УДК 004.043

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ И ОЦЕНКА ЕЁ ПРИМЕНИМОСТИ НА ИНДУСТРИАЛЬНЫХ КЕЙСАХ

Лазебный В. В.[1], Федоров Д. А.[1]
Научный руководитель – канд. техн. наук, доцент Федоров Д. А.[1]
[1]Университет ИТМО
vllazebnyi@gmail.com

Введение

Задача автоматизации сбора и структурирования информации из разрозненных источников остается актуальной на протяжении многих лет [1]. Традиционные методы парсинга часто неустойчивы к изменениям структуры веб-страниц и динамически изменяемых сайтов, что делает их уязвимыми и требовательными в постоянной поддержке. Альтернативные методы извлечения данных, например использование больших языковых моделей или IE на основе машинного обучения, дают более качественный результат, но требуют тщательной и долгой настройки, а также создают риски появления галлюцинаций и увеличения стоимости разрабатываемой системы [2, 3]. Так, ключевой задачей настоящей работы стала разработка автоматизированной системы сбора и обработки данных, которая позволила бы обеспечить высокий уровень надежности и устойчивости системы к изменениям. Более того, важным требованием стало снижение риска возникновения галлюцинаций и стоимости разработки.

Основная часть

В работе предложен гибридный метод обработки неструктурированных данных, который обеспечивает полный цикл работы: от сбора до последующей валидации. Он включает в себя следующие шаги:

1) Комбинированный сбор данных для парсинга из различных веб-источников, где в зависимости от типа источника определяется свой метод извлечения информации (динамический или статический). Этот шаг позволяет снизить чувствительность к изменениям.

2) Семантическая дедупликация с использованием эмбеддингов, где при превышении порога схожести рассматриваемый объект признается семантическим дубликатом и отфильтровывается. Это позволяет не только экономить вычислительные ресурсы LLM, но и выявлять ситуативно схожие документы (например, новости об одном событии).

3) Приведение данных к заданной JSON-структуре, состоящее из извлечения сущностей, их сопоставления с полями схемы и нормализации.

4) Выполнение нескольких запросов к большой языковой модели с разными параметрами температуры и сравнение их эмбеддингов. Данный шаг позволяет проверить ответ на наличие галлюцинаций, а последующая проверка ключей – на соответствие целевому JSON-формату.

Сочетание перечисленных шагов снижает долю ручной проверки и делает результат ближе к требованиям промышленной эксплуатации [3]. Новизна заключается в комбинации перечисленных подходов с последующим доказательством эффективности.

Выводы

В результате была разработана автоматизированная система для агрегации неструктурированных данных. Также в рамках тестирования предложенной технологии были проведены эксперименты, которые подтвердили точность и надежность алгоритма.

Литература

1. Boegershausen J., Datta H., Borah A., Stephen A. Fields of Gold: Scraping Web Data for Marketing Insights // Journal of Marketing. – 2022. – Vol. 86.
2. Juraj Vladika, Ihsan Soydemir, Florian Matthes. Correcting Hallucinations in News Summaries: Exploration of Self-Correcting LLM Methods with External Knowledge. – arXiv:2506.19607, 2025.
3. Remadi, A., Hage, K., Hobeika, Y., & Bugiotti, F. To prompt or not to prompt: Navigating the use of Large Language Models for integrating and modeling heterogeneous data // Data Knowl. Eng. – 2024. – Vol. 152.