ПРОЕКТ МЭРА МОСКВЫ

ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ

РАЗВИТИЕ ЧЕЛОВЕЧЕСКОГО КАПИТАЛА

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ГЛАВАРХИВ МОСКВЫ

PM x FE

09. Сервис извлечения и индексирования информации из образов архивных документов (Ретроконверсия)

КОМАНДА «PM x FE»

ПРОЕКТ МЭРА МОСКВЫ
ДЕПАРТАМЕНТ ПРЕДПРИНИМАТЕЛЬСТВА И ИННОВАЦИОННОГО РАЗВИТИЯ ГОРОДА МОСКВЫ
РАЗВИТИЕ ЧЕЛОВЕЧЕСКОГО КАПИТАЛА
ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

О команде

- Город: Москва
- Количество человек: 2
- Капитан команды: Вязьмин Максим

Наименование задачи:

Сервис извлечения и индексирования информации из образов архивных документов (Ретроконверсия)

Описание решения:

Локальное веб-приложение для распознавания рукописного и машинописного текста.

Решение предназначено для автоматизации перевода бумажных документов в цифровой формат с последующей индексацией.

Максим Вязьмин

Никита Вишневский

Как вы планируете дальше использовать или развивать ваше решение:

1. Разработка базы-данных для сохранения прогресса и повышения удобства сортировки.
2. Автоматизация индексации ключевых параметров.
3. Доработка механизма обучения.
4. Доработка интерфейса.

КОМАНДА «PM x FE»

Максим Вязьмин

- Капитан, проджект-менеджер
- @maxim_vyazmin
- +7(916)497-22-14

Никита Вишневский

- Frontend-разработчик
- @Lugertr
- +7(916)135-06-38

PM x FE

IT- PROJECTS 4 LIFE

Наши ценности

- ❖ Потребность заказчика – наша забота
- ❖ Качество и скорость в ограниченное время
- ❖ Невыполнимых задач не бывает

Наш девиз

«2 человека тоже могут свернуть гору»

КОМАНДА «PM x FE»

Краткая история команды:

Мы дружим со школьных времен, в целом по жизни часто кооперировались для каких-то задач или вопросов. В этот раз решили впервые поучаствовать в хакатоне, так как каждый из нас имеет отношение к IT (один к менеджер, другой frontend-разработчик). Стало интересно сможем ли мы в сжатые сроки и с ограниченными ресурсами выполнить сложную задачу.

Почему вы выбрали именно эту задачу из предложенных на хакатоне?

Мы выбрали эту задачу, так как ранее не сталкивались в своей профессиональной деятельности с ИИ. Подробно изучив описание задач, и оценив свои возможности, способности и знания решили, что эта задача была бы оптимальным вариантом опробовать свои силы в разработки продуктов с ИИ + это было наиболее близким к нашему профессиональному опыту

С какими основными сложностями или вызовами вы столкнулись и как их преодолели?

Наиболее сложным моментом по ходу разработки стало требование использования отечественных//open-source решений при сохранении качества продукта и одновременно выполнения требований технического задания. Для решений этой проблемы мы использовали open-source решений с открытой лицензией, также была разработана свой модель обучения/дообучения модели компьютерного зрения, так как имеющиеся решения базировались на разработках «нежелательных» организаций.

Также одной из главных проблем стал вопрос обучения OCR-модели, так как наш датасет состоял из 72 тысяч экземпляров рукописей и требовалось на мощностях домашних ПК запускать серии обучения. Для решения проблемы мы пользовались скриптами ограничивающими производительность занятую на обучение, а также системой чекпоинтов и сохранений (к сожалению мощности Яндекс Клауд мы получили за 2-3 дня до дедлайна).

ОПИСАНИЕ РЕШЕНИЯ

Описание веб-приложения

Веб-сервис представляет собой комплексное решение для автоматизированного распознавания и обработки архивных документов. На текущий момент успешно реализовано и функционирует ядро системы, включающее мощный веб-интерфейс для загрузки изображений, отслеживания задач в реальном времени и, что наиболее важно, — инструмент верификации. Данный инструмент позволяет эксперту сопоставлять распознанный текст с оригинальными фрагментами изображения и вносить правки, обеспечивая высочайшую точность результата.

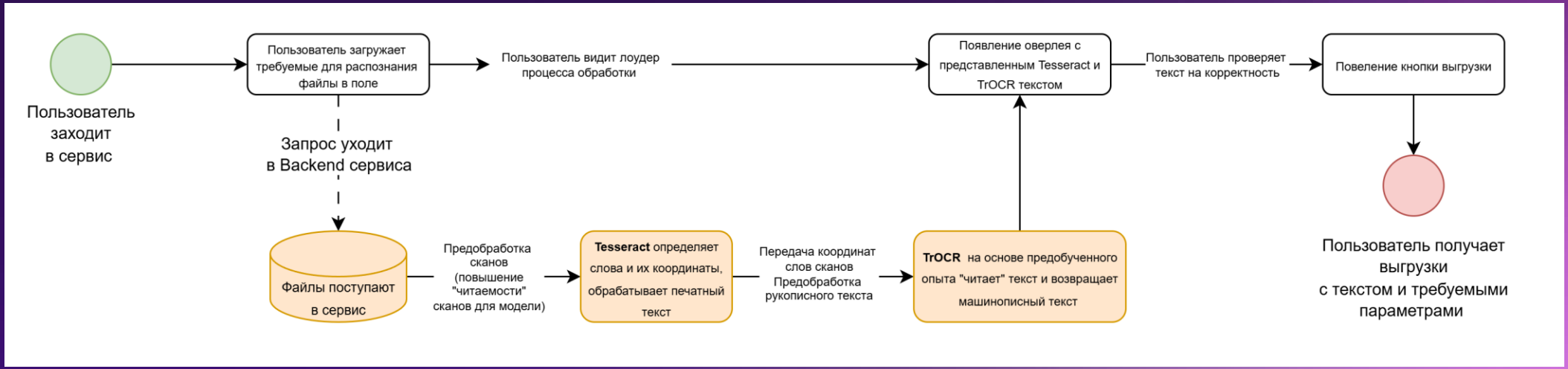
Ключевым преимуществом решения является его полное соответствие строгим требованиям информационной безопасности. Система функционирует в изолированном контуре, гарантируя полную конфиденциальность обрабатываемых данных. Технологический стек построен на открытом программном обеспечении. В основе обработки лежат сложные алгоритмы нормализации изображений и распознавания текста, которые успешно справляются с печатными и рукописными материалами, включая документы с дореволюционной орфографией.

На данный момент основные модули распознавания и верификации полностью интегрированы и протестированы.

Представленный функционал

1. Загрузка изображений;
2. Предобработка изображения (отдельные для выявления печатного и для рукописного текста);
3. Определение координат слов на изображении;
4. Определение текста и конвертация в машинописный текст;
5. Редактирования полученного текста;
6. Выгрузка результатов.

Упрощенный процесс работы приложения



ТЕХНОЛОГИЧЕСКИЙ СТЕК

01

Язык программирования: Python

- Простота в обращении, большая гибкость и хорошая адаптируемость

02

Frontend: Angular

- Высокая скорость работы, практичность, легкость в обращении и доработке

03

Виртуальное окружение: Ubuntu

- Высокая совместимость с серверным оборудованием
- Гибкость при установке и настройке библиотек

04

OCR-модель для печатного текста: Tesseract

- Функционал определения координат слов
- Высокий уровень обработки печатного текста

05

OCR-модель для рукописного текста: TrOCR

- Определение рукописного текст
- Функционал обучения модели

УНИКАЛЬНОСТЬ РЕШЕНИЯ

Приложение реализовано полностью на базе open-source решений

Tesseract – модель от Google, поддерживающая более 100 языков, распространяется по свободной лицензии

TrOCR – модель от Microsoft. В отличие от традиционных OCR-систем, основанных только на сверточных нейросетях, TrOCR сочетает Vision Transformers (ViT) с последовательным моделированием, что позволяет анализировать контекст и пространственные отношения в тексте. Распространяется по свободной лицензии.

Ключевые преимущества:

- Точность на уровне человека: Эффективно распознает рукописный и искаженный текст
- Многоязычная поддержка: Предобученные модели для английского, французского, немецкого и других языков
- Комплексное решение: Объединяет обнаружение и распознавание текста в едином процессе
- Простая интеграция: Работает на базе библиотеки Hugging Face Transformers

Для реализации требования «реализация решения на свободном ПО с открытым ПО» был разработан **собственный механизм обучения модели TrOCR** , так как альтернативные инструменты не отвечают требованиям, например популярным механизмом к обучению модели является механизм другой модели PyTorch (разработанной экстремистской организацией Meta).

ПЛАНЫ ПО РАЗВИТИЯ РЕШЕНИЯ

Автоматизация индексации

- для повышения скорости обработки файлов и экономии ресурсов (человеческих и временных)

Разработка БД

- для сохранения прогресса и повышения удобства сортировки обработанных документов

Доработка механизма обучения

- доработка инструмента обучения для повышения результатов

Доработка интерфейса

- улучшение пользовательского опыта