Федеральное агентство связи
Уральский технический институт связи и информатики (филиал)
ФГБОУ ВО «Сибирский государственный университет
телекоммуникаций и информатики» в г. Екатеринбурге
(УрТИСИ СибГУТИ)

Шаблон кейса на Всероссийский хакатон связи 2025

ООО «К ТЕЛЕКОМ»

Разработка сервиса автоматического извлечения данных из платежных счетов в формате PDF

Кондюрин Сергей Денисович, ведущий специалист учебного центра

Екатеринбург
2025

1 Цель кейса: Разработать специализированный сервис (телеграм-бот), способный автоматически извлекать структурированную информацию из сканов платежных документов (PDF) для уменьшения временных затрат на ручную обработку финансовых данных.

2 Требования:

- Сервис должен быть реализован в виде телеграм-бота.
- Бот должен позволять пользователям загружать PDF-файлы счетов.
- Сервис должен возвращать структурированные данные в формате JSON с сохранением иерархии документа.
- Система должна корректно обрабатывать различные форматы платежных документов.
- Сервис должен автоматически пропускать нераспознаваемые поля.
- Должны быть предоставлены Jupyter ноутбуки, демонстрирующие процесс обучения модели и полученную точность.

3 Необходимое ПО (со ссылками) *если потребуется***:**

Python 3.x или C++ (для модели)

Go или Python (для бэкенда)

Docker, Docker Compose (для контейнеризации)

Библиотеки для обработки PDF и компьютерного зрения (например, PyMuPDF, Tesseract OCR, OpenCV) – конкретные технологии на выбор команды.

4 Исходные данные:

Для обучения и тестирования модели необходимо использовать датасеты с аннотированными счетами в формате PDF. (Конкретные ссылки на датасеты или примеры данных должны быть предоставлены отдельно).

5 Задания:

1. Разработать и обучить модель машинного обучения для распознавания и извлечения данных из PDF-счетов.

2. Реализовать backend-часть сервиса на предпочтительном языке (Go или Python).
3. Интегрировать модель в backend-сервис.
4. Реализовать телеграм-бота для взаимодействия с пользователем.
5. Обеспечить контейнеризацию решения с использованием Docker и Docker Compose.
6. Подготовить документацию и Jupyter ноутбуки с демонстрацией обучения и метрик модели.

6 Требуемый результат выполнения задания:

- Рабочий телеграм-бот, развернутый в контейнере.
- API для загрузки и обработки PDF-файлов.
- Структурированные данные в формате JSON, извлеченные из загруженных счетов.
- Jupyter ноутбуки с кодом обучения и валидации модели.
- Исходный код проекта в репозитории (например, GitHub).

7 Литература (если есть):

Документация по использованию Tesseract OCR.

Руководства по разработке телеграм-ботов.

Документация по Docker и Docker Compose.

(Дополнительная литература может быть предоставлена экспертом)