



SL SOFT

×



SOiCA

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

Гиперавтоматизация как результат НОВЫХ ВОЗМОЖНОСТЕЙ



- **Гиперавтоматизация** – 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



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



Распознавание



Поиск
и извлечение

Распознавание – движок OCR

**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

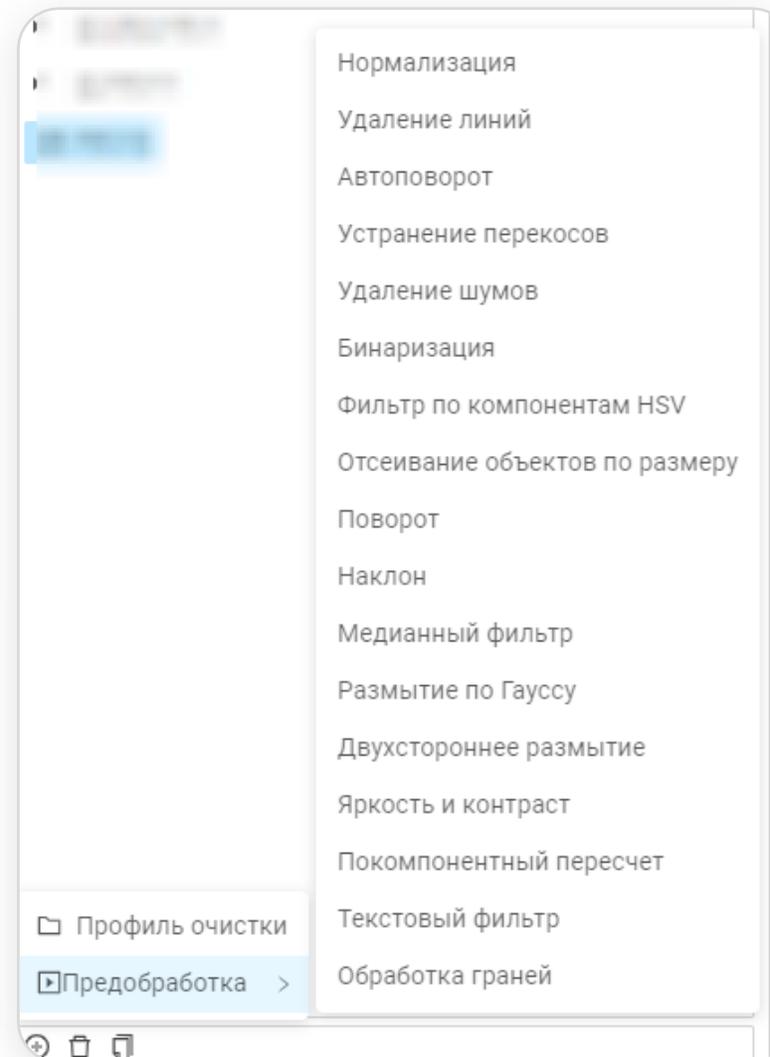
Для улучшения изображений в SOiCA
есть **> 15 инструментов.**

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

Новая версия



Старая версия



Распознавание рукописного текста

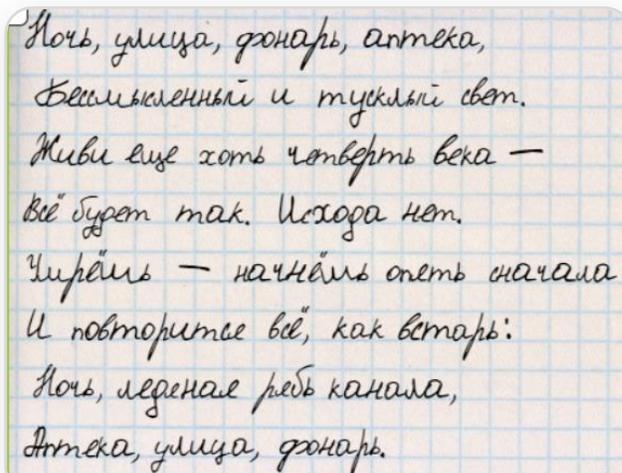
Параметры:

1 Точность*

- по словам: 80%
- по символам: 95%

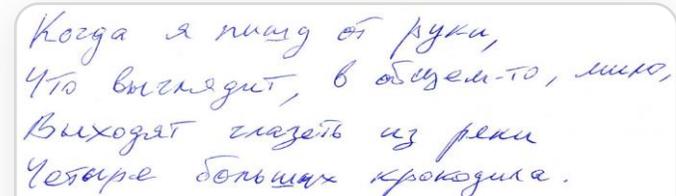
2 Скорость обработки (1 страница)

- От 7 сек CPU
- 0,3 сек GPU

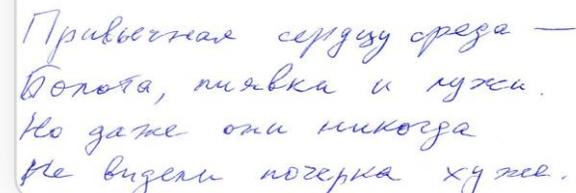


Ночь, улица, фонарь, аптека,
Бессмысленный и тусклый свет.
Живи еще хоть четверть века —
Всё будет так. Исхода нет.
Умрешь — начнём опять начала
И повторится всё, как встарь:
Ночь, ледяная рябь канала,
Аптека, улица, фонарь.

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 глаза из реки
Четыре больших крокодила.



Привычная сердцу среда —
Болота, пиявка и лужи.
Но даже они никогда
Не видели почерка хуже.

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

Извлечение данных

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

1 таблиц

2 печатей

3 гербов

4 линий

5 чекбоксов

Настройки

Имя профиля распознавания SOICA

Отступ слева 0 Отступ сверху 0 Ширина 100 Высота 100

Доп. языки Использовать постсортировку Разгруппировать слова Модель движка SOICAII

Использовать автоповорот Использовать контраст

Алгоритм сегментации NNET

Распознавать гербы
 Распознавать штрихкоды
 Распознавать таблицы
 Распознавать сетки данных
 Распознавать линии
 Распознавать точечные линии
 Распознавать печати
 Минимальный радиус печати в % 0
 Максимальный радиус печати в % 30
 Искать треугольную печать в центре

Распознавать штампы
 Распознавать лица
 Распознавать подписи
 Распознавать чекбоксы

Режим распознавания CRNN
Режим перерасознавания NONE
Режим сегментации при перерасознавании SINGLELINE

Уточнение области при перерасознавании
 Использовать алгоритмическое распознавание цифр
 Извлекать

Обрабатывать точечные линии
 Обрабатывать таблицы
 Извлекать структуру документов
 Перечитывать
 Изменять размер изображения
 Использовать сегментацию Угол для искаженных слов 0,00
 Использовать распознавание
 Использовать перспективу
 Использовать нормализацию
 Распознавать вертикальный текст
 Использовать словари

Использовать фильтр сглаживания Использовать деление по спецификации
Уровень сглаживания 75 Использовать исправление орфографии
Радиус сглаживания 9 Использовать морфологию
 Использовать нормализацию контраста Проверка 1 или 4 алг. методом
 Очистка от мусора Разрывы технологической линии

Поиск и извлечение данных

- Вот так мы ищем данные с помощью алгоритмов

Предпросмотр Soica

дата SF

Акт № 2173 от 23 апреля 2011 г.

Исполнитель: 000 "Перевозки"

Заказчик: 000 "Любимый Клиент"

№	Наименование работ, услуг	Кол-во	Ед.	Цена без НДС	Сумма без НДС
1	Предоставление вагонов	1	шт.	49 247,00	49 247,00
2	Ж.д. тариф по отправленным вагонам	1	шт.	67 553,00	67 553,00
3	Плата за утепление вагона	1	шт.	19 372,88	19 372,88
4	Вознаграждение компании	1	шт.	28 983,05	28 983,05
Итого по документу					165 155,93

Всего оказано услуг 4, на сумму 173 860,00 руб.

Предпросмотр Soica

Заказчик SF

Акт № 2173 от 23 апреля 2011 г.

Исполнитель: 000 "Перевозки"

Заказчик: 000 "Любимый Клиент"

№	Наименование работ, услуг	Кол-во	Ед.	Цена без НДС	Сумма без НДС	Ставка
1	Предоставление вагонов	1	шт.	49 247,00	49 247,00	0'
2	Ж.д. тариф по отправленным вагонам	1	шт.	67 553,00	67 553,00	0'
3	Плата за утепление вагона	1	шт.	19 372,88	19 372,88	18'
4	Вознаграждение компании	1	шт.	28 983,05	28 983,05	18'
Итого по документу					165 155,93	

Всего оказано услуг 4, на сумму 173 860,00 руб.
Сто семьдесят три тысячи восемьсот шестьдесят рублей 00 копеек

Вышеперечисленные услуги выполнены полностью и в срок. Заказчик претензий по оказанию услуг не имеет.

Исполнитель _____ Куницына А. Л. Заказчик _____
подпись, ра
м.п. м.п.

поиск заказчика (Область ключевого слова)

Имя локатора: поиск заказчика

Описание локатора: поиск заказчика

Работа с таблицами

Искать альтернативы подполя: Организации Результат

по отношению к: 1

лучшим альтернативам подполя: в заказчик Результат

Отступ сверху: -0,594 Отступ слева: 1,133

Ширина: 4,56 Высота: 2,765

Направление многостраничной области: None

Полное попадание

Инвертировать выбор альтернатив Наследовать профиль от ключа

выб заказчика SF

Акт № 2173 от 23 апреля 2011 г.

Исполнитель: 000 "Перевозки"

Заказчик: 000 "Любимый Клиент"

№	Наименование работ, услуг	Кол-во	Ед.	Цена без НДС	Сумма без НДС	Ставка
1	Предоставление вагонов	1	шт.	49 247,00	49 247,00	0'
2	Ж.д. тариф по отправленным вагонам	1	шт.	67 553,00	67 553,00	0'
3	Плата за утепление вагона	1	шт.	19 372,88	19 372,88	18'
4	Вознаграждение компании	1	шт.	28 983,05	28 983,05	18'
Итого по документу					165 155,93	

Всего оказано услуг 4, на сумму 173 860,00 руб.
Сто семьдесят три тысячи восемьсот шестьдесят рублей 00 копеек

Вышеперечисленные услуги выполнены полностью и в срок. Заказчик претензий по оказанию услуг не имеет.

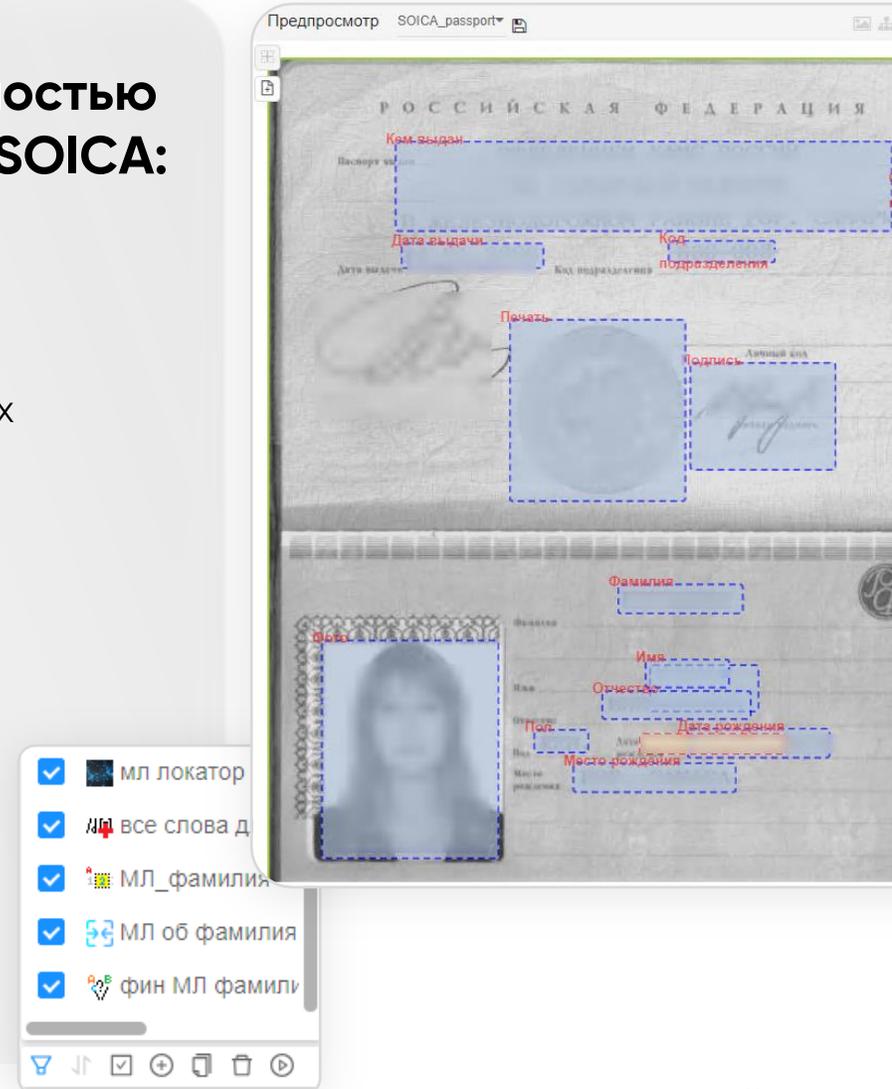
Исполнитель _____ Куницына А. Л. Заказчик _____
подпись, ра
м.п. м.п.

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...

- 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