DEDOC: СИСТЕМА ИЗВЛЕЧЕНИЯ СТРУКТУРЫ ДОКУМЕНТОВ

Dedoc — универсальная открытая система для приведения документов к единому формату. Автоматически извлекает логическую структуру, таблицы и метаинформацию. Содержимое документов представляется в виде дерева, кодирующего заголовки и списки различного уровня вложенности. Dedoc может встраиваться как отдельный компонент в системы анализа структуры и содержимого документов.

ИСП РАН

ОСОБЕННОСТИ И ПРЕИМУЩЕСТВА

Dedoc реализован на языке Python. Работает со слабоструктурированными форматами данных (doc*, odt, xls*, csv, txt, json). Позволяет добавлять плагины, в частности, пакет расширений Docreader для работы с изображениями (png, jpg и др.), архивами (zip, rar и др.), pdf, html.Извлечение структуры документа проводится в полностью автоматическом режиме вне зависимости от типа входных данных, с извлечением метаинформации и разного вида форматирования текста.

DEDOC — ЭТО:

- Расширяемость за счет гибкого добавления поддержки новых форматов документов и простоты изменения выходного формата данных.
- Поддержка извлечения структуры вложенных документов различных форматов.
- Извлечение разного вида форматирования текста (отступы, шрифты, жирность, размер шрифта и др.).
- Добавление правил корректировки списков, неправильно составленных в документе (с опечатками).
- Извлечение табличной информации из xml формата doc*.

DOCREADER — ЭТО

- Работа с изображениями сканированных документов различного назначения (технические задания, нормативно-правовые акты, научные отчёты и статьи) и гибкая настройка под документы новой предметной области.
- Работа с pdf-документами — как с текстовым слоем, так и без.
- Распознавание физической структуры и текста ячеек сложных многостраничных таблиц с границами на изображениях с помощью методов контурного анализа; определение ориентации таблиц на изображении.
- Работа с активно развивающимся движком оптического распознавания символов OCR Tesseract компании Google в совокупности с использованием методов предварительной обработки изображений.
- Использование современных методов машинного обучения для определения ориентации документов и извлечения иерархической структуры на основе классификации строк извлеченных признаков из изображений документов.

ДЛЯ КОГО ПРЕДНАЗНАЧЕНА СИСТЕМА DEDOC?

- Разработчики прикладных систем анализа содержимого документов и документооборота.
- Разработчики интеллектуального анализа текста документов.
- Разработчики систем автоматической обработки текстов.

ПОДДЕРЖИВАЕМЫЕ ФОРМАТЫ

Русский и английский.

СХЕМА РАБОТЫ