OCR mit Tesseract

Sven Hettwer

Software Engineer

12.01.2020

Agenda

- Sakuli OCR Demo
- Wer ist dieser Tesseract überhaupt und was kann der Typ?
- Tesseract im Node.js Umfeld
- Stärken und Schwächen von Tesseract
- Automatische Aufbereitung von Bildschirminhalten

Sakuli OCR Demo

Wer ist dieser Tesseract überhaupt und was kann der Typ?

- Software zur Texterkennung
- Geschrieben in C++
- Für mehr als 100 Sprachen und verschiedene Schriften
- Entwickelt zwischen 1984 und 1994 bei Hewlett-Packard
- Zielgeräte: Scanner mit OCR-Funktionalität
- Übernahme durch Google in 2005

https://de.wikipedia.org/w/index.php?title=Tesseract_(Software)&oldid=197191775

Wer ist dieser Tesseract überhaupt und was kann der Typ?

- Seit 2015 öffentlich entwickelt auf GitHub
- Seit 2016 Einsatz neuronaler Netze
- hOCR und alto Ausgabe
- Unter Linux verfügbar in allen gängigen package repistories
- CLI gibt’s auch!

```
tesseract imagename outputbase [-l lang] [--oem ocrenginemode] [--psm pagesegmode] [configfiles...]
```

https://de.wikipedia.org/w/index.php?title=Tesseract_(Software)&oldid=197191775

Tesseract im Node.js Umfeld

- tesseract.js
  - Emscripten port von Tesseract-OCR
  - Zuverlässigkeit bei Erkennungen deutlich unter Tesseract-OCR
  - Cool für web-apps mit „on the fly“ oder „realtime“ OCR
  - Performance gefühlt leicht schlechter – Suchen dauern länger
  - Wird maintained
- Alles andere (was wir gefunden haben)
  - CLI-Wrapper für Tesseract-OCR
  - D.h. inkl. hard dependency auf Tesseract-OCR

https://www.npmjs.com/search?q=tesseract

Stärken und Schwächen von Tesseract

- Disclaimer: Nur unsere Erfahrungen!
- Stärken:
  - Texte mit hohen Kontrasten
  - Texte mit klarem, strukturiertem Textverlauf (z.B. Bücher)
  - Texte mit klarer Schrift
- Hilfreich sind:
  - Bilder mit hoher DPI
  - Bilder mit Meta-Informationen
  - Bilder mit klarem Rand / Abgrenzung des Texts

Stärken und Schwächen von Tesseract

- Schwächen:
  - Unstrukturierte Texte
  - Texte mit kleiner Schriftgröße
  - Schlechte Kontraste
- Erkennungsgenauigkeit sinkt wenn mehr als zwei „Schwächen“ zusammen kommen
- Hinderlich sind:
  - Fehlende Metainformationen

Automatische Aufbereitung von Bildschirminhalten

- Vorteile:
  - Erhöht „unter Umständen“™ die Genauigkeit
- Nachteile
  - Kostet Ressourcen
  - Mehrere Analysen durch Tesseract sind notwendig
- Alternativ mit „page segmentation mode“ experimentieren
- Lohnt sich auf jeden Fall, wenn:
  - Typen von Dokumenten bekannt sind
  - Große Mengen von Seiten mit den selben Charakteristiken zu digitalisieren sind

Vielen Dank!

ConSol
Consulting & Solutions Software GmbH

St.-Cajetan-Straße 43
D-81669 München
Tel.: +49-89-45841-100
info@consol.de
www.consol.de
Twitter: @consol_de