OCR + OpenSource = ?

Вести с фронта открытых OCR решений

Что такое OCR?

Оптическое распознавание символов (англ. optical character recognition, OCR) — механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные, использующихся для представления символов в компьютере (например, в текстовом редакторе).

Задача

- Сохранить структуру документа
- Качество распознавания
- Распознавать только текст
- Экспорт в различные форматы
- Работа со сканерами
- И чтобы всё получалось само собой :-)

Сложность задачи

- Предобработка изображений
- Требуется хорошее железо для хорошего и БЫСТРОГО результата
- Сфера довольно сложна и наукоёмка

OCR движки

- Tesseract
- Cuneiform
- OCRAD

Tesseract

- Написан на C++
- Активно развивается(хоть и медленно)
- Обеспечивает очень крутое качество
- Имеет встроенную обработку изображений (плохую)
- Медленно работает
- Лучшее, что есть в Open Source

Cuneiform

- Написан на C/C++
- Написан русскими :-)
- Изначально коммерческий, потом выбросили в Open Source
- В целом слабее Tesseract по качеству
- Очень мёртв

OCRAD

- Написан на C++
- Разработан под эгидой GNU
- Довольно простой движок с ожидаемым для него качеством (хуже Tesseract)
- Скорее мёртв чем жив

OCR GUI

- gImageReader
- VietOCR
- OCRFeeder
- YAGF, UFOCR

gImageReader

- Написан на C++
- Два интерфейса: Qt и GTK
- Tesseract
- Пока что нет обработки изображений
- Импорт PDF, DJVU, множества форматов изображений. Экспорт в plain text, PDF, ODT
- Ведётся активная разработка

VietOCR

- Написан на Java
- Имеет обработку изображений (Leptonica)
- Имеет приятный интерфейс
- Заточен под вьетнамский язык
- Статус: неизвестно

OCRFeeder

- Написан на Python
- Приятный интерфейс
- Умеет экспорт в ODT
- Статус: мёртв

Мобильные решения

- TextFairy
- OpenNote Scanner

YAGF, UFOCR

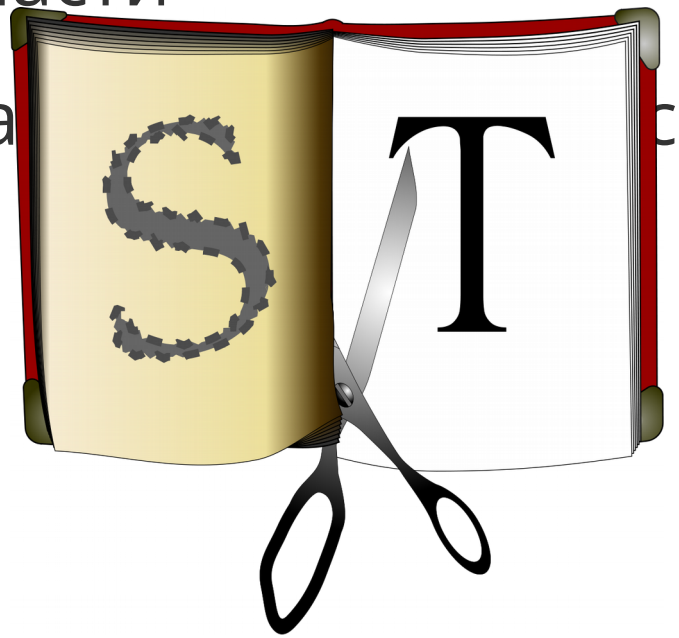
- Написан на C/C++
- Имеет интуитивный интерфейс
- Имеет обработку изображений (самописную и с кучей багов)
- Tesseract + Cuneiform
- Давно заброшен

Обработка изображений (приложения)

- ScanTailor
- Unpaper
- ImageMagick

ScanTailor

- Написан на C++
- Имеет GUI
- Умеет выравнивать изображения, делить изображения на части
- Обработка изобра сная
- Заброшен :(



Unpaper

- Написан на C
- Умеет в
- Все алгоритмы самописные
- Умер
- Можно почерпнуть порядок запуска алгоритма

ImageMagick

- Написан на C
- Имеет большое количество алгоритмов
- Проект с давней историей и проверенный временем (есть также форк GraphicsMagick)
- Использование: скрипты :-)
- Живее всех живых

Обработка изображений (библиотеки)

- OpenCV
- Leptonica
- Cimg
- Наколенные наброски с StackOverflow

OpenCV

- Написан на C++
- Предназначен для написания своих алгоритмов
- Из коробки есть неплохой набор уже готовых вещей (мало? OpenCV_contrib)
- Живой, и умирать не собирается

Leptonica

- Написан на C
- Предназначен именно как сборник полезных алгоритмов
- Имеет ряд уникальных алгоритмов
- Живой, но разработка идёт медленно

CImg

- Написана на C/C++
- Имеет неплохой набор уже готовых алгритмов
- Используется в GIMP

Исправление текста

- Встроенные вещи в OCR движки
- LanguageTool
- Hunspell, QtSpell, Aspell
- Написание своих средств

LanguageTool

- Написан на Java
- Поддерживает большое количество языков
- Умеет исправлять не только банальные опечатки, но и грамматические ошибки
- Standalone, LibreOffice, Chromium
- Активно развивается

Проприетарные решения

OCR

- Abbyy FineReader
- Omnipage

Image processing

- LEADTOOLS

Внимание!

Спасибо за внимание!