Texterkennung

Die SQUEEZE Texterkennung basiert auf der OCR Engine Tesseract. Tesseract wurde ausgewählt, da es unheimlich viele Konfigurationsmöglichkeiten bietet. SQUEEZE wurde so auf Tesseract abgestimmt, dass die bestmöglichen Ergebnisse für die meisten maschinell erstellten Dokumente erzielt werden.

Texterkennung bei sehr großen Dokumenten

SQUEEZE ist in der Lage auch mit mehreren hundert Seiten umzugehen. In diversen Tests wurden Dokumente mit mehr als 700 Seiten verarbeitet und analysiert. Dabei sind jedoch lange Laufzeiten zu berücksichtigen. Bei einer konservativen Berechnung gehen wir von 15 Sekunden je Seite aus. Nimmt man also ein Dokument mit 700 Seiten, so ist mit einer Verarbeitungszeit von 700 Seiten x 15 Sekunden zu rechnen. Das entspricht einer Verarbeitungszeit von knapp 3 Stunden.

Alternative OCR Engines

SQUEEZE unterstützt zusätzlich zu Tesseract weitere Lizenzpflichtige OCR Engines. Dazu gehören die **Abbyy** und **IRIS** OCR Engines. Sollte der Bedarf bestehen eine dieser alternativen Engines nutzen zu wollen, so ist dies natürlich möglich führt jedoch zu höheren Kosten, da die Engines kostenpflichtig sind.

Texterkennung bei PDFs mit Volltext (native PDFs)

Da die Bereitstellung von Rechnungen per E-Mail zum Standard geworden ist, wird die Texterkennung mittels OCR immer seltener erforderlich. Die per E-Mail bereitgestellten Dokumente beinhalten häufig "native" PDFs, die den Volltext bereits enthalten. Dieser Text kann von SQUEEZE ebenfalls verarbeitet werden. In diesem Fall wird keine Texterkennung mittels OCR ausgeführt, der Text des PDFs wird stattdessen analysiert und ausgewertet. Dabei bleiben die geometrischen Informationen der Texte erhalten und können so auch im Viewer markiert und selektiert werden.

Es gibt eine Besonderheit die zu Berücksichtigen ist, bei gescannten Dokumenten, welche bereits einen Volltext enthalten, welche von einer anderen OCR Engine erstellt wurden. Bei diversen Tests wurde festgestellt, dass die genutzten OCR Engines nicht immer ein zufriedenstellendes Ergebnis liefern. Aus diesem Grund wird bei gescannten Dokumenten immer der Volltext verworfen und durch SQUEEZE und Tesseract neu erstellt. Nur so kann eine einheitliche Qualität der OCR gewährleistet werden.

Revision #3
Created 2020-03-04 19:18:30 UTC by Phillip Langer
Updated 2020-03-10 09:43:22 UTC by Jasmin Ruß