УДК 004

Волосников Павел Дмитриевич,
студент,
Институт радиоэлектроники и информационных технологий,
ФГАОУ ВО «Уральский федеральный университет имени первого ПрезидентаРоссии Б.Н.Ельцина»
г. Екатеринбург, Российская Федерация

Сагилова Эйла Кайратовна,
студент,
Институт радиоэлектроники и информационных технологий,
ФГАОУ ВО «Уральский федеральный университет имени первого ПрезидентаРоссии Б.Н.Ельцина»
г. Екатеринбург, Российская Федерация

ОБОСНОВАНИЕ ПРИМЕНЕНИЯ LLM ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПОИСКА ДУБЛЕЙ ПРИ СОЗДАНИИ ИНСТРУМЕНТА НОРМАЛИЗАЦИИ ДАННЫХ НСИ

Аннотация:
В данной статье приведено обоснование применения LLM для поиска дублей в процессе разработки инструмента нормализации данных нормативно-справочной информации (НСИ). Проведен анализ существующих исследований, изучена литература по использованию LLM в контексте НСИ, приведены наукометрические показатели и обоснован приоритет применения LLM.

Ключевые слова:
Дубликаты, большие языковые модели, нормативно-справочная информация (НСИ), нормализация, обработка данных.

Введение

В условиях роста объемов данных задача поиска дубликатов становится критически важной, поскольку традиционные алгоритмы часто не справляются с разнообразием формулировок и контекстов, что приводит к искажению аналитических выводов. Использование больших языковых моделей (LLM) представляет собой перспективный подход, способный повысить точность и эффективность обработки данных.

В данной статье представлены наукометрические показатели, подтверждающие актуальность темы, что позволяет обосновать необходимость дальнейшего изучения. Рассматривается использование больших языковых моделей (LLM) в обработке нормативной справочной информации (НСИ). В заключении обосновывается применение LLM для решения задачи поиска дубликатов в процессе разработки инструментов нормализации данных НСИ. Аргументируется, что LLM способны повысить точность и эффективность обработки данных, что делает их незаменимыми в современных системах управления информацией.

Материалы и методы

Проведен поиск исследований на тему «LLM НСИ» и выбраны следующие статьи для анализа и обоснование предложенного решения – использования LLM для решения задачи поиска дублей при создании инструмента нормализации данных НСИ.

В статье «OpenRefine и другие альтернативные MS Excel инструменты нормализации справочников для Экспертов НСИ» рассматриваются некоммерческие программные решения, предназначенные для выполнения типовых операций по нормализации справочников, которые могут быть использованы экспертами в области нормативно-справочной информации (НСИ). Наш подход отличается от предложенных подходов тем, что мы интегрируем инструмент на основе больших языковых моделей (LLM) непосредственно в систему НСИ. Это позволяет автоматизировать процессы нормализации данных на более глубоком уровне, обеспечивая не только ускорение обработки информации, но и повышение ее точности и качества.

В статье «Как с помощью ML сократить время нормализации справочников номенклатуры с 8 часов до 30 минут?» поднимается актуальная проблема, связанная с качеством данных в системах нормативно-справочной информации (НСИ). Однако наш подход отличается от предложенного подхода статьи тем, что мы применяем модели на основе больших языковых моделей (LLM). В отличие от традиционных методов ML, которые могут быть ограничены в своей способности учитывать контекст и семантические связи между данными, LLM обладают высокой степенью гибкости и адаптивности.

В статье «Эффективное управление НСИ: зачем нужны данные из внешних источников» поднимается вопрос о значимости интеграции внешних источников данных для формирования и обогащения основных данных организаций. Однако стоит отметить, что наш подход отличается от традиционных подходов тем, что мы применяем большие языковые модели (LLM), специально ориентированные на обработку естественного языка.

В статье «Безопасность данных при использовании LLM в разговорных ассистентах» рассматриваются ключевые аспекты обеспечения безопасности данных в контексте внедрения моделей с большими языковыми моделями (LLM). Наш подход отличается тем, что LLM используется не только как инструмент для

предоставления ответов на запросы сотрудников, но и как встроенное средство для автоматического поиска дубликатов данных. В этом контексте вопросы к системе формируются автоматически в соответствии с потребностями решаемой задачи. Такой подход позволяет значительно повысить эффективность работы с разнородными данными, обеспечивая более точное выявление дублей и минимизируя вероятность ошибок, связанных с ручным вводом или анализом данных.

Результаты и обсуждение

Наукометрические показатели соответствующего направления исследований представлены на рисунках 1 и 2.

Рисунок 1 – Наукометрические показатели

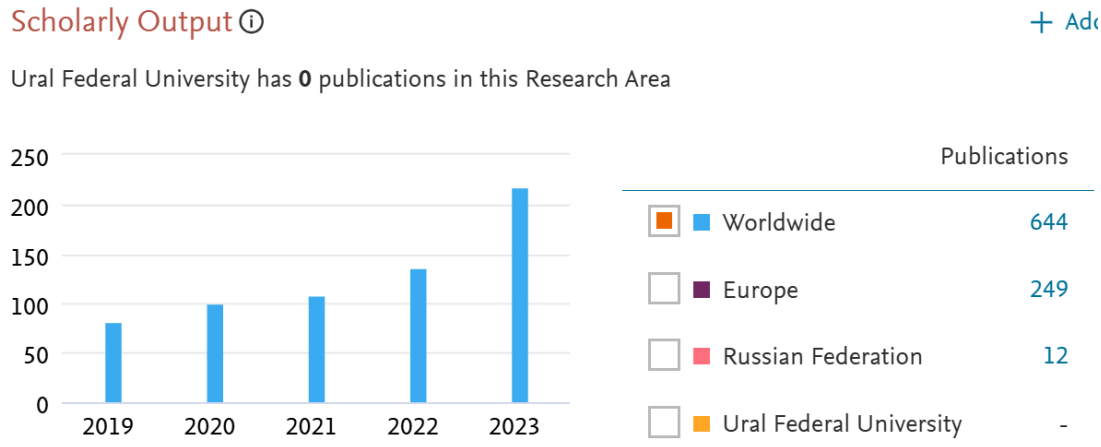

Рисунок 2 – График наукометрических показателей

Таким образом, растущие оценки показателей свидетельствуют о значимости выбранной темы исследования.

В результате проведенного исследования был предложен подход, который отличается рядом ключевых особенностей:

1. Поиск дублей осуществляется с помощью технологии (LLM). Применение LLM обеспечивает более глубокое понимание контекста и семантики данных, что позволяет эффективно идентифицировать схожие записи, даже если они представлены в различных формулировках или форматах. Это, в свою очередь, минимизирует вероятность пропуска значимых дублей, которые могут быть упущены при использовании традиционных методов поиска.

2. Запуск запроса LLM, разбор результатов и запись результатов в НСИ осуществляется автоматизировано средствами модуля. Автоматизация данных процессов не только сокращает временные затраты на выполнение операций, но и снижает вероятность человеческой ошибки, что является важным аспектом при работе с большими объемами данных. Внедрение автоматизированного подхода позволяет обеспечить непрерывность и стабильность работы модуля, а также повышает его общую эффективность.

4. Выводы

Актуальность использования больших языковых моделей (LLM) в научном сообществе подтверждается ростом наукометрических показателей и продолжающимся исследованием их применения для решения различных задач. Применение LLM для поиска дубликатов в нормативно-справочной информации (НСИ) демонстрирует высокую точность, полноту и способность учитывать контекст, что значительно улучшает качество обработки данных. В будущем рекомендуется создать инструмент нормализации данных НСИ непосредственно с использованием LLM.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Корниенкова, А. М. Применение LLM при разработке системы автоматизированного тестирования / А. М. Корниенкова // A Posteriori. – 2024. – № 6. – С. 7-9. – EDN BDGUFN.
2. Михайлович, Е. В. Большие языковые модели (LLM) - краткий обзор / Е. В. Михайлович // Исследование и проектирование интеллектуальных систем в автомобилестроении, авиастроении и машиностроении : VIII Всероссийская научно-практическая конференция с международным участием, Таганрог, 05 апреля 2024 года. – Таганрог: ДиректСайнс (ИП Шкуркин Дмитрий Владимирович), 2024. – С. 116-125. – EDN GAIWCP.
3. OpenRefine и другие альтернативные MS Excel инструменты нормализации справочников для Экспертов НСИ. URL: https://habr.com/ru/articles/786288/ (дата обращения: 10.11.2024)
4. Эффективное управление НСИ: зачем нужны данные из внешних источников. URL: https://companies.rbc.ru/news/CvmWcYlnL5/effektivnoe-upravlenie-nsi-zachem-nuzhnyi-dannyie-iz-vneshnih-istochnikov/ (дата обращения: 10.11.2024)
5. Безопасность данных при использовании LLM в разговорных ассистентах. URL: https://vocamate.ru/articles/bezopasnost-dannykh-pri-ispolzovanii-llm-v-razgovornykh-assistentakh (дата обращения: 10.11.2024)

Sagilova Eila Kairatovna,
student,
Graduate School of Economics and Management,
Ural Federal University named after the first President of Russia B.N.Yeltsin,
Yekaterinburg, Russian Federation

Volosnikov Pavel Dmitrievich,
student,
Graduate School of Economics and Management,
Ural Federal University named after the first President of Russia B.N.Yeltsin,
Yekaterinburg, Russian Federation

JUSTIFICATION OF THE USE OF LLM TO SOLVE THE PROBLEM OF SEARCHING FOR DUPLICATES WHEN CREATING AN MDM DATA NORMALIZATION TOOL

Abstract:
This article provides a rationale for using LLM to find duplicates in the process of developing an MDM data normalization tool. The analysis of existing studies was carried out, the literature on the use of LM in the context of MDM was studied, scientometric indicators were given and the relevance of using LLM was substantiated.

Keywords:
Duplicates, large language models, Master Data Management (MDM), normalization, data processing.