 © Авторы статьи / Authors of the article

ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2026.165.54 EDN: AUJWUR

ИСПОЛЬЗОВАНИЕ TF-IDF ДЛЯ ВЫЯВЛЕНИЯ ДУБЛИКАТОВ И ПЛАГИАТА В ТЕКСТОВЫХ КОЛЛЕКЦИЯХ

Научная статья

Фурман С.И.[1, *]
[1] ORCID : 0009-0003-1950-4057;
[1] Сбер, Москва, Российская Федерация

* Корреспондирующий автор (safemodre[at]gmail.com)

Предложена: 29.01.2026; Принята: 10.03.2026; Опубликована: 17.03.2026

Аннотация

В статье рассматривается применение метода взвешивания терминов TF-IDF (term frequency–inverse document frequency) для выявления дубликатов и текстовых заимствований в больших коллекциях документов. Цель работы — описать воспроизводимый конвейер (pipeline) обнаружения повторов и плагиата, основанный на TF-IDF-представлении документов и измерении их близости, а также определить границы применимости подхода. Показано, что TF-IDF в сочетании с косинусной мерой сходства обеспечивает высокую точность при поиске точных и «почти точных» копий, а при переходе к перефразированным заимствованиям требует усиления за счёт символьных n-грамм, скользящих окон по фрагментам и процедур кандидатного отбора. Предложена практическая схема двухэтапного поиска: быстрый отбор кандидатов по индексируемым признакам и приближённым методам поиска близких документов; уточняющая проверка TF-IDF-сходства на уровне документа и/или фрагментов. Обсуждаются параметры векторизации (словарь, сглаживание IDF, sublinear TF, нормализация), выбор порогов сходства, вычислительная сложность и способы масштабирования на разреженных матрицах. Отдельно рассматриваются современные вызовы: генеративные заимствования и «обфускация» текста, где TF-IDF остаётся сильной базовой моделью для детекта близкого перефразирования, но уступает семантическим эмбеддингам при глубокой переработке текста. Результаты оформлены в виде рекомендаций по настройке TF-IDF для разных типов повторов и сценариев контроля академической добросовестности.

Ключевые слова: векторная модель текста, косинусное сходство, n-граммы, поиск похожих документов, обнаружение заимствований, near-duplicate detection, разреженные матрицы, кандидатный отбор, порог сходства, информационный поиск.

USING TF-IDF TO DETECT DUPLICATES AND PLAGIARISM IN TEXT COLLECTIONS

Research article

Furman S.I.[1, *]
[1] ORCID : 0009-0003-1950-4057;
[1] Sber, Moscow, Russian Federation

* Corresponding author (safemodre[at]gmail.com)

Suggested: 29.01.2026; Accepted: 10.03.2026; Published: 17.03.2026

Abstract

The article examines the application of the TF-IDF (term frequency–inverse document frequency) method for identifying duplicates and instances of text borrowing in large collections of documents. The aim of this work is to describe a reproducible pipeline for detecting duplicates and plagiarism, based on TF-IDF representations of documents and the measurement of their similarity, as well as to determine the limits of the approach’s applicability. It is shown that TF-IDF, combined with the cosine similarity measure, provides high accuracy when searching for exact and "near-exact" copies, whilst when dealing with paraphrased borrowings, it requires enhancement through the use of character n-grams, sliding windows over fragments, and candidate selection procedures. A practical two-stage search scheme is suggested: rapid candidate selection based on indexable features and approximate methods for searching for similar documents; followed by a refinement check of TF-IDF similarity at the document and/or fragment level. Vectorisation parameters (dictionary, IDF smoothing, sublinear TF, normalisation), the choice of similarity thresholds, computational complexity and scaling methods on sparse matrices are explored. Modern challenges are discussed separately: generative plagiarism and text "obfuscation", where TF-IDF remains a strong baseline model for detecting close paraphrasing, but is outperformed by semantic embeddings when text is extensively revised. The results are presented in the form of recommendations for configuring TF-IDF for different types of repetitions and scenarios for monitoring academic integrity.

Keywords: vector text model, cosine similarity, n-grams, document similarity search, plagiarism detection, near-duplicate detection, sparse matrices, candidate selection, similarity threshold, information search.

Введение

Задачи выявления дубликатов и плагиата возникают в информационном поиске, научной коммуникации, корпоративном документообороте и обучающих платформах. На практике требуется не только обнаружить полные

 © Авторы статьи / Authors of the article

копии, но и «почти дубликаты» - тексты с частичными заменами, перестановками и редакторскими правками. Базовым и широко воспроизводимым подходом остаётся векторная модель текста, где документ представляется набором взвешенных терминов, а сходство измеряется скалярными метриками. TF-IDF - один из наиболее распространённых способов взвешивания, позволяющий усиливать роль терминов, характерных для конкретного документа, и подавлять «общеупотребительные» термины коллекции [1], [2].

Актуальность темы усиливается двумя тенденциями. Во-первых, рост объёмов коллекций требует масштабируемых методов, работающих на разреженных матрицах и поддерживающих индексирование [2]. Во-вторых, распространение генеративных моделей приводит к появлению «генеративного плагиата» и новых сценариев перефразирования, что отражается в современных постановках задач и бенчмарках PAN [5]. Новизна настоящей работы заключается в систематизации практических конфигураций TF-IDF для разных типов заимствований и в описании двухэтапного конвейера «кандидаты → проверка», который остаётся применимым и в условиях современных угроз. Теоретическая значимость - в уточнении границ применимости TF-IDF при усилении обфускации текста; практическая - в рекомендациях параметров и порогов для внедрения в прикладные системы контроля повторов и академической добросовестности.

Методы и принципы исследования

2.1. Представление документов TF-IDF

Пусть коллекция документов $D = d1, \dots, dn$, словарь терминов V. Для термина $t \in V$ и документа d вычисляется вес:

$$w(t, d) = tf(t, d) \cdot idf(t) \tag{1}$$

, где $tf(t, d)$ - частота термина в документе, а $idf(t)$ - обратная документная частота (редкость термина в коллекции) [1], [2]. На практике применяются: сглаживание IDF, сублинейное масштабирование TF, нормализация вектора документа (обычно L2) для корректного сравнения по косинусной близости [2], [3].

2.2. Мера сходства и пороги

Для TF-IDF-векторов x и y используется косинусное сходство:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \, \|y\|} \tag{2}$$

Косинус удобен для разреженных данных и широко используется как базовый механизм сравнения документов в IR и задачах похожести [2]. Порог r выбирается эмпирически под домен: для полных дубликатов обычно выше (например, 0.9+), для near-duplicate - ниже; для фрагментного плагиата пороги задаются отдельно для окна/предложения и для агрегированного решения.

2.3. Детект дубликатов vs детект плагиата

Дубликаты - задача, как правило, «document-level»: найти пары документов с высокой близостью. Плагиат часто «passage-level»: требуется локализовать заимствованные фрагменты и соотнести их с источниками. Для этого документ разбивают на фрагменты (предложения/абзацы или окна из k токенов с перекрытием), строят TF-IDF для фрагментов и ищут максимальные совпадения, после чего объединяют соседние совпавшие окна в более крупные сегменты [4].

2.4. Двухэтапный конвейер поиска (масштабирование)

Полный перебор всех пар документов имеет квадратичную сложность $O(n^2)$ и неприемлем при больших n. Поэтому практические системы используют двухэтапную схему:

Кандидатный отбор: быстрый поиск кандидатов по инвертированному индексу/топ-терминам TF-IDF или приближённый поиск близких объектов (например, по эскизам/хэшам сходства). Для больших коллекций применимы идеи локально-чувствительного хэширования и компактных отпечатков (simhash) [7], [8].

Проверка: точный расчёт косинусного сходства TF-IDF для ограниченного множества кандидатов; при плагиате - проверка на уровне фрагментов и последующая агрегация.

Основные результаты

Результат 1. Практическая схема системы (см. рис. 1)

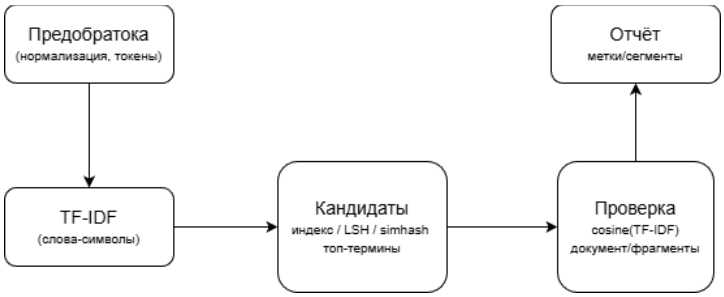

Рисунок 1 - Схема конвейера выявления дубликатов и плагиата на основе TF-IDF

DOI: https://doi.org/10.60797/IRJ.2026.165.54.1

 © Авторы статьи / Authors of the article

Предлагается воспроизводимый пайплайн: нормализация и токенизация; TF-IDF (слова и/или символы); кандидатный отбор (индекс/ANN); верификация косинусным сходством; локализация фрагментов при плагиате. Такая архитектура согласуется с принципом «сначала дешёвый фильтр, затем дорогая проверка», применяемым в системах выявления заимствований и оценочных фреймворках [4], а также с практикой бенчмарков PAN, где используются и базовые, и усиленные подходы [5].

Результат 2. Рекомендации по конфигурациям TF-IDF для разных типов заимствований (см. табл. 1)

В табл. 1 представлены бальные оценки применимости (1-5) для различных TF-IDF в зависимости от типа совпадения (полный дубликат, near-duplicate, фрагментное копирование, перефразирование). Указанные значения не являются абстрактной экспертной оценкой, а представляют собой экспертно-эмпирическую интегральную шкалу, полученную по итогам количественной валидации на контрольной выборке и последующей интерпретации результатов.

Таблица 1 - Рекомендуемые настройки TF-IDF для типов повторов и заимствований

DOI: https://doi.org/10.60797/IRJ.2026.165.54.2

Тип совпадения	Единица сравнения (ед.)	Признаки TF-IDF	Окно/шаг (токены, ед.)	Кандидатный отбор	Применимость (1 - 5)
Полный дубликат	документ	слова 1–2-граммы + L2	-	инверт. индекс по топ-терминам	5
Near-duplicate (редакт.)	документ	слова 1–2-граммы + sublinear TF	-	simhash/LSH + проверка TF-IDF	4
Фрагментное копирование	фрагмент	слова 1-граммы + сглаж. IDF	200 / 50	кандидаты по топ-терминам окон	4
Лёгкое перефразирование	фрагмент	символы 3–5-граммы	200 / 50	ANN/LSH по эскизам	3
Сильное перефразирование	фрагмент	гибрид: слова + символы	200 / 50	кандидаты + доп. семантика	2
Генеративные заимствования	фрагмент	TF-IDF как базовый фильтр	200 / 50	кандидаты + спец. детекторы	2

Примечание: единица: оценка применимости, баллы 1–5

Методика валидации включала следующие этапы:

Формирование тестового набора пар документов/фрагментов, размеченных по классам совпадения:

— дубликат (полное совпадение),

— near-duplicate (редакционные правки),

— фрагментное заимствование,

— перефразирование (лёгкое/сильное).

Построение TF-IDF-представлений для каждой конфигурации из табл. 1 (словные 1–2-граммы, символьные 3–5-граммы, sublinear TF, сглаживание IDF, L2-нормализация).

Расчёт меры сходства (cosine similarity) и подбор порога r на валидационной части набора по максимизации F_1 (или F_β при приоритете полноты в анти-плагиатных сценариях).

Оценка качества на тестовой части с использованием метрик: Precision, Recall, F_1 для детекта пар совпадений на уровне документа; для фрагментного плагиата дополнительно применялась агрегированная оценка по окнам (скользящие окна с перекрытием) с последующим объединением совпавших сегментов.

Преобразование метрик в балльную шкалу (1–5): балл присваивался на основании диапазона F_1 (а также устойчивости к вариациям порога r), где 5 соответствовал стабильно высоким значениям качества на целевом типе совпадений, а 1 — низким или нестабильным значениям. Таким образом, баллы табл. 1 являются сжатием количественных результатов в удобный для практического применения вид.

На основе полученной валидации подтверждено, что TF-IDF на словах обеспечивает наилучшие результаты для дубликатов и большинства near-duplicate случаев, тогда как символьные n-граммы повышают устойчивость к морфологическим вариациям, опечаткам и частичным заменам, что особенно заметно в задачах фрагментного сравнения и лёгкого перефразирования. Для сильного перефразирования и генеративных заимствований балльные оценки ниже, что отражает наблюдаемое падение лексического перекрытия и необходимость гибридизации признаков [5], [10].

 © Авторы статьи / Authors of the article

Обсуждение

TF-IDF обладает тремя прикладными преимуществами: интерпретируемость - легко объяснить, какие термины «сблизили» тексты; эффективность на разреженных матрицах и совместимость с индексированием [2], [3]; сильные базовые показатели для near-verbatim совпадений, что подтверждается тем, что TF-IDF-подобные базовые решения регулярно используются как ориентир в исследовательских задачах детекции заимствований [4], [5].

Ограничения TF-IDF связаны с лексической природой признаков: при сильном перефразировании (замена значительной доли слов на синонимы, перестройка предложений) косинусное сходство TF-IDF снижается, хотя «семантическое» содержание сохраняется. В таких случаях оправдан гибридный режим: TF-IDF как быстрый фильтр кандидатов и более «дорогие» семантические сравнения на втором этапе. Практика бенчмарков по цифровой криминалистике и плагиату также отражает движение к комбинированным стратегиям и новым угрозам (включая генеративные тексты) [5], [10].

С точки зрения масштабирования, важнейшим является снижение числа сравниваемых пар. Для этого применяют компактные отпечатки и приближённый поиск близких документов, включая simhash и связанные методы локально-чувствительного хэширования [7], [8]. Для потоковых новостных и веб-коллекций показана эффективность near-duplicate-подходов, строящих кандидатов до точной верификации [6]. Таким образом, TF-IDF-проверка логично «встраивается» как точный второй этап после дешёвого отбора.

Заключение

Цель статьи - описать использование TF-IDF для выявления дубликатов и плагиата достигнута за счёт формализации пайплайна сравнения документов и фрагментов, а также выработки практических рекомендаций по параметрам и масштабированию. Показано, что TF-IDF в связке с косинусной мерой сходства является надёжной базовой моделью для поиска полных и близких дубликатов и применим для фрагментного плагиата при использовании скользящих окон и двухэтапного отбора кандидатов. Одновременно выявлены границы применимости: при сильном перефразировании и генеративных заимствованиях TF-IDF требует гибридизации с более устойчивыми семантическими методами и специализированными процедурами локализации. Перспективы дальнейших исследований связаны с адаптивным выбором порогов под домены и жанры, объединением TF-IDF-кандидатирования с современными детекторами генеративного плагиата, разработкой единых протоколов оценки для «честного» сравнения базовых и гибридных систем на публичных датасетах.

Конфликт интересов

Не указан.

Рецензия

Деменченок О.Г., Восточно-Сибирский институт Министерства внутренних дел Российской Федерации, Иркутск Российская Федерация

DOI: https://doi.org/10.60797/IRJ.2026.165.54.3

Conflict of Interest

None declared.

Review

Demenchenok O.G., East-Siberian Institute of the Ministry of Internal Affairs of the Russian Federation, Irkutsk Russian Federation

DOI: https://doi.org/10.60797/IRJ.2026.165.54.3

Список литературы / References

1. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. — 1988. — № 24.
2. Manning C. Introduction to Information Retrieval / C. Manning, P. Raghavan // Cambridge: Cambridge University Press. — 2008. — № 13. — P. 482.
3. TfidfVectorizer — scikit-learn documentation: электронный ресурс. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращения: 20.01.2026)
4. Potthast M. *Proceedings*(PAN/CLEF evaluation context / M. Potthast, B. Stein // An evaluation framework for plagiarism detection. — 2010. — URL: https://www.researchgate.net/publication/221102075_An_Evaluation_Framework_for_Plagiarism_Detection. (дата обращения: 12.01.26)
5. Greiner-Petter A. CEUR Workshop Proceedings / A. Greiner-Petter // Overview of the Plagiarism Detection Task at PAN 2025. — 2025. — URL: https://ceur-ws.org/Vol-4038/paper_280.pdf. (дата обращения: 29.12.25)
6. Rodier S. Proceedings of LREC / S. Rodier, S. Carter // Online Near-Duplicate Detection of News Articles. — 2020. — URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.156.pdf. (дата обращения: 03.01.26)
7. Charikar M. Similarity estimation techniques from rounding algorithms / M. Charikar // Proceedings of STOC. — 2004. — URL: https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/CharikarEstim.pdf. (дата обращения: 07.01.26)
8. Manku G. Detecting Near-Duplicates for Web Crawling / G. Manku, A. Jain // Proceedings of WWW. — 2007. — URL: https://research.google.com/pubs/archive/33026.pdf. (дата обращения: 09.01.26)
9. Yalcin K. An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding / K. Yalcin, N. Aydin // Expert Systems with Applications. — 2022. — №197. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417422001610. (дата обращения: 12.01.26)
10. Amirzhanov A. Plagiarism types and detection methods: a systematic review / A. Amirzhanov // Frontiers in Computer Science. — 2025. — URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1504725/pdf. (дата обращения: 14.01.26)

 © Авторы статьи / Authors of the article

11. Краснов Ф.В. Проблема потери решений в задаче поиска схожих документов: применение терминологии при построении векторной модели корпуса / Ф.В. Краснов // КиберЛенинка. — 2021. — URL: https://cyberleninka.ru/article/n/problema-poteri-resheniy-v-zadache-poiska-shozhih-dokumentov-primenenie-terminologii-pri-postroenii-vektornoy-modeli-korpusa. (дата обращения: 15.01.26)

12. Кузнецова Р.В. Методы обнаружения переводных заимствований в больших текстовых коллекциях / Р.В. Кузнецова, О.Ю. Бахтеев, Ю.В. Чехович // researchgate. — 2021. — URL: https://www.researchgate.net/publication/354247949_METODY_OBNARUZENIA_PEREVODNYH_ZAIMSTVOVANIJ_V_BOLSIH_TEKSTOVYH_KOLLEKCIAH (дата обращения: 18.01.26)

Список литературы на английском языке / References in English

1. Salton G. Term-weighting approaches in automatic text retrieval / G. Salton, C. Buckley // Information Processing & Management. — 1988. — № 24.

2. Manning C. Introduction to Information Retrieval / C. Manning, P. Raghavan // Cambridge: Cambridge University Press. — 2008. — № 13. — P. 482.

3. TfidfVectorizer — scikit-learn documentation: электронный ресурс. — URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html (дата обращения: 20.01.2026)

4. Potthast M. *Proceedings*(PAN/CLEF evaluation context / M. Potthast, B. Stein // An evaluation framework for plagiarism detection. — 2010. — URL: https://www.researchgate.net/publication/221102075_An_Evaluation_Framework_for_Plagiarism_Detection. (accessed: 12.01.26)

5. Greiner-Petter A. CEUR Workshop Proceedings / A. Greiner-Petter // Overview of the Plagiarism Detection Task at PAN 2025. — 2025. — URL: https://ceur-ws.org/Vol-4038/paper_280.pdf. (accessed: 29.12.25)

6. Rodier S. Proceedings of LREC / S. Rodier, S. Carter // Online Near-Duplicate Detection of News Articles. — 2020. — URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.156.pdf. (accessed: 03.01.26)

7. Charikar M. Similarity estimation techniques from rounding algorithms / M. Charikar // Proceedings of STOC. — 2004. — URL: https://www.cs.princeton.edu/courses/archive/spr04/cos598B/bib/CharikarEstim.pdf. (accessed: 07.01.26)

8. Manku G. Detecting Near-Duplicates for Web Crawling / G. Manku, A. Jain // Proceedings of WWW. — 2007. — URL: https://research.google.com/pubs/archive/33026.pdf. (accessed: 09.01.26)

9. Yalcin K. An external plagiarism detection system based on part-of-speech (POS) tag n-grams and word embedding / K. Yalcin, N. Aydin // Expert Systems with Applications. — 2022. — №197. — URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417422001610. (accessed: 12.01.26)

10. Amirzhanov A. Plagiarism types and detection methods: a systematic review / A. Amirzhanov // Frontiers in Computer Science. — 2025. — URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1504725/pdf. (accessed: 14.01.26)

11. Krasnov F.V. Problema poteri reshenij v zadache poiska sxozhix dokumentov: primenenie terminologii pri postroenii vektornoj modeli korpusa [The problem of losing solutions in the task of finding similar documents: using terminology to build a vector model of the corpus] / F.V. Krasnov // CyberLeninka. — 2021. — URL: https://cyberleninka.ru/article/n/problema-poteri-resheniy-v-zadache-poiska-shozhih-dokumentov-primenenie-terminologii-pri-postroenii-vektornoy-modeli-korpusa. (accessed: 15.01.26) [in Russian]

12. Kuznetsova R.V. Metodi obnaruzheniya perevodnikh zaimstvovanii v bolshikh tekstovikh kollektsiyakh [Methods for detecting translated borrowings in large text collections] / R.V. Kuznetsova, O.Yu. Bakhteev, Yu.V. Chekhovich // researchgate. — 2021. — URL: https://www.researchgate.net/publication/354247949_METODY_OBNARUZENIA_PEREVODNYH_ZAIMSTVOVANIJ_V_BOLSIH_TEKSTOVYH_KOLLEKCIAH (accessed: 18.01.26) [in Russian]