Задание 2: Оцифровка чеков и извлечение данных

Суть задачи

Разработать систему, которая автоматически обрабатывает изображения чеков и извлекает текст и ключевые данные.

В реальных задачах цифровизации документов требуется быстро и массово переводить изображения в структурированные данные. При этом документы могут отличаться по качеству, структуре и содержанию.

Система должна:

- извлекать текст из изображений;
- находить ключевые поля;
- работать на потоке документов;
- обеспечивать баланс между точностью и скоростью.

Особенность задачи: необходимо не просто распознать текст, а построить **устойчивый и производительный pipeline обработки документов**.

Входные данные

Датасет: SROIE

Форматы:

- `.jpg`, `.png`

Документы могут содержать:

- шум;
- наклон;
- различное качество;
- различную структуру.

Требования к решению

- OCR / VLM / гибрид;
- HTTP API:

```
POST /scan_receipts
```

- пакетная обработка;
- автоматическая работа.

Формат результата

```
{
  "results": [
    {
      "doc_id": "",
```

```
      "status": "ok",
      "full_text": "",
      "fields": {
        "company": "",
        "date": "",
        "total": "",
        "address": ""
      },
      "metadata": {
        "method": "",
        "duration_ms": 0
      }
    }
  ]
}
```

Критерии оценки

- Точность извлечения — 40%
- Производительность — 30%
- Полнота обработки — 15%
- Эффективность ресурсов — 10%
- Обоснованность решения — 5%

Ожидаемый результат

Система, способная быстро и точно извлекать данные из изображений чеков и масштабироваться на большие объёмы. Приветствуются доработки, повышающие удобство и практическую ценность решения.