УДК 004.91

Анализ методов получения табличных данных из PDF документов

Смирнов Н.А. (ИТМО)
Научный руководитель – старший преподаватель, Береснев А.Д. (ИТМО)

Введение. PDF (Portable Document Format) является универсальным форматом для представления документов, включающих в себя текст, изображения, мультимедийные элементы, ссылки на веб-страницы и другие компоненты. Основное предназначение формата заключается в представлении и просмотре документов, а не в их непосредственном редактировании. Эта особенность формата обусловлена его проектированием как средства для сохранения и точной передачи визуального оформления документов на различных платформах и устройствах. Преобразование в PDF обычно происходит из других форматов и предполагает финальный этап подготовки документа, когда все необходимые изменения уже выполнены, и требуется зафиксировать окончательный вид документа для распространения, публикации или архивирования. При разработке медицинской информационной системы для “НМИЦ онкологии имени Н.Н. Петрова” одной из задач являлась извлечение данных из статистических таблиц, представленных в формате PDF. В связи с сложностью структуры PDF формата, обработка и анализ таких данных является актуальной проблемой.

Основная часть. Структура PDF документа состоит из четырех основных частей [1]:

1) заголовок: отвечает за идентификацию PDF документа,
2) тело: иерархически организованные блоки объектов, которые определяют содержимое и внешний вид документа,
3) таблица перекрестных ссылок: таблица, отслеживающая местоположение каждого объекта в теле документе,
4) трейлер: элемент в конце файла, предоставляющий информацию, необходимую для нахождения других структурных элементов документа.

Существует два типа PDF документов: текстовые и отсканированные. Для текстовых документов в работе рассматриваются два основных способа получения данных из PDF таблиц:

1) Stream: основан на предположении, что текст, составляющий строки и столбцы таблицы расположен на одинаковом расстоянии друг от друга. Эффективен для обработки таблиц без четких границ,
2) Lattice: основан на анализе линий и перекрестий при помощи алгоритмов компьютерного зрения, реализованных в библиотеке OpenCV.

Для получения данных из отсканированных PDF документов в работе рассмотрено два метода:

1) Table Transformer (TATR): инновационная модель искусственного интеллекта, разработанная Microsoft AI для распознавания и извлечения таблиц из различных документов, включая PDF-файлы и изображения [2].
2) Paddle OCR: открытое программное обеспечение для оптического распознавания символов (OCR), разработанным корпорацией Baidu [3].

В работе приведен практический пример применения способа Lattice и Paddle OCR

Выводы. Проведен анализ методов получения табличных данных из PDF документов, составлен ряд рекомендаций для выбора наиболее оптимального способа. В качестве метода для работы с статистическими данными выбран Lattice.

Список использованных источников:

1. ISO 32000-1:2008(en) Document management — Portable document format — Part 1: PDF 1.7. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:32000:-1:ed-1:v1:en

2. Smock, B., Pesala, R., & Abraham, R. (2022b). PubTables-1M: Towards comprehensive table extraction from unstructured documents.
3. Paddle OCR. URL: https://github.com/PaddlePaddle/PaddleOCR