевого мониторинга // Инженерный вестник Дона, 2015, № 4. URL: ivdon.ru/ru/magazine/archive/n4y2015/3409/.

9. Rodionov D., Lyamin B., Konnikov E., Obukhova E., Golikov G., Polyakov P. Integration of Associative Tokens into Thematic Hyperspace: A Method for Determining Semantically Significant Clusters in Dynamic Text Streams // Big Data Cogn. Comput., 2025, Vol. 9, № 8, P. 197, DOI: 10.3390/bdcc9080197.

10. Tan L., Yi J., Yang F. Improving Performance of Massive Text Real-Time Classification for Document Confidentiality Management // Applied Sciences, 2024, Vol. 14, № 4, P. 1565, DOI: 10.3390/app14041565.

11. Денискин А.В., Федосин С.А., Плотникова Н.П. Применение иерархической кластеризации DIANA для улучшения качества классификации текста // Инженерный вестник Дона, 2023, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2023/8662/.

12. Родионов Д.Г., Конников Е.А., Пашинина П.А., Шаныгин С.И. Тематическое моделирование информационной среды медиакомпаний: инструментальный комплекс LDA-TF-IDF // Мягкие измерения и вычисления, 2024, Т. 76, № 3, С. 72–84.

13. Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., Li Y., Zhao L. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, and challenges // Multimedia Tools and Applications, 2019, Vol. 78, pp. 15169–15211.

14. Albalawi R., Yeap T.H., Benyoucef M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis // Frontiers in Artificial Intelligence, 2020, Vol. 3, P. 42.

15. Конников Е.А. Система анализа эффекта информационного импульса в цифровой среде // Естественно-гуманитарные исследования, 2024, № 5 (55), С. 176–182.

References

1. Saygin A.A., Fedosin S.A. Inzhenernyj vestnik Dona, 2025, No. 1. URL: ivdon.ru/ru/magazine/archive/n1y2025/9741/.

2. Korchagin S.A., Serdechnyy D.V., Okunev A.I., Andriyanov N.A. Inzhenernyj vestnik Dona, 2025, No. 3. URL: ivdon.ru/ru/magazine/archive/n3y2025/9914/.

3. Atoum I., Otoom A.A. Electronics, 2024, Vol. 13, № 23, P. 4843, DOI: 10.3390/electronics13234843.

4. Saygin A.A., Fedosin S.A. Inzhenernyj vestnik Dona, 2024, No. 7. URL: ivdon.ru/ru/magazine/archive/n7y2024/9381/.

5. Mimno D., Wallach H.M., Talley E., Leenders M., McCallum A. Proceedings of the 2011 conference on empirical methods in natural language processing. 2011. P. 262–272.

6. Alangari H., Algethami N. Applied Sciences, 2024, Vol. 14, № 23, P. 11350, DOI: 10.3390/app142311350.

7. Mitrov G., Stanoev B., Gievska S., Mirceva G., Zdravevski E. Information, 2022, Vol. 8, № 9, P. 110, DOI: 10.3390/info8090110.

8. Nosko V.I., Svechkarev V.P., Rozin M.D. Inzhenernyj vestnik Dona, 2015, No. 4. URL: ivdon.ru/ru/magazine/archive/n4y2015/3409/.

9. Rodionov D., Lyamin B., Konnikov E., Obukhova E., Golikov G., Polyakov P. Big Data Cogn. Comput., 2025, Vol. 9, № 8, P. 197, DOI: 10.3390/bdcc9080197.

10. Tan L., Yi J., Yang F. Applied Sciences, 2024, Vol. 14, № 4, P. 1565, DOI: 10.3390/app14041565.

11. Deniskin A.V., Fedosin S.A., Plotnikova N.P. Inzhenernyj vestnik Dona, 2023, No. 9. URL: ivdon.ru/ru/magazine/archive/n9y2023/8662/.

12. Rodionov D.G., Konnikov E.A., Pashinina P.A., Shanygin S.I. Myagkie izmereniya i vy`chisleniya, 2024, Vol. 76, No. 3, pp. 72–84.

13. Jelodar H., Wang Y., Yuan C., Feng X., Jiang X., Li Y., Zhao L. Multimedia Tools and Applications, 2019, Vol. 78, pp. 15169–15211.

14. Albalawi R., Yeap T.H., Benyoucef M. Frontiers in Artificial Intelligence, 2020, Vol. 3, P. 42.

15. Konnikov E.A. A Estestvenno-gumanitarnye issledovaniya, 2024, No. 5 (55), pp. 176–182.

Дата поступления: 2.08.2025

Дата публикации: 25.09.2025