Надежность, том 26, №1, 2026 / Dependability 1, 2026
Вопросы автоматизации и управления процессами на транспорте

УДК 656.2 : 004.912
Оригинальная статья
https://doi.org/10.21683/1729-2646-2026-26-1-12-20

Text mining журнала ДУ-46 с применением частотных методов обработки текста

Text mining of the DU-46 inspection log using frequency-based text processing methods

Канарский В.А.[1]*
Kanarsky V.A.[1]

[1] *Дальневосточный государственный университет путей сообщения, Российская Федерация, Хабаровск*
[1] *Far Eastern State Transport University, Russian Federation, Khabarovsk*
* *vkanarsky@ro.ru*

Канарский В.А.

Резюме. *Для принятия обоснованных решений по устранению сбоев и неисправностей, возникших на объектах железнодорожной инфраструктуры, необходим оперативный доступ к информации о ранее выявленных нарушениях и динамике их устранения. Технические журналы, такие как ДУ-46, содержат ценные сведения о состоянии этих объектов (путей, стрелочных переводов, светофоров, фидеров, контактной сети и др.), однако в моменте они практически не используются при анализе причин вновь возникающих отказов.* **Цель***: разработка алгоритма обработки записей журнала ДУ-46, позволяющего по запросу оператора получать сведения о предыдущих неисправностях или выполненных работах по конкретным объектам инфраструктуры.* **Методы***: предобработка текста, лемматизация с использованием морфологического анализатора М. Коробова, частотный анализ текста, TF-IDF, L2-нормализация, вычисление косинусного расстояния и сортировка результатов.* **Результат:** *создан прототип приложения, позволяющий осуществлять поиск релевантных записей и отображать метрику сходства между запросом и найденными фрагментами, которые помимо всего прочего могут носить рекомендательный характер и использоваться для установления причин возникших сбоев.* **Заключение:** *использование оперативных журналов осмотра в сочетании с методами интеллектуального анализа текста (text mining) может стать основой для построения рекомендательных систем и систем поддержки принятия решений в техническом обслуживании объектов ж.д. инфраструктуры.*

Abstract. *For making well-founded decisions on the elimination of failures and malfunctions occurring in railway infrastructure facilities, prompt access to information on previously identified faults and the dynamics of their resolution is essential. Inspection logs such as DU-46 contain valuable data on the condition of these facilities (tracks, turnouts, signals, power supply, contact lines, etc.); however, they are hardly used in practice when analyzing the causes of newly emerging failures.* **Aim.** *To develop an algorithm for processing DU-46 log records that allows operators, upon request, to obtain information on previous malfunctions or maintenance activities on specific infrastructure objects.* **Methods.** *Text preprocessing, lemmatization using M. Korobov's morphological analyzer, frequency-based text analysis, TF–IDF, L2 normalization, cosine similarity calculation, and result sorting.* **Result.** *A prototype application has been developed that enables search for relevant records and displays a similarity metric between the query and the retrieved fragments, which, among other things, may serve as a recommendatory function for determining the causes of failures.* **Conclusion.** *The use of operational inspection logs in combination with text mining methods can form the basis for building recommendation systems and decision support systems in the maintenance of railway infrastructure facilities.*

Ключевые слова: *железнодорожный транспорт, журнал, ДУ-46, текст-майнинг, векторное представление, TF-IDF, косинусное расстояние, интеллектуальный поиск*
Keywords: *railway transport, inspection log, DU-46, text mining, vector representation, TF-IDF, cosine distance, intelligent search*

Для цитирования: *Канарский В.А. Text mining журнала ДУ-46 с применением частотных методов обработки текста // Надежность. 2026. №1 С. 12-20. https://doi.org/10.21683/1729-2646-2026-26-1-12-20*
For citation: *Kanarsky, V.A. Text mining of the DU-46 inspection log using frequency-based text processing methods. Dependability 2026;1: 12-20. https://doi.org/10.21683/1729-2646-2026-26-1-12-20*

Поступила: *26.08.2025* / **После доработки:** *29.09.2025* / **К печати:** *01.02.2026*
Received on: *26.08.2025* / **Revised on:** *29.09.2025* / **For printing:** *01.02.2026*

Введение

Для достижения целевого состояния системы обеспечения гарантированной безопасности перевозочного процесса ОАО «РЖД» выделяет 3 основных потенциальных области риска, к которым относятся:

– нормативная и техническая документация;

– состояние технических средств;

– технологическая дисциплина.

О надежности обеспечения перевозочного процесса, в первую очередь, судят на основании информации о состоянии технических средств, которые формируются на основании данных учета отказов технических средств и *контроля за их устранением*, а также *учета результатов осмотров* объектов инфраструктуры и подвижного состава. Технология сбора информации по отказам технических средств должна максимальным образом базироваться на данных объективных источников информации – показаниях систем мониторинга, результатах расшифровки скоростемерных лент, архивах микропроцессорных систем управления и систем обеспечения безопасности движения (ДЦ, ДК), данных вагонов-путеизмерителей и других мобильных средств измерений [1].

Современные производственные системы все шире внедряют подходы *предиктивной аналитики*, позволяющие выявлять неисправности, прогнозировать отказы и оценивать состояние оборудования. Так Надежкин В.А. рассматривает применение предиктивной аналитики к рельсовым цепям и показывает, что такой подход позволяет выявлять скрытые аномалии и формировать систему раннего предупреждения отказов, недоступную традиционным комплексам диагностики [2]. Минтус А.Н. описывает структуру подобной системы для устройств сигнализации, централизации и блокировки (СЦБ), включающую модули сбора, обработки и прогнозирования данных [3]

Все эти решения положительным образом сказываются на проведении мероприятий по техническому обслуживанию (ТО). Несмотря на эти разработки, большинство подходов в этой области опирается на числовые данные от датчиков и полевых устройств для проведения анализа.

Бушуев С.В. подчеркивает, что современные системы диспетчерской и электрической централизации содержат подсистемы архивирования, в которых хранятся оперативная поездная обстановка и приказы операторов. Но такие данные используются только при разборе нештатных ситуаций – системного анализа архивной информации для поиска скрытых закономерностей и извлечения знаний, полезных для оптимизации процесса перевозок нет [4].

Андронов И.А., Сидоренко В.Г. описывают тенденции в автоматизации управленческого документооборота. Среди предложений отмечается применение *интеллектуальных механизмов категоризации и поиска.* Также авторы статьи отмечают, что большинство современных систем документооборота ограничиваются базовыми функциями маршрутизации и хранения документов, в то время как потребности управленцев смещаются в сторону поддержки принятия решений [5].

Для обоснованного и своевременного упреждающего воздействия необходимо владеть информацией в реальном масштабе времени для прослеживания динамики изменения ситуации, причем не только о выявленных нарушениях, но и о динамике их устранения. Текстовые данные из оперативных журналов ТО содержат одни из самых важных сведений о функционировании систем и компонентов, но до сих пор рассматриваются как «черные дыры», то есть хранят ценные данные, неиспользуемые при принятии решений [6].

На железнодорожных станциях существует целый комплекс форм оперативного внутреннего учета, в котором фиксируются любые события, происходящие на путях, в посту электрической централизации, с устройствами сигнализации, централизации и блокировки (светофоры, стрелочные переводы, рельсовые цепи и т.п.), энергоснабжением, контактной сетью, записываются временя отправления и прибытия и т.п. Такие документы являются источником ценной информации, объясняющей снижение эффективности поездной работы, а также причины нарушения безопасности движения поездов.

Журнал ДУ-46

Одной из важнейших форм внутреннего учета на станциях является журнал осмотра путей, стрелочных переводов, устройств СЦБ, связи и контактной сети – журнал ДУ-46. Состояние этих объектов железнодорожной инфраструктуры, как правило, контролируется вручную в ходе плановых проверок и предупредительных мероприятий, сведения о результатах, внезапные отказы или неисправности фиксируются в указанном журнале. Записи в ДУ-46 ведутся под строгим контролем и безошибочно, причем сам журнал является важнейшим документом и в случае экстренных ситуаций служит доказательством причастности того или иного работника к происшествию, связанному с эксплуатацией железнодорожного транспорта в целом.

Таким образом, журнал осмотра ДУ-46 предназначен:

– для записей в него обо всех возникающих на станции и прилегающих к станции перегонах неисправностей устройств СЦБ, пути, контактной сети и других в хронологическом порядке;

– для записей о результатах периодических осмотров устройств на станции;

– для регистрации в нем в хронологическом порядке всех выполняемых на станции работ по техническому обслуживанию, устранению неисправностей, ремонту, строительству, реконструкции устройств СЦБ, пути и контактной сети;

– для записей в обязательном порядке о производстве путевых работ на станционных путях и перегонах вблизи станции (1-х участков удаления), требующих огражде-

ния сигналами остановки или уменьшения скорости, о снятии напряжения и подаче его в контактную сеть станции, об открытии и закрытии движения по путям и стрелочным переводам и др.;

– для записей о производстве работ, связанных с выключением устройств СЦБ из централизации [1].

Как следует из рис. 1, заполняемыми полями являются дата и время проводимого мероприятия, описание запланированной работы либо обнаруженной неисправности, сроки окончания/устранения, принятые меры, а также фамилии сотрудников и их подписи.

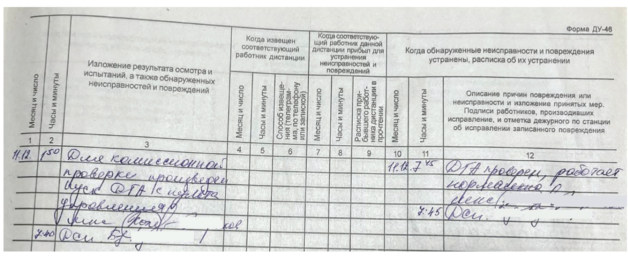


Рис. 1. Пример записи из журнале ДУ-46

В целом записи в журнале ДУ-46 имеют обобщенные формулировки, представленные в Приложении № 5 к Инструкции по обеспечению безопасности движения поездов при технической эксплуатации устройств и систем СЦБ [7]. Однако исходя из местных условий и эксплуатируемых устройств, эти формулировки дополняются конкретными номера стрелок, наименования участков пути или светофоров.

В настоящей статье предлагается реализовать модуль рекомендаций на основе архивных журналов ДУ-46 с некоторой станции «А» (262 записи) и некоторой станции «Б» (124 записи). Наименование станций не раскрываются по коммерческим причинам.

Первичная обработка фрагментов ДУ-46

В ходе первичного визуального анализа журналов были выявлены следующие текстовые особенности:

– дополнительное указание времени в часах и минутах в столбце 3, вдобавок к указанному в столбцу 2 времени (в др. записях как на рис. 1);

– изобилие сокращений (ДСП, ДГА, ИРМ, ПОБ, ПСГО, РЦНС, КТСМ и т.п.);

– наличие узкоспециализированной терминологии («пульт-табло», «маневровый», «путеизмерительная», «курбель»);

– изобилие числовых описаний разделов иных документов, номеров стрелок, поездов и т.п.;

– изобилие пунктуации: кавычки, точки, символы «№» и «/»;

– наличие имен собственных.

Таким образом, анализируемый корпус текста нуждается в серьезной предварительной обработке (рис. 2):

– приведении слов к одному регистру;

– удалении специальных символов и знаков препинания;

– разделении (токенизации) корпуса текста на отдельные слова (токены);

– выявление среди токенов предлогов, союзов, имен собственных и их последующее удаление;

– приведение оставшихся слов к единой форме.

Слова «светофора» и «светофоров» обозначают один и тот же объект, но без приведения к одной форме они будут считаться различными словами. Существует два подхода к данной задачи – это стемминг и лемматизация.

Алгоритмы стемминга позволяет получить корень слова урезая суффиксы и окончания и является наиболее легковесным вариантом. Лемматизация – это более существенная процедура анализа, которая в большинстве случаев позволяет получить начальную форму слова (лемму): «перегорел», «перегорела» – начальная форма: «перегореть». Подход с использованием лемматизации вычислительно более затратный, однако при этом предпочтительнее из-за со сложной морфологии русского языка [8], а также из-за обилия специализированных терминов журнала ДУ-46.

Для нахождения леммы был выбран морфологический анализатор pymorphy3 Михаила Коробова [9]. На рис. 2 продемонстрирован процесс первичной обработки одной записи:

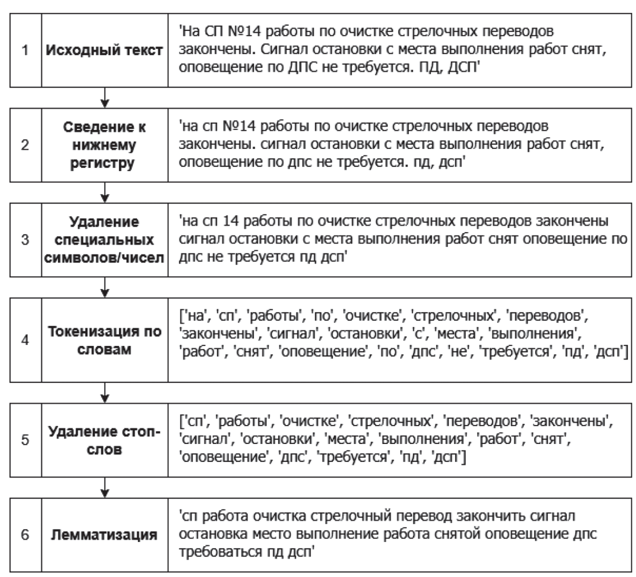
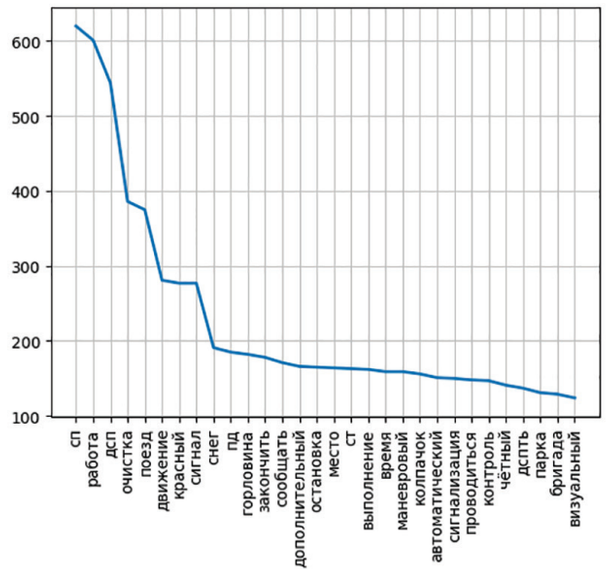


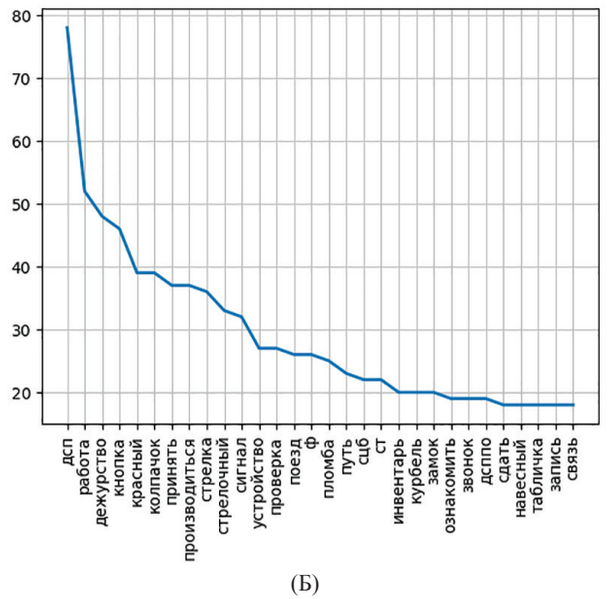
Рис. 2. Процесс обработки текста одной из записей в ДУ-46

После приведения слов к единой форме имеется возможность выполнить частотный анализ по журналам А и Б (рис. 3 и 4).

Согласно рис. 3, наиболее часто встречающимися словами являются «ДСП», «СП», «работа» и «очистка». Сокращение «ДСП» присутствует во всех записях журнала, поскольку каждая из них заверяется дежурным по станции, находившимся на смене в момент события. Общеупотребляемым также является слово



(А)



(Б)

Рис.3. Частотный анализ слов записей журнала «А» и «Б»

«работа», тогда как термины «снег» и «очистка» преимущественно связаны с зимним периодом ведения журнала. Таким образом, в рамках данного исследования указанные слова были отнесены к стоп-словам, поскольку они не отражают уникальных закономерностей в рассматриваемом корпусе записей. Вместе с тем при анализе данных, охватывающих более протяжённые временные интервалы (например, за год), целесообразно учитывать формулировки, связанные с сезонными видами работ.

Проводя более детальный анализ корпусов записей А и Б без учета стоп-слов были выявлены следующие показатели (табл. 1).

Анализ табл. 1 показывает, что распределение частей речи в обоих журналах имеет схожий характер. Однако в журнале А доля уникальных слов составляет лишь 4,4%, что свидетельствует о частом повторении одних и тех же формулировок и, как следствие, об однотипности выполняемых работ. При ограничении объема журнала А до 2000 слов доля уникальной лексики составила бы около 7%, что почти в три раза меньше, чем в журнале Б. Последний, напротив, демонстрирует более высокую вариативность словаря, что отражает разнообразие описываемых случаев и ситуаций на станции.

Тем не менее, для дальнейшего этапа исследования записи обоих журналов были агрегированы в один набор данных.

Унитарное кодирование и отбор признаков

Для обработки текста необходимо преобразовать его слова в численное представление. Одним из наиболее простых и распространенных подходов является использование унитарного (от англ. *one-hot encoding*) кодирования. После проведения лемматизации формируется словарь S, включающий все слова получившегося корпуса. В рамках унитарного кодирования каждое слово представляется вектором длиной, равной размеру словаря, где единичное значение занимает позицию, соответствующую данному слову (рис. 4).

$$\text{бригада} = [1\ 0\ 0\ 0\ 0\ 0\ \ldots\ 0],$$

$$\text{видимость} = [0\ 1\ 0\ 0\ 0\ 0\ \ldots\ 0].$$

Рис. 4. Пример унитарного представления слов

Таким образом, каждая запись журнала может быть представлена в виде «мешка слов» (англ. bag of words) – вектора, элементы которого соответствуют словам из словаря. Значения элементов вектора отражают количество вхождений соответствующего слова в данную запись. На примере, представленном на рис. 5, это можно проиллюстрировать следующим образом.

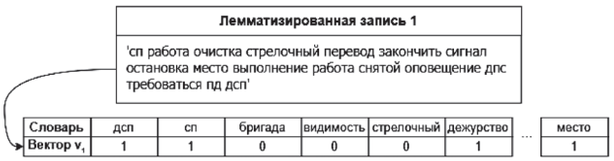


Рис. 5. Процесс преобразования записи 1 из ДУ-46 в вектор $\mathbf{v}_1$

Табл. 1. Морфологический анализ текста из журналов ДУ-46

Журнал	Всего слов	% уникальных слов	Распределение по частям речи, %		
			Существит.	Прилагател.	Глаголов
А	12860	4,4	65,74	18,31	15,95
Б	2034	20	66,8	16,62	16,56

Недостатком такого подхода (рис. 5) является высокая размерность векторов **v**, равная размеру словаря S. В результате формируется значительное количество нулевых значений, так как каждая отдельная запись содержит гораздо меньше слов, чем весь словарь, и векторы получаются разреженными. Это особенно заметно для редких слов, встречающихся лишь однажды. Однако в некоторых задачах сохранение таких слов в векторах может быть полезным, поскольку они позволяют фиксировать уникальные совпадения между записями. Наиболее часто употребляемые слова (как отмечалось ранее), такие как «ДСП», «СП», «работа», «очистка», были внесены в список *стоп-слов*.

На этапе очистки 3 (рис. 2) при построении вектора числовые значения, которыми изобилуют записи ДУ-46, были проигнорированы. Хотя в целом такие данные в обобщенном (*агрегированном*) виде могут нести полезную информацию и тем самым способствовать более корректной оценке близости текстов. Поэтому перед проведением вычислительных экспериментов в векторные представления был добавлен еще один признак – количество числовых значений, упомянутых в записи (номера стрелок, светофоров, фидеров, время, количество курбелей и т.п.) (переменная n_number) (рис. 6).

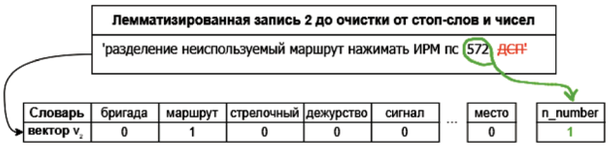


Рис. 6. Иллюстрация уточнения вектора признаком подсчета чисел

Векторизация записей ДУ-46 методом TF-IDF

Как уже отмечалось записи ДУ-46 основываются на обобщенных формулировках [6] и отличаются друг от друга некоторыми ключевыми словами. Для такого формата текста не очень подходит простой подсчет слов методом мешка слов. Для ключевых слов можно установить их важность с помощью *частоты термина и обратной частоты документа* (от англ. Term Frequency-Inverse Document Frequency, TF-IDF) [10].

Частота использования термина t в документе d определяется простым соотношением:

$$TF(t,d) = \frac{n_{t\in d}}{N_d}, \qquad (1)$$

где $n_{t\in d}$ – количество термина t в документе d;

N_d – количество всех слов в документе d.

Обратная частота документа позволяет оценить вес (важность) термина t исходя из всего набора документов D:

$$IDF(t,D) = \log\frac{N_{t\in D}}{count\left(d_i \in D \text{ если } t \in d_i\right)}, \qquad (2)$$

где $N_{t\in D}$ – количество документов, в которых встречается термин t;

count(…) – функция, которая возвращает количество документов d, в которых есть термин t.

Таким образом, уникальность термина t в документе с поправкой на его «встречаемость» во всем наборе записей D можно выразить следующим образом:

$$TF - IDF = TF \times IDF.^{1} \qquad (3)$$

Оценим важность термина «стрелочный» в документе (рис. 6) без стоп-слов:

$$TF\left(<\text{стрелочный}>, \text{рис.}1\right) = \frac{1}{13};$$

$$IDF\left(<\text{стрелочный}>, \text{ДУ-46}\right) = \log\frac{400}{46} \approx 2{,}16;$$

$$TF - IDF\left(<\text{стрелочный}>, \text{рис.}1\right) = \frac{1}{13}\cdot 1{,}85 \approx 0{,}166.$$

Следуя такому алгоритму, термин «сигнал» из этой же записи, встречающийся 255 раз во всем имеющимся фрагменте ДУ-46 будет иметь вес 0,034. Получается в тексте рис. 5 слово «стрелочный» в 5 раз важнее слова «сигнал».

Таким образом, выполнив TF-IDF (по реализации, приведенной в [11]) для всех слов каждой из записей ДУ-46, получим более релевантные векторные представления (рис. 7). В разделе 3 мы также условились обогатить эти векторы дополнительным признаком n_number. В процессе подсчета числовых значений выяснилось, что максимальное количество чисел, встречающихся в одной заметке журнала, составляет 13, а минимальное – 0.

В завершение процедуры можно выполнить нормализацию всех строк таблицы записей, что позволит привести каждый вектор к единичной длине. Это достигается делением ненормализованного вектора записи на его евклидову норму (4):

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2} = \frac{\mathbf{v}}{\sqrt{v_1^2 + v_2^2 + .. + v_n^2 + v_{n\_number}^2}} = \frac{\mathbf{v}}{\sqrt{\sum_{i=1}^{n} v_i^2}}, \qquad (4)$$

где **v** – вектор одной записи ДУ-46;

$\|\ \|_2$ – обозначение евклидовой нормы (L2) вектора;

n – максимальное кол-во слов, содержащихся в словаре;

v_1, v_2 … v_n – компоненты вектора, относящиеся к словам из словаря, рассчитанные по методу TF-IDF;

v_{n_number} – подсчитанное количество неизвестных слов и чисел в записи.

[1] Стоит уточнить, что окончательная реализация TF-IDF в вычислительном пакете [11] отличается от каноничной формулы (2) добавлением слагаемых «+1». Это делается для обеспечения численной стабильности предотвращает деление на 0 для терминов, которые не встретились ни в одном документе. Подробнее [11]

Лемматизирования запись 1 с подсчетом числовых значений

'~~сп 14 работа очистка~~ стрелочный перевод закончить сигнал остановка место выполнение работа снятой оповещение дпс требоваться пд ~~дсп~~'

	бригада	маршрут	стрелочный	перевод	дежурство	сигнал	...	место	n_number
запись 1	0	0	0.31	0,19	0	0.015		0,26	0,082
запись 2	0	0.45	0	0		0		0	0,124
....									
запись n	0	0	0	0	0.12	0.024		0	0

Рис. 7. Таблица нормализованных векторизованных записей ДУ-46

Таким образом, мы обеспечиваем одинаковую длину векторов **v** данных (строк) при сохранении их направления. Данное решение имеет особую ценность для текстов, потому что нас интересует больше схожесть записей по содержанию, а не по их по длине[1].

Алгоритм поиска релевантных записей

Получив векторное представление записей ДУ-46, можно организовать поиск интересуемых событий, происшествий или выполненных работ по следующему алгоритму (рис. 8):

Производя поиск релевантных фрагментов по ДУ-46 запрос q, подаваемый пользователем, должен пройти те же этапы обработки, что и исходный набор данных – сначала выполнить очистку текста с использованием уже сформированного списка стоп-слов, затем преобразовать запрос в векторное представление на основе существующего словаря и нормализовать его по формуле (4).

[1] Подробнее в техническом блоге https://blog.milvus.io/ai-quick-reference/how-does-vector-normalization-affect-embeddings?utm_source=chatgpt.com

Для выявления фрагментов записей, наиболее близких по содержанию к запросу q, рассчитывается косинусное расстояние между вектором запроса **q** и векторами записей $\mathbf{v_i}$ из предварительной сохраненной матрицы (рис. 7):

$$\text{cosine distance} = 1 - \frac{\mathbf{q} \cdot \mathbf{v}_i}{\|\mathbf{q}\| \cdot \|\mathbf{v}_i\|}. \quad (5)$$

С учетом заблаговременно проведенной нормализации выражение (5) можно упростить до:

$$\text{cosine distance} = 1 - \mathbf{q} \cdot \mathbf{v}_i. \quad (6)$$

После получения списка косинусных расстояний, рассчитанных от векторного представления запроса **q** до каждого векторного представления предыдущей записи $\mathbf{v}_i$ ДУ-46, выполняется сортировка данного списка по возрастанию. Таким образом можно выбрать первые наименьшие косинусные расстояния, полученные от тех записей, которые наиболее релевантны запросу (рис. 8).

Заинтересованным читателям статьи подобный алгоритм может показаться излишним усложнением, ведь в любых текстовых редакторах уже имеются встроенные функции поиска по словам. Действительно, в строку поиска достаточно ввести основу ключевого слова, например «освещ», чтобы найти связанные заметки. Однако в реальности при работе с журналом ДУ-46 количество записей только за один год на одной станции может достигать нескольких сотен, и даже тысяч. В таких условиях оператору приходится просматривать большое число фрагментов, содержащих нужную основу, но не

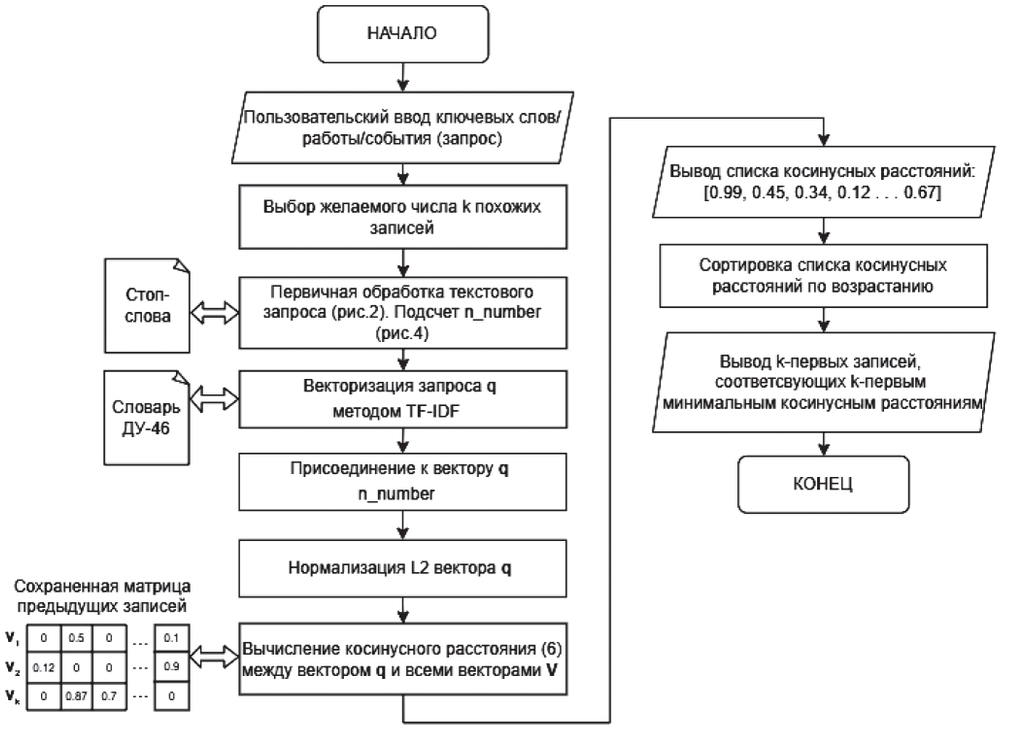
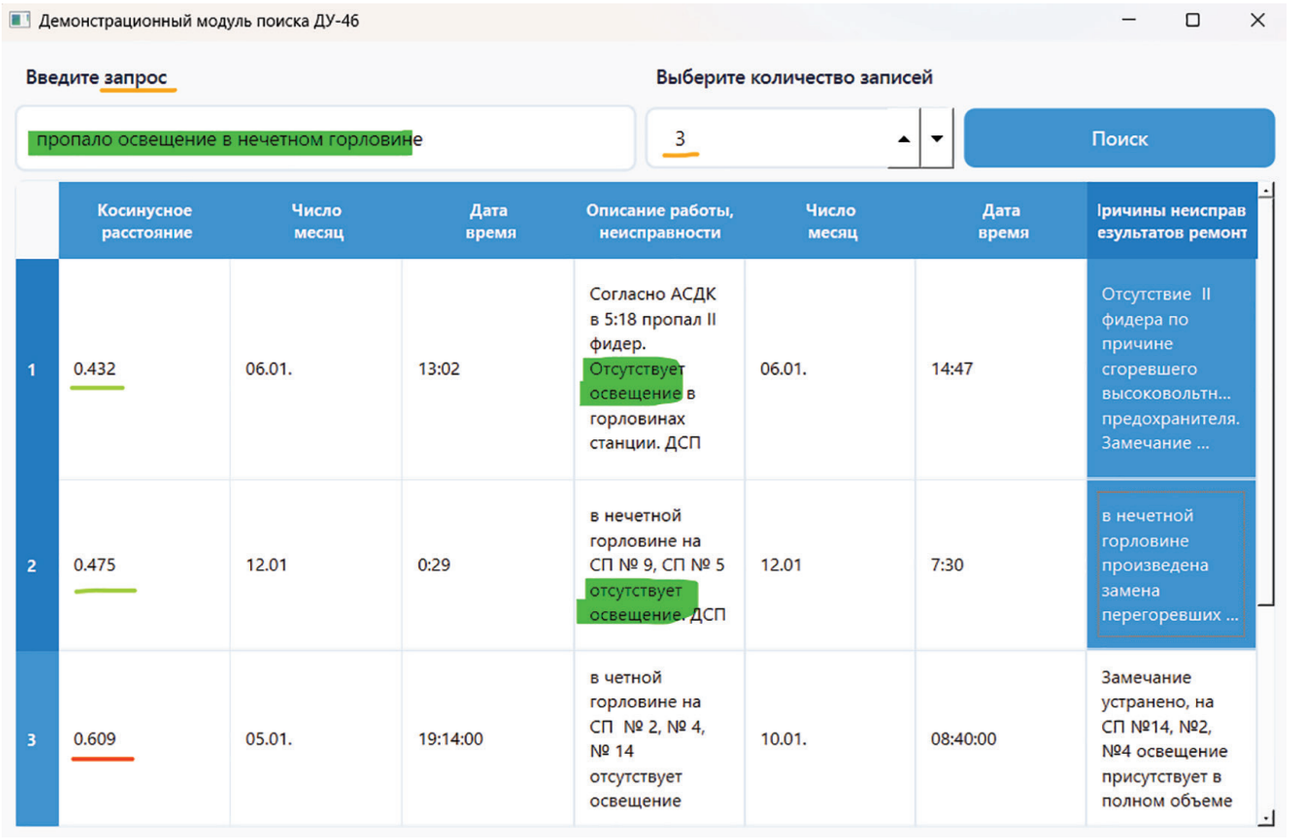


Рис. 8. Алгоритм поиска релевантных записей по запросу



	Косинусное расстояние	Число месяц	Дата время	Описание работы, неисправности	Число месяц	Дата время	Іричины неисправ езультатов ремонт
1	0.432	06.01.	13:02	Согласно АСДК в 5:18 пропал II фидер. Отсутствует освещение в горловинах станции. ДСП	06.01.	14:47	Отсутствие II фидера по причине сгоревшего высоковольтн... предохранителя. Замечание ...
2	0.475	12.01	0:29	в нечетной горловине на СП № 9, СП № 5 отсутствует освещение. ДСП	12.01	7:30	в нечетной горловине произведена замена перегоревших ...
3	0.609	05.01.	19:14:00	в четной горловине на СП № 2, № 4, № 14 отсутствует освещение	10.01.	08:40:00	Замечание устранено, на СП №14, №2, №4 освещение присутствует в полном объеме

Рис. 9. Демонстрационное приложение на основе ДУ-46 и алгоритма рис. 8

всегда относящихся к искомой ситуации. Это делает простой поиск малоэффективным.

Предложенный алгоритм поиска релевантных записей (рис. 8) опирается не только на наличие отдельных слов, но и на совпадение набора слов из запроса с текстами в журнале. Это позволяет отбирать записи, которые содержат все ключевые термины запроса и, таким образом, являются более близкими по содержанию. Его преимущество особенно заметно при работе с большими объемами данных (несколько тыс. заметок), когда формируется еще более представительный словарь с учетом сложных терминов, технического жаргона и отраслевых сокращений. Важным является то, что алгоритм опирается именно на *историю поломок конкретной станции*, что делает рекомендации максимально прикладными: оператор (электромеханик) получает ответы, основанные не на абстрактных примерах, а на уже происходивших событиях в данной инфраструктуре.

Даже в рамках данного исследования, основанного на 400 записях, алгоритм показал эффективность. На рис. 9 приведен пример: в левой колонке выводится косинусное расстояние (чем меньше число, тем ближе найденный фрагмент), в центральной части отображаются описания происшествий, а в правой – соответствующие решения.

Таким образом, оператор (электромеханик) получает не просто совпадения по словам, а *практическую рекомендацию, извлеченную из реального опыта* устранения неисправностей на его станции.

Заключение

Учитывая современные тенденции цифровой трансформации транспортной отрасли, внедрение интеллектуальных алгоритмов обработки документации (text mining) на станциях может повысить готовность технического персонала к устранению неисправностей и восстановлению работоспособности объектов железнодорожной инфраструктуры. Даже простейшие алгоритмы, основанные на обработке записей о предыдущих сбоях, позволяют формировать полезные подсказки для специалистов. Создание же полноценной системы поддержки принятия решений требует применения более сложных агентных ИИ-систем, которые, однако, нуждаются в значительных вычислительных ресурсах и подвержены трудностям при интерпретации нестандартных ситуаций, включая риск генерации некорректных ответов.

Вместе с тем предложенный подход имеет ряд ограничений. Наибольшую сложность представляет работа с ошибками и опечатками в текстах: такие слова попадают в словарь в искаженном виде и могут не находиться при поиске по запросу, что снижает качество вычислений сходства. В дальнейшем исследования будут направлены на расширение корпуса данных за счет большего количества записей из журналов ДУ-46, а также на применении улучшенной модели векторизации текста FastText. Данная технология особенно интересна, так как благодаря использованию n-грамм позволяет учитывать близость

слов даже при их неправильном написании, при этом всё ещё оставаясь вычислительно легкой в сравнении с современными большими языковыми моделями.

Особую ценность для будущих исследований представляет подключение других форм документации, например журнала учета выполненных работ на объектах СЦБ и связи (ШУ-2), что позволит формировать более полное представление о техническом состоянии инфраструктуры и создать более «сильную» рекомендательную систему.

Следует отметить, что представленная тема исследования отвечает положениям о цифровой трансформации транспортной отрасли до 2030 г. [12] в части развития цифрового электронного документооборота. Однако развитие данного направления именно в железнодорожной отрасли во многом сдерживается традиционной практикой ведения оперативных журналов в бумажном виде на многих участках железных дорог. При этом уже существуют примеры применения электронных форм, в частности, использование журнала ДУ-46 в составе системы ЕКАСУИ (Единая корпоративная автоматизированная система управления инфраструктурой) [13].

Список литературы

1. Информационные технологии на железнодорожном транспорте : учебное пособие : в 3 частях / Л.И. Папировская, Д.Н. Франтасов, Е.А. Часовских, М.Н. Липатова. Самара: СамГУПС, 2020. Часть 2: Информационные технологии в системе обеспечения движения поездов. 2020. 156 с.

2. Надежкин В.А. О возможности применения технологий искусственного интеллекта для определения и прогнозирования технического состояния устройств железнодорожной автоматики и телемеханики / В.А. Надежкин, С.А. Надежкина, А.Р. Мусин // Известия Петербургского университета путей сообщения. 2025. Т. 22. № 2. С. 484-491. DOI: 10.20295/1815-588X-2025-2-484-491 EDN: XONKFC

3. Минтус А.Н. Разработка системы предиктивной аналитики для технического обслуживания устройств сигнализации, централизации и блокировки / А.Н. Минтус, С.И. Кучеренко, И.Г. Герасина // Сборник научных трудов Донецкого института железнодорожного транспорта. 2025. № 1(76). С. 59-64. EDN: TBRCBY

4. Анализ загрузки путевого развития станции (по данным архивов систем централизаций стрелок и сигналов) / С.В. Бушуев, Б.В. Рожкин, А.А. Блюдов, Н.С. Голочалов // Вестник Уральского государственного университета путей сообщения. 2021. № 2(50). С. 30-44. DOI: 10.20291/2079-0392-2021-2-30-44 EDN: XAKKGM

5. Андронов И.А. Анализ современных подходов к построению систем электронного документооборота / И.А. Андронов, В.Г. Сидоренко // Интеллектуальные транспортные системы : Материалы IV Международной научно-практической конференции, Москва, 22 мая 2025 года. Москва: Российский университет транспорта (МИИТ), 2025. С. 42-48. DOI: 10.30932/9785002587582-2025-42-48 EDN: EWUCGV

6. Sundaram S., Zeid A. Technical language processing for Prognostics and Health Management: applying text similarity and topic modeling to maintenance work orders // Journal of Intelligent Manufacturing. 2024. Vol. 36. Pp. 1637–1657. DOI: 10.1007/s10845-024-02323-4

7. Инструкция по обеспечению безопасности движения поездов при производстве путевых работ : инструкция ОАО «РЖД» № ЦШ-530 : утв. ОАО «РЖД» 20.09.2011 г. № 2055р. (ред. от 18.09.2020). URL: https://itt-54.ru/wp-content/uploads/2023/12/Инструкция-ОАО-РЖД-Инструкция-по-обеспечению-БД-N-ЦШ-530-11от-20.09.-2011-г.-№-2055р.pdf (дата обращения: 22.08.2025).

8. Русаков А.М. Тестирование программных библиотек лемматизации текстов на Python / А.М. Русаков, П.А. Полянская // Наукосфера. 2023. № 11-1. С. 210-218. DOI: 10.5281/zenodo.10136822 EDN: EJFGZF

9. Коробов М. и др. Морфологический анализатор и генератор для русского и украинского языков // Материалы Международной конференции «Анализ изображений, социальных сетей и текстов» (International Conference on Analysis of Images, Social Networks and Texts) / Communications in Computer and Information Science. 2015. С. 320–332. DOI: 10.1007/978-3-319-26123-2_31

10. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing // Communications of the ACM. 1975. Vol. 18. No. 11. Pp. 613–620. DOI:10.1145/361219.361220

11. TfidfTransformer [Электронный ресурс] // scikit-learn developers. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html. (дата обращения: 23.08.2025).

12. Об утверждении стратегического направления в области цифровой трансформации транспортной отрасли РФ до 2030 г.: распоряжение Правительства РФ от 03.11.2023 г. № 3097р // Министерство транспорта Российской федерации. URL: https://mintrans.gov.ru/documents/2/12953 (дата обращения: 26.08.2025).

13. Зудина А. Еще одна безбумажная технология. // Гудок : газета «Волжская магистраль». 17.03.2023. № 9. URL: https://gudok.ru/zdr/168/?page_print_=Y&ID=1629697 (дата обращения 24.08.2025).

References

1. Papirovskaya L.I., Frantasov D.N., Chasovskikh E.A., Lipatova M.N. [Information technologies in railway transport: a textbook: in 3 parts. Part 2: Information technology in the train traffic management system]. Samara: SamGUPS; 2020. (in Russ.)

2. Nadezhkin V.A., Nadezhkina S.A., Musin A.R. Artificial Intelligence Technologies for Evaluation and Prediction of the Technical Condition of Railway Automation and Telemechanic Devices. Proceedings of Petersburg State Transport University, 2025, vol. 22, iss. 2, pp. 484–491. (In Russian) DOI: 10.20295/1815-588X-2025-2-484-491. (in Russ.)

3. Mintus A.N., Kucherenko S.I., Gerasina I.G. Development of a predictive analytics system for the technical maintenance of signaling, centralization, and blocking devices. *Sbornik nauchnykh trudov Donetskogo instituta zheleznodorozhnogo transporta* 2025;1(76):59-64. EDN: TBRCBY. (in Russ.)

4. Bushuev S.V., Rozhkin B.V., Bludov A.A., Golochalov N.S. Analysis of the station's track development load (according to the archives of railway switches and signal centralized systems). *Herald of the Ural State University of Railway Transport* 2021;2(50):30-44. (in Russ.)

5. Andronov I.A., Sidorenko V.G. Analysis of modern approaches to building electronic document management systems. In: Intelligent Transportation Systems: Proceedings of the IV International Research and Practice Conference. Moscow; May 22, 2025. Moscow: Russian University of Transport (MIIT); 2025. Pp. 42-48. (in Russ.) DOI: 10.30932/9785002587582-2025-42-48 EDN: EWUCGV

6. Sundaram S., Zeid A. Technical language processing for Prognostics and Health Management: applying text similarity and topic modeling to maintenance work orders. *Journal of Intelligent Manufacturing* 2024;36:1637–1657. DOI: 10.1007/s10845-024-02323-4.

7. [Manual for ensuring train traffic safety in the course of track work operations: JSC RZD manual no. TSH-530: approved by Order of JSC RZD dated 20.09.2011 no. 2055r. (as amended on 18.09.2020)]. (accessed: 22.08.2025). Available at: https://itt-54.ru/wp-content/uploads/2023/12/Инструкция-ОАО-РЖД-Инструкция-по-обеспечению-БД-N-ЦШ-530-11от-20.09.-2011-г.-№-2055р. pdf. (in Russ.)

8. Rusakov A.M., Polyanskaya P.A. Testing software libraries for lemmatization of texts in Python. *Naukosfera* 2023;11-1:210-218. DOI: 10.5281/zenodo.10136822 EDN: EJFGZF. (in Russ.)

9. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: International Conference on Analysis of Images, Social Networks and Texts / Communications in Computer and Information Science 2015;320–332. DOI: 10.1007/978-3-319-26123-2_31. (in Russ.)

10. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing. *Communications of the ACM* 1975;18(11):613–620. DOI:10.1145/361219.361220.

11. TfidfTransformer. scikit-learn developers. (accessed: 23.08.2025). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html.

12. [On the approval of the strategic direction in the field of digital transformation of the transportation industry of the Russian Federation to 2030: Decree of the Government of the Russian Federation dated 03.11.2023 no. 3097r. Ministry of Transport of the Russian Federation]. (accessed: 08/26/2025). Available at: https://mintrans.gov.ru/documents/2/12953. (in Russ.)

13. Zudina A. [Another paperless technology]. *Gudok: Volzhskaya magistral* 2023;9. (accessed 08/24/2025). Available at: https://gudok.ru/zdr/168/?page_print_=Y&ID=1629697. (in Russ.)

Сведения об авторе

Вадим Андреевич Канарский – кандидат технических наук, доцент, и.о. заведующего кафедрой «Автоматизированные, телекоммуникационные и электротехнические системы» в Дальневосточном государственном университете путей сообщения, Far Eastern State Transport Universuty; г. Хабаровск, Российская Федерация: e-mail: vkanarsky@ro.ru; SPIN-код 3411-0352

About the Author

Vadim A. Kanarsky, Candidate of Technical Sciences, Senior Lecturer, Acting Head of the Department of «Automated, Telecommunication and Electrical Engineering Systems», Far Eastern State Transport University; Khabarovsk, Russian Federation, e-mail: vkanarsky@ro.ru. SPIN-code: 3411-0352

Вклад автора в статью

В.А. Канарским были исследованы внутренние процессы учета работ и прошествий на объектах железнодорожного транспорта, проанализированы некоторые тенденции, характерные в сфере интеллектуальных транспортных систем и документооборота. Автором была собрана база документов из архивов действующих станций, разработан алгоритм, использующий эти документы как источник ценных знаний о состоянии объектов инфраструктуры. Вся необходимая предобработка, проводимые вычислительные эксперименты и создание прототипа приложения выполнены автором самостоятельно.

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов