

РАЗРАБОТКА БИБЛИОТЕКИ ДЛЯ АВТОМАТИЗАЦИИ ОЦЕНКИ КАЧЕСТВА ДАННЫХ

Макунина А.А. (ИТМО)

Научный руководитель – ассистент ФПИИ Чернышева А.В. (ИТМО)

Введение

Данные являются одним из основных факторов обеспечения конкурентоспособности и эффективного управления организацией. Поэтому поддержание качества данных [1] представляет собой не только полезную, но и необходимую практику. Неточности в хранилищах данных могут привести к негативным последствиям, таким как денежные потери, потеря доверия клиентов и сбои в бизнес-процессах [2, 3].

Цель данной работы – разработать библиотеку для комплексной проверки данных на соответствие критериям [4] качества данных, а также уведомления о найденных нарушениях и рекомендаций по их исправлению с использованием возможностей больших языковых моделей (LLM, Large Language Model).

Основная часть

Был проведен анализ существующих инструментов для проверки качества данных [5-11], в ходе которого выяснилось, что часть решений не предоставляет возможности по интерпретации обнаруженных аномалий, не интегрируется в желаемые процессы обработки данных или требует слишком сложных настроек. Некоторые инструменты сосредоточены лишь на базовых проверках, в то время как другие не имеют гибких механизмов уведомления о проблемах. Чтобы устранить эти недостатки, была спроектирована и разработана собственная библиотека.

Разработанная библиотека содержит набор проверок для реляционных и нереляционных источников, которая нацелена на высокую точность проверок, производительность и простоту интеграции пользователем. Для начала работы достаточно лишь заполнить конфигурационный файл или воспользоваться автоматической генерацией настроек, что обеспечивает быстрое внедрение в существующие процессы. Система запускается из командной строки короткими командами и содержит как базовые проверки, такие как уникальность значений и отсутствие лишних пробелов, так и продвинутые, включая обнаружение выбросов и анализ с использованием регулярных выражений. Система дополнительно отправляет уведомления о выявленных проблемах через Telegram и создает аналитические отчеты, отражающие все выявленные несоответствия. Библиотека может сформулировать предполагаемую причину ошибки и предложить варианты устранения при обнаружении неверных данных с использованием механизмов на основе LLM-модели gpt-4o от OpenAI.

Благодаря своей архитектуре библиотека легко интегрируется в CI/CD и ETL/ELT-процессы, обеспечивает масштабирование и дает возможность адаптации под конкретные бизнес-процессы.

Выводы

Была разработана и протестирована библиотека для автоматизированной оценки качества данных, демонстрирующая высокую эффективность в выявлении проблем с данными. Это решение упрощает процессы мониторинга данных за счет автоматической генерации уведомлений и подсказок с использованием большой языковой модели, что позволяет оперативно реагировать на проблемы с данными. Такой комплексный подход позволяет эффективно контролировать качество данных, что делает библиотеку актуальной и перспективной для использования в работе с хранилищами данных.

Список использованных источников

1. ГОСТ Р ИСО 8000-2—2019. Управление данными и информацией. Качество данных. Часть 2. Метаданные: концепции и принципиальные положения. – М.: Стандартинформ, 2019. – 22 с.
2. Иванов В. Н., Суворов А. В. Современные проблемы развития российского здравоохранения. Часть 2 [Электронный ресурс] / Проблемы прогнозирования. – 2022. – № 1 (190). – URL: <https://cyberleninka.ru/article/n/sovremennye-problemy-razvitiya-rossiyskogo-zdravoohraneniya-chast-2> (дата обращения: 08.02.2025).
3. 9 ошибок селлеров в работе на маркетплейсах [Электронный ресурс] / E-pepper. – URL: <https://e-pepper.ru/news/9-oshibok-sellerov-v-rabote-na-marketpleysakh.html> (дата обращения: 08.02.2025).
4. Критерии качества данных [Электронный ресурс] / Loginom. – URL: <https://loginom.ru/blog/data-quality-criteria> (дата обращения: 08.02.2025).
5. Great Expectations [Электронный ресурс] / Great Expectations. – URL: <https://greatexpectations.io/> (дата обращения: 08.02.2025).
6. Pandas Profiling [Электронный ресурс] / Pandas Profiling. – URL: <https://pandas-profiling.ydata.ai/> (дата обращения: 08.02.2025).
7. Deepchecks [Электронный ресурс] / Deepchecks. – URL: <https://deepchecks.com/> (дата обращения: 08.02.2025).
8. Evidently AI [Электронный ресурс] / Evidently AI. – URL: <https://evidentlyai.com/> (дата обращения: 08.02.2025).
9. YData [Электронный ресурс] / YData. – URL: <https://ydata.ai/> (дата обращения: 08.02.2025).
10. SodaSQL [Электронный ресурс] / SodaSQL. – URL: <https://soda.io/products/soda-sql> (дата обращения: 08.02.2025).
11. Data Ladder [Электронный ресурс] / Data Ladder. – URL: <https://dataladder.com/> (дата обращения: 08.02.2025).