УДК 004.8

РАЗРАБОТКА МОДУЛЯ ИЗВЛЕЧЕНИЯ СУЩНОСТЕЙ ИЗ НЕСТРУКТУРИРОВАННЫХ МЕДИЦИНСКИХ ДОКУМЕНТОВ

Зеленин Д.С. (ИТМО)
Научный руководитель – инженер ФПИН ИТМО Лаврова А. К. (ИТМО)

Введение. Система здравоохранения представляет собой сложную информационную систему. Мединские данные хранятся и передаются разными способами, где-то используется структурированный специальный медицинский формат CDA, а где-то все сводится к PDF, PNG, DOCX. В качестве одного из подходов к решению проблемы разрозненности данных используются OCR и NER подходы. Такие решения активно развиваются для англоязычных текстов, однако для русскоязычного сегмента практически отсутствуют. [1].

Основная часть. При проектировании системы был выбран модульный подход. На первом этапе определяется тип документа, если данные относятся к PDF и PNG, то документы передаются в модуль OCR [2], далее в модуль извлечения медицинских сущностей (NER). Следующий этап - это приведение необходимых данных в формат МКБ-10. Важно отметить, что благодаря модульному подходу каждый модуль можно изменять независимо от другого. Для реализации лучшего решения по соотношению скорости и качества было протестировано несколько OCR решений, таких как: PaddleOCR, DeepSeek-OCR, EasyOCR и другие. Для модуля NER тестировались RuBERT, Qwen 1.5 и другие. Модуль OCR и NER были протестированы на искусственно созданном датасете. Для отслеживания экспериментов использовался ml-flow.

Результатом работы является сформированный файл в формате JSON, что дает возможность последующей интеграции с внешними сервисами, с таким как телемедицинский ассистент. Решение позволяет снизить долю ручной обработки медицинской документации, повысить структурированность данных.

Выводы. В ходе работы была спроектирована и разработана модульная система извлечения сущностей из медицинских документов, основанная на применении методов OCR и машинного обучения. Разработанная архитектура обеспечивает последовательную обработку медицинских документов различных форматов, как частично структурированных так и неструктурированных.

Список использованных источников:

1. Ронжин Л.В., Астанин П.А., Раузина С.Е., Ядгарова П.А., Зарубина Т.В. Разработка сервиса для автоматического извлечения именованных сущностей из неструктурированных медицинских русскоязычных текстов. // Сибирский журнал клинической и экспериментальной медицины. 2025;40(2):201-210. URL: https://doi.org/10.29001/2073-8552-2025-40-2-201-210 (дата обращения: 21.12.2025).
2. Давлетов А. Р. СОВРЕМЕННЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ И ТЕХНОЛОГИЯ OCR ДЛЯ АВТОМАТИЗАЦИИ ОБРАБОТКИ ДОКУМЕНТОВ // Вестник науки. 2023. №10 (67). URL: https://cyberleninka.ru/article/n/sovremennye-metody-mashinnogo-obucheniya-i-tehnologiya-ocr-dlya-avtomatizatsii-obrabotki-dokumentov (дата обращения: 22.12.2025).