УДК 004.8

ВАЛИДАЦИЯ ШИФРОВ В ДОКУМЕНТАХ С НЕТИПИЧНЫМ ПОВТОРЯЮЩИМСЯ ПАТТЕРНОМ ЦИФР И БУКВ НА ОСНОВЕ OCR МОДЕЛЕЙ

Петрова В.В. (ИТМО)
Научный руководитель – кандидат технических наук, старший научный сотрудник Ходненко И.В.
(ИТМО)

Введение. Автоматизация нормоконтроля документов требует высокой точности распознавания структурных элементов, включая шифры с нестандартными паттернами (например, комбинации букв, цифр и знаков препинания). Для распознавания шифров используется оптическое распознавание символов (OCR) [1]. Стандартные OCR-модели не справляются с распознаванием нестандартных шрифтов, используемых в шифрах. Для постобработки распознанного текста с целью коррекции ошибок неприменимы традиционные методы нормализации текста (например, исправление опечаток, приведение к единому регистру) из-за специфики шифров.

Основная часть. Для решения задачи валидации шифров с нетипичными паттернами был разработан поэтапный подход.

1) Анализ документов и генерация датасета. Проведен анализ исходных документов, в результате которого были определены все шрифты шифров. На их основе с использованием инструмента TextRecognitionDataGenerator [2] был сгенерирован специализированный датасет, включающий различные комбинации символов (буквы, цифры, знаки препинания) в разных шрифтах.
2) Дообучение модели. OCR-модель EasyOCR была дообучена на сгенерированном датасете, что значительно повысило точность распознавания. Однако в процессе тестирования выявилась проблема: модель путала символы "0" (ноль) и "O" (буква), а также некорректно распознавала знаки препинания (например, точку и запятую). На тестовой выборке из 2203 изображений штампов неправильно распознаны «0» и «O» два и более раза в одном шифре наблюдалась в 24,1% случаев (531 изображение).
3) Дополнительное обучение на проблемных символах. Был создан отдельный датасет, содержащий последовательности с нулями, буквами "O" и знаками препинания. После дополнительного обучения точность распознавания возросла: ошибка «0»/«O» два и более раза сократилась до 0,14% случаев (3 изображения).

Выводы. Предложенный подход демонстрирует эффективность адаптации OCR-моделей с использованием специализированных датасетов. Разработанная методика может быть применена для повышения точности распознавания других категорий текстовых данных с нестандартными структурами.

Список использованных источников:

1. Маслов И.А. Оптическое распознавание символов в информационных системах и проблемы внедрения // E-Scio. 2023. №3 (78). URL: https://cyberleninka.ru/article/n/opticheskoe-raspoznavanie-simvolov-v-informatsionnyh-sistemah-i-problemy-vnedreniya (дата обращения: 15.02.2025).
2. TextRecognitionDataGenerator: документация [Электронный ресурс]. – Режим доступа: https://textrecognitiondatagenerator.readthedocs.io/en/latest/index.html (дата обращения: 13.02.2025)