SL SOFT × SOiCA

Распознавание документов. Как современные OCR-технологии с ИИ меняют подход

Гипераавтоматизация как результат новых возможностей

SL SOFT × SOiCA

- **Гипераавтоматизация** — быстрое выявление и автоматизация как можно большего числа бизнес- и ИТ-процессов: **«все, что может быть автоматизировано, должно быть автоматизировано!»**

- **Инструменты для анализа процессов и задач:** Process-mining, аналитика больших данных

- **Средства автоматизации**, снижающие трудоемкость: RPA, Low-Code/No-Code, iPaaS

- **Решения для поддержки бизнес-логики:** iBPMs, управление бизнес-правилами и пр.

- **Технологии искусственного интеллекта и машинного обучения:** обработка естественного языка (NLP), аналитика неструктурированных данных, OCR, цифровые сотрудники и чат-боты

SL SOFT × SOiCA

Платформа SOICA

Распознавание, извлечение и обработка данных из скан-образов и цифровых копий структурированных и неструктурированных документов любого типа

Импорт данных

Захват данных из любых источников: писем, вложений, фотографий, архивов и др.

Предварительная обработка

Бинаризация и очистка при использовании 18 встроенных фильтров

Распознавание и классификация

Полнотекстовое и атрибутивное распознавание данных из любых типов документов

Контроль

Проверка корректности данных, поиск ошибок, контроль комплектности

Модификация

Добавление штрихкодов, факсимиле, деперсонализация, обработка изображений

Экспорт данных

Интеллектуальная маршрутизация данных, интеграция с ECM, CRM, ERP и др.

Где в SOICA ИИ

SL SOFT × SOiCA

Распознавание

Поиск и извлечение

Распознавание – движок OCR

SL SOFT × SOiCA

Движок OCR – это, как минимум:

1. Движок детектирования
2. Движок распознавания

Качество распознавания и функциональность движка влияют на:

1. Необходимые алгоритмы пред/пост обработки документов
2. Сложность настройки извлечения данных
3. Время обработки

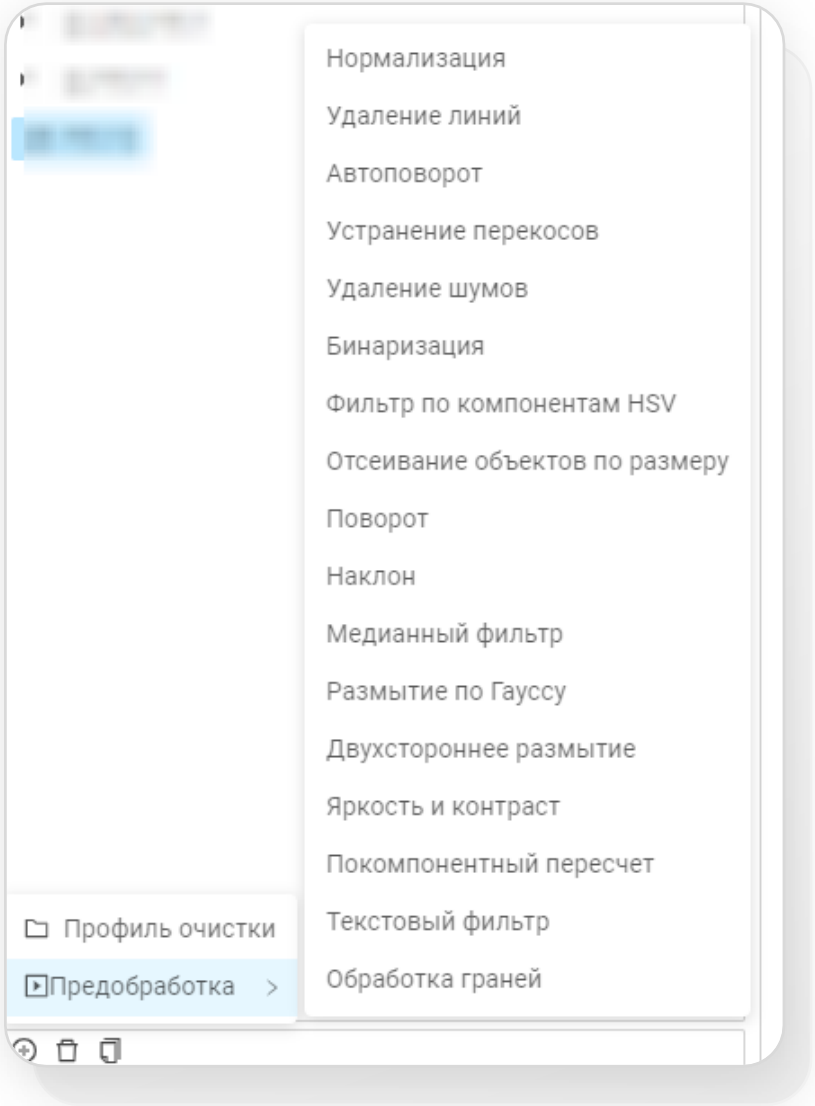
Предобработка

SL SOFT × SOiCA

Для улучшения изображений в SOICA есть > 15 инструментов.

Благодаря такой предобработке система может:

- Выделять на изображении только нужную информацию по цвету, размеру и так далее
- Нивелировать влияние плохого качества изображения на извлекаемые атрибуты
- Видоизменять оригинальное изображение под конкретную задачу



Работа движка

SL SOFT × SOiCA

Новая версия

Старая версия

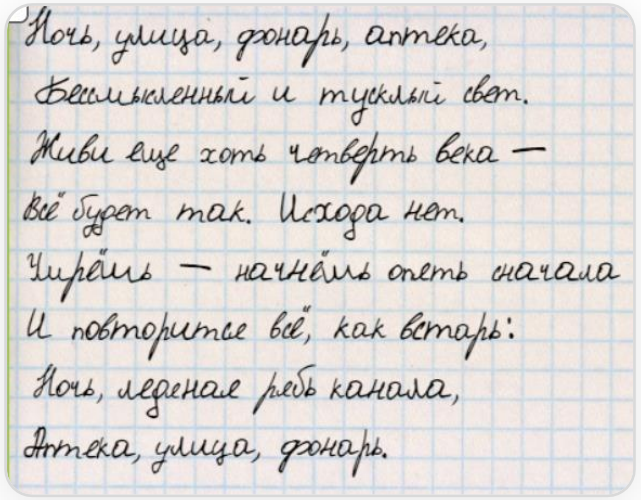
Распознавание рукописного текста

SL SOFT × SOiCA

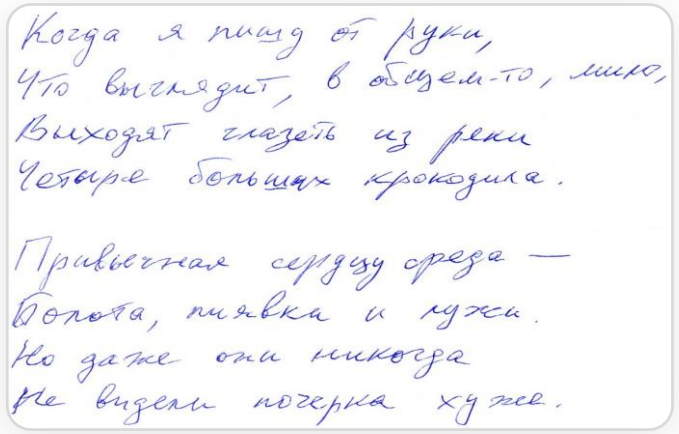
Параметры:

1. **Точность***
   - по словам: 80%
   - по символам: 95%

2. **Скорость обработки (1 страница)**
   - От 7 сек CPU
   - 0,3 сек GPU



Ночь, улица, фонарь, аптека,
Бессмысленный и тусклый свет.
Живи еще хоть четверть века
Всё будет так. Исхода нет.
Умрешь начнём опять начала
и повторится всё, как встарь:
Ночь, ледяная рябь канала,
Аптека, улица, фонарь.



Когда я пишу от руки,
Что выглядит, в общем-то, миро,
Выходят глазать из реки
Четыре больших прокодила.
Привычная сердцу среда
Болота, пиявка и лужи.
Но даже они никогда
Не видели почерка ху же.

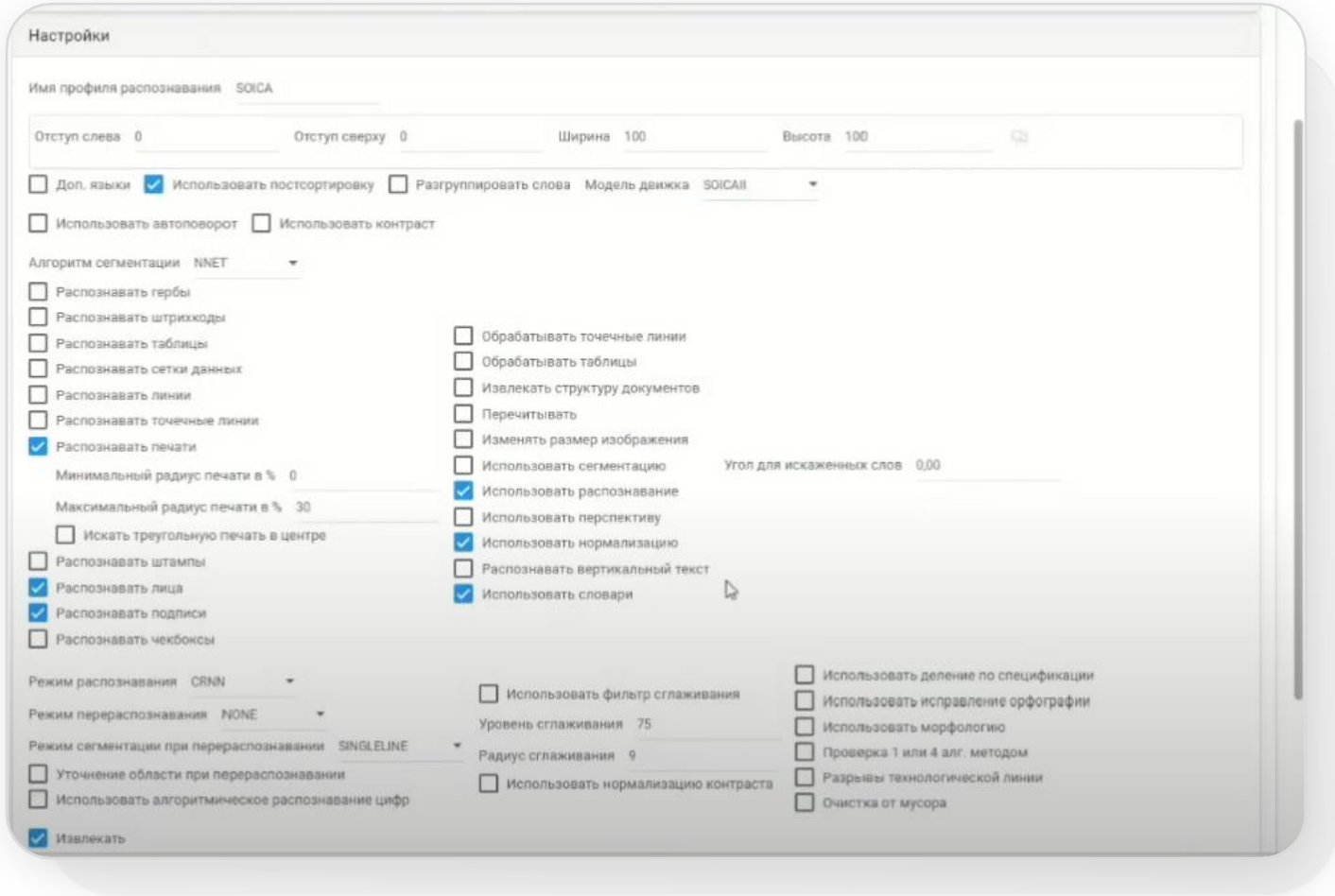
*Замерено на датасете
https://www.kaggle.com/datasets/constantinwerner/cyrillic-handwriting-dataset/data

Извлечение данных

SL SOFT × SOiCA

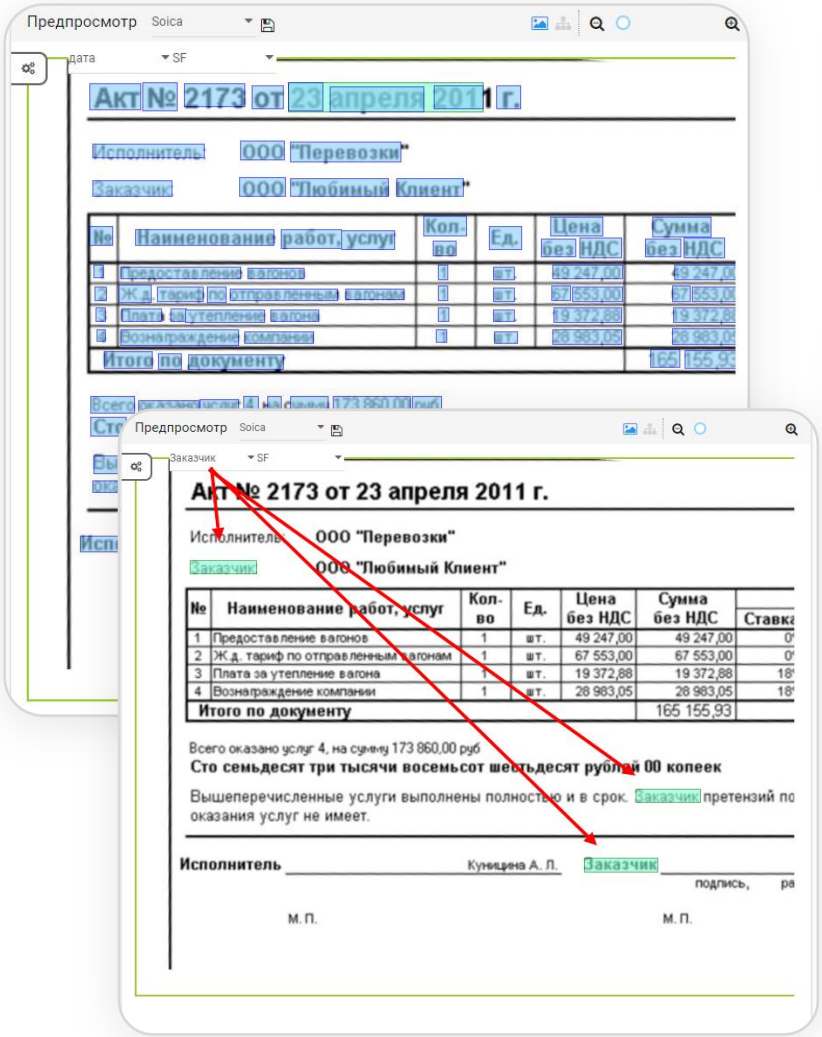
Расширенная функциональность движка SOICA II – поиск не только текста, но и:

1. таблиц
2. печатей
3. гербов
4. линий
5. чекбоксов

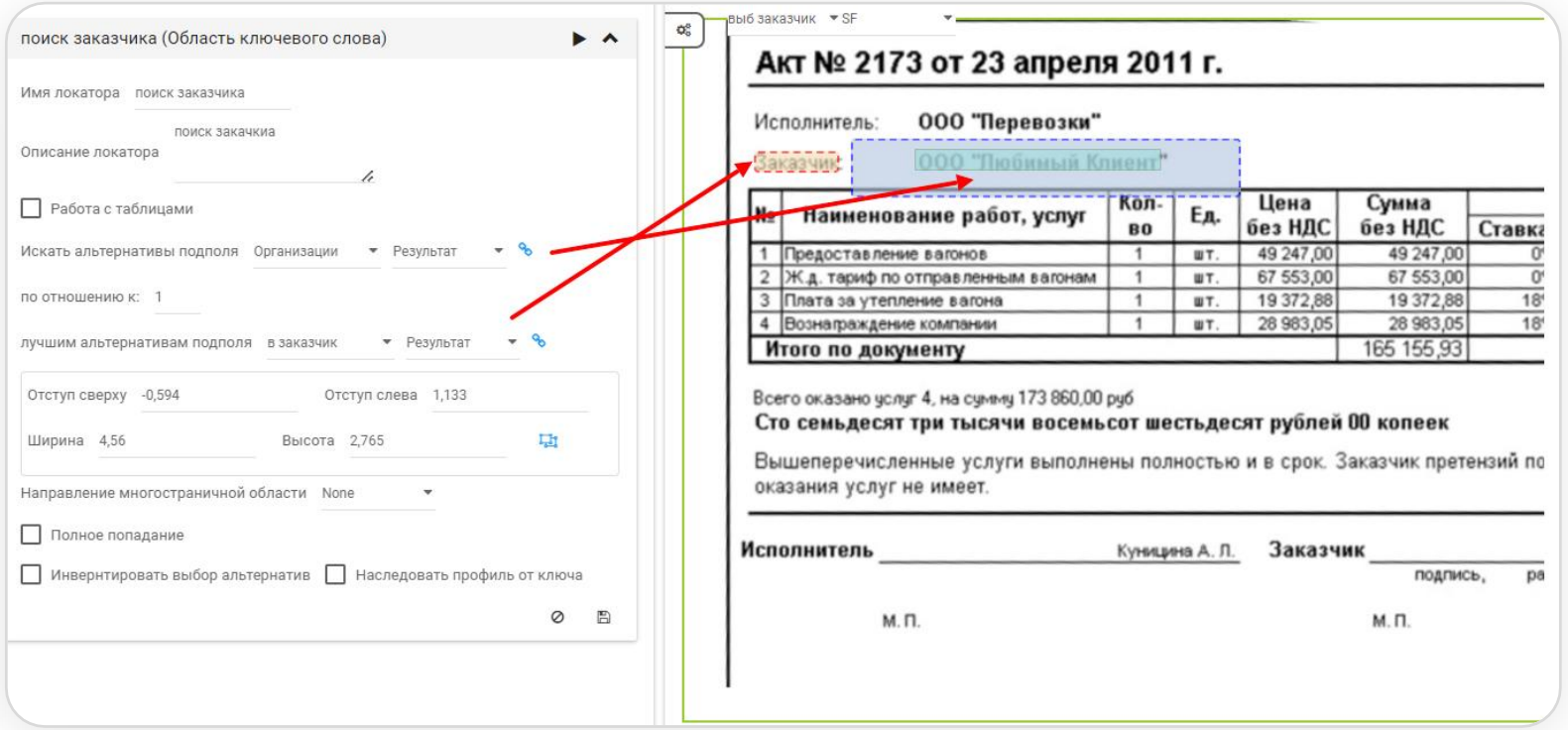


Поиск и извлечение данных

SL SOFT × SOiCA



- Вот так мы ищем данные с помощью алгоритмов



ML-локатор. Демо.

- Выполняет автоматическое детектирование области поиска целевых данных без привязки к ключам (текст, фото, печать, картинка…).
- Можно использовать несколько моделей для поиска различного набора атрибутов (в рамках класса документов).
- Легко создавать, обучать и до обучать нейронные сетки под конкретные классы документов.
- Для обучения модели нужно разметить от 50 примеров
- Возможность дообучения моделей.
- Время детектирования областей <1 секунды.
- Время обучения – 1,5-2 часа (CPU – Xeon Gold -).

Инструмент полностью встроен в среду SOICA:

1. Разметка
2. Обучение
3. Использование моделей в нескольких проектах
4. Перенос моделей между серверами
5. Перенос датасетов вместе с моделями (опционально)

Извлечение с помощью LLM. В следующей версии...

SL SOFT × SOiCA

- Настройка извлечения данных с помощью текстового промта

Textbox

Ниже представлен текст. Тебе необходимо проанализировать его и ответить на вопрос: напиши все консерванты использующиеся в продукте исходя из текста, который будет дальше. Ответ оформи в XML формате.

Instruction:
Ответ предоставь на русском языке. Ответ должен содержать только названия и отвечать на приведенный выше вопрос. Не уточняй ответ. Не дополняй ответ комментариями. Не переспрашивай и не задавай уточняющий вопрос. Не извиняйся. Не давай определения консервантам.
Пример XML:

```
<консерванты>
  <консервант> </консервант>
  <консервант> </консервант>
</консерванты>
```

ответ:

```
### Response:
<консерванты>
 <консервант>E202</консервант>
 <консервант>E211</консервант>
</консерванты>
```

Вопросы?

SL SOFT × SOiCA

Команда SOICA

info@slsoft.ru

- Александр Павлов
- Георгий Гончаров

Спасибо за внимание