Кейс №2: Система конвертации PDF документов в текстовые файлы

1. Введение

В данном кейсе необходимо разработать систему для конвертации PDF документов в текстовые файлы с сохранением структурных элементов. Особое внимание уделяется корректному извлечению и преобразованию таблиц в структурированные таблицы, а также схем и диаграмм в соответствующие форматы (например, в виде изображений или векторных схем). Система предназначена для обработки базы данных документов компании, в которой большинство PDF не содержит текстового слоя, а часть документов имеет текстовый слой.

2. Цель проекта — Автоматическая конвертация PDF: разработать алгоритмы для извлечения текста из PDF документов, как с текстовым слоем, так и без него (с применением OCR).

- Сохранение структурных элементов: обеспечить корректное преобразование таблиц в структурированные таблицы и схем/диаграмм в соответствующие форматы с сохранением визуальной структуры.
- Интеграция с базой данных: организовать обработку документов из существующей базы данных компании, гарантируя высокую точность и воспроизводимость результатов.
- Предоставление возможности ручной корректировки: разработать интерфейс для просмотра и редактирования результатов конвертации, если автоматический процесс требует доработки.
- Оценка стоимости решения: провести анализ затрат на разработку, внедрение, эксплуатацию, поддержку и масштабирование системы, включая оценку затрат на облачные сервисы, лицензии и трудовые ресурсы.
- Оптимизация и масштабируемость: обеспечить обработку большого объёма документов с минимальными задержками и возможностью дальнейшего масштабирования системы.

3. Функциональные возможности

3.1 Конвертация PDF документов

• Обработка документов с текстовым слоем: извлечение текста, форматирование и сохранение структуры исходного документа.

• Обработка документов без текстового слоя: применение OCR для распознавания текста, извлечение информации и последующее форматирование.

• Корректное преобразование таблиц: извлечение табличных данных с сохранением ячеечной структуры, разбиение на строки и столбцы.

• Извлечение схем и диаграмм: определение графических элементов (схем, диаграмм) и их сохранение в виде изображений или векторных схем с сохранением оригинальной композиции.

3.2 Интеграция с базой данных

• Импорт документов из существующей базы данных компании.

• Обработка различных типов PDF файлов (с текстовым слоем и без него) в едином рабочем процессе.

• Экспорт результатов конвертации в виде текстовых файлов с дополнениями (таблицы, схемы).

ID	ИМЯ	ВОЗРАСТ	ДАТА РОЖДЕНИЯ	ЗАРПЛАТА ($)	АКТИВЕН	КОММЕНТАРИЙ
1	Иван Петров	28	1995-04-12	3500.50	Да	Хороший работник
2	Мария Сидорова	34	1989-07-23	4200.00	Нет	В отпуске по уходу
3	Петр Иванов	45	1978-11-05	5100.75	Да	Лидер команды
4	Ольга Кузнецова	29	1994-02-18	3800.00	Да	Требует повышения
5	Александр Павлов	31	1992-09-30	4100.25	Нет	Уволился вчера
6	Елена Григорьева	27	1996-05-14	3300.00	Да	Новый сотрудник
7	Дмитрий Смирнов	37	1986-08-22	4700.50	Да	Отличные навыки
8	Анна Ковалева	26	1997-12-01	3600.00	Да	Требуется обучение
9	Сергей Морозов	42	1981-03-10	4900.00	Нет	На больничном
10	Татьяна Федорова	33	1990-06-15	4300.75	Да	Профессионал своего дела