Федеральное государственное автономное образовательное учреждение высшего образования
«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет Информационных технологий

Кафедра «Информатика и информационные технологии»

Направление подготовки/ специальность: Информационные системы и технологии

ОТЧЕТ

по проектной практике

Студент: Болотная Дарья Сергеевна Группа: 241-334

Место прохождения практики: Московский Политех, кафедра ИиИТ

Отчет принят с оценкой _____ Дата _____

Руководитель практики: Рябчикова Анна Валерьевна

Москва 2025

ОТЧЕТ ПО ПРОЕКТНОЙ ПРАКТИКЕ

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В ходе проектной практики был разработан программный продукт — сканер документов на основе технологий компьютерного зрения и оптического распознавания текста (OCR). Проект включал в себя создание двух версий системы: первая обрабатывает изображения в формате JPG, вторая — документы PDF.

Актуальность проекта обусловлена тем, что в повседневной жизни у людей часто возникает необходимость быстро оцифровать бумажный документ или извлечь текст из его фотографии. Применение данного программного обеспечения позволяет упростить этот процесс и получить качественный результат без привлечения дорогостоящих коммерческих решений.

ОБЩАЯ ИНФОРМАЦИЯ О ПРОЕКТЕ

Название проекта: «Сканер документов с функцией распознавания текста»

Цель проекта: Разработка универсального инструмента для сканирования документов из изображений и PDF-файлов с возможностью их преобразования в текстовую форму.

Задачи проекта:

- Реализация алгоритма детектирования границ документа на изображении.
- Перспективное преобразование для выравнивания и обрезки документа.
- Улучшение читаемости документа с помощью пороговой фильтрации.
- Применение OCR для извлечения текста.
- Поддержка форматов JPG и PDF.
- Сохранение результата в текстовый файл.

ОБЩАЯ ХАРАКТЕРИСТИКА ДЕЯТЕЛЬНОСТИ ОРГАНИЗАЦИИ

Наименование заказчика: ООО «ДокФлоу»

Сфера деятельности:

Офисная фирма, предоставляющая услуги по документообороту и сопровождению делопроизводства для малого и среднего бизнеса. Специализируется на подготовке, хранении, передаче и цифровизации деловых документов.

Организационная структура:

- **Генеральный директор** — общее руководство компанией.
- **Отдел документационного обеспечения** — прием, классификация и регистрация документов.
- **Отдел цифровизации** — сканирование, распознавание текста, архивация.
- **ИТ-отдел** — поддержка внутренних систем, сопровождение программного обеспечения.
- **Отдел продаж и клиентской поддержки** — работа с заказами и клиентами.

Проблематика и актуальность проекта:

Организация ежедневно работает с большим количеством бумажных документов. Возникает потребность в автоматизированном решении для быстрой и качественной оцифровки документов с последующим распознаванием текста и возможностью сохранить результат в различных форматах. Это позволит ускорить обработку, снизить нагрузку на сотрудников и повысить точность хранения данных.

ОПИСАНИЕ ЗАДАНИЯ ПО ПРОЕКТНОЙ ПРАКТИКЕ

Постановка задачи: Разработать скрипт, который:

- Загружает документ из изображения или PDF;
- Автоматически определяет контуры страницы;
- Выравнивает и корректирует изображение;
- Выполняет распознавание текста (OCR);
- Сохраняет извлеченный текст в файл;
- Показывает результаты визуально.

Описание командной работы:

В составе команды было 2 участника. Я отвечала за:

- Обработку изображений и PDF-файлов;
- Настройку библиотеки OCR (Tesseract);
- Реализацию алгоритмов на Python;
- Тестирование корректности работы на различных документах.

Партнер по команде занимался:

- Поиском оптимальных параметров фильтрации и преобразования;
- Поиском библиотек для обработки PDF (Poppler);
- Подготовкой документации и скриншотов.

Для координации использовался Google Docs и Discord, сроки соблюдались с использованием Trello. Возникали сложности с настройкой Poppler и определением контура документа в PDF, решалось путем ручной замены изображений.

Формирование компетенций:

- Прокачаны навыки OpenCV и работы с изображениями.
- Освоена библиотека pytesseract и технология OCR.
- Улучшены умения работы в команде, планирования и самоменеджмента.

ОПИСАНИЕ ДОСТИГНУТЫХ РЕЗУЛЬТАТОВ

Реализация версии 1 (JPG)

Основные этапы работы программы:

1. **Загрузка изображения:**
2. image = cv2.imread("document.jpg")

   image = imutils.resize(image, height=500)
3. **Предобработка:**
   - Преобразование в оттенки серого;
   - Размытие (Gaussian Blur);
   - Детекция границ (Canny).
4. **Поиск документа:** Используется контур с четырьмя вершинами — предполагаемый прямоугольник документа.
5. **Перспективное преобразование:** Прямоугольник трансформируется в ровное изображение документа.
6. **Пороговая фильтрация:** Используется адаптивный метод для улучшения четкости текста.
7. **Распознавание:** Текст извлекается с помощью Tesseract.
8. **Сохранение:** Текст сохраняется в файл scanned_text.txt.

Реализация версии 2 (PDF)

1. **Конвертация PDF в изображения:**

   pages = convert_from_path(pdf_path, dpi=300, poppler_path=poppler_path)

   Используется библиотека pdf2image, необходим Poppler.
2. **Цикл по страницам:** Каждая страница обрабатывается как изображение — так же, как в первой версии.

3. **Распознавание текста и сохранение:** Текст добавляется в итоговую строку и сохраняется в файл scanned_from_pdf.txt.

Актуальность и пользовательская ценность:

- Упрощает и ускоряет процесс оцифровки бумажных документов.
- Позволяет сразу получать редактируемый и машиночитаемый текст.
- Автоматически сохраняет результат в удобном формате.
- Устраняет необходимость вручную вырезать, поворачивать или улучшать изображения — все эти функции реализуются автоматически.
- Уменьшает нагрузку на сотрудников и сокращает количество технических ошибок.

Преимущества:

- **Интеллектуальная обработка изображений:** автоматическое кадрирование, выравнивание, фильтрация шума, улучшение контраста.
- **OCR-распознавание текста:** поддержка русского и английского языков, возможность извлекать текст с различных форматов и шрифтов.
- **Гибкость форматов:** обработка и изображений (JPG), и документов (PDF) с сохранением результата в формате TXT.
- **Минимальные требования к оборудованию:** работает на стандартных ПК и ноутбуках, не требует дорогостоящих промышленных сканеров.
- **Простота внедрения:** не требует глубокой технической подготовки для начала работы, подходит для небольших офисов и индивидуального использования.
- **Бесплатность и открытость:** решение построено на бесплатных библиотеках и не требует покупки лицензий.

ЗАКЛЮЧЕНИЕ

В ходе выполнения проекта были достигнуты все поставленные цели: реализован программный продукт, успешно обрабатывающий как изображения, так и PDF-документы. Было получено ценное практическое знание по работе с OpenCV, Tesseract, PDF2Image. Работа в команде позволила лучше организовать время и улучшить коммуникационные навыки.

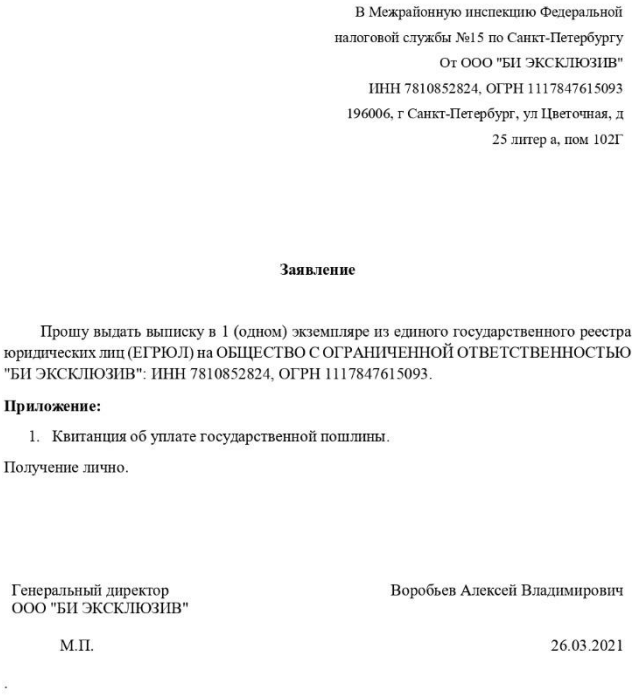
Продукт может быть полезен самым разным категориям пользователей — от студентов и преподавателей до офисных работников, бухгалтеров, юристов и специалистов, работающих с бумажными документами. Он позволяет быстро и удобно преобразовывать физические документы в цифровую форму без необходимости приобретать дорогостоящее профессиональное оборудование или устанавливать сложное программное обеспечение. Благодаря простоте использования, широким функциональным возможностям и ориентации на повседневные задачи, разработанное решение способствует повышению эффективности работы и упрощает процесс цифровизации даже для неопытных пользователей.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. OpenCV Documentation — https://docs.opencv.org/
2. Pytesseract GitHub — https://github.com/madmaze/pytesseract
3. Poppler Project — https://poppler.freedesktop.org/
4. PDF2Image Documentation — https://github.com/Belval/pdf2image
5. Python Official Documentation — https://docs.python.org/3/

ПРИЛОЖЕНИЯ

- Пример сканируемого изображения (JPG), взятого из открытого источника.



В Межрайонную инспекцию Федеральной
налоговой службы №15 по Санкт-Петербургу
От ООО "БИ ЭКСКЛЮЗИВ"
ИНН 7810852824, ОГРН 1117847615093
196006, г Санкт-Петербург, ул Цветочная, д
25 литер а, пом 102Г

Заявление

Прошу выдать выписку в 1 (одном) экземпляре из единого государственного реестра юридических лиц (ЕГРЮЛ) на ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "БИ ЭКСКЛЮЗИВ": ИНН 7810852824, ОГРН 1117847615093.

Приложение:

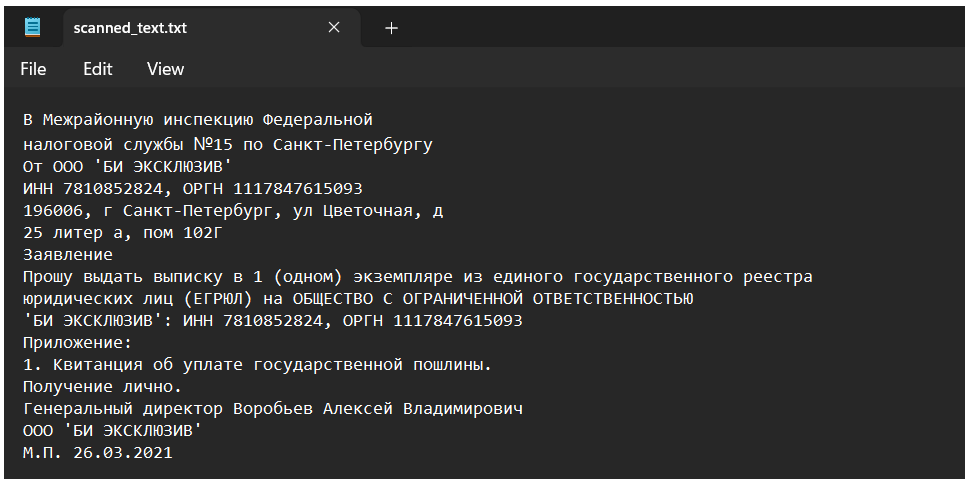
1. Квитанция об уплате государственной пошлины.

Получение лично.

Генеральный директор
ООО "БИ ЭКСКЛЮЗИВ" Воробьев Алексей Владимирович

М.П. 26.03.2021

- Текст, извлеченный из данного изображения.



- Фрагменты кода (см. раздел «Описание достигнутых результатов»).

Подтверждаю, что отчет выполнен лично и соответствует требованиям практики

Болотная Дарья Сергеевна, 03.06.2025

Подпись: