

Stapelklasseneigenschaften

OCR-Texterkennung

Allgemeines

Die OCR Extraktion ist ein elementarer Teil der Squeeze Software. Dieser Kernbereich der Software ist mit verschiedenen Einstellungen versehen, die das Ergebnis der Extraktion tangieren. Im folgenden Artikel gehen wir auf die Besonderheiten und die Anforderungen der unterschiedlichen Eigenschaften ein.

Welche Arten der OCR unterstützt Squeeze?

Grundsätzlich unterscheiden wir im Kontext von Squeeze zwischen dem Einsatz einer OCR basierend auf den Ressourcen der lokalen Maschine und dem Einsatz eines Remote-OCR-Dienstes.

Was beinhaltet meine Standardversion von Squeeze?

Im Auslieferungszustand ist Squeeze mit einer lokal verfügbaren OCR-Engine ausgestattet. Auf Kundenwunsch können unsere Berater bei einer Squeeze Installation ab der Version 2.4 eine Remote-OCR aktivieren, die mithilfe von AI bessere Ergebnisse liefern kann.

Allgemeine Stapelklassen-Eigenschaften

OCREngine (ab Squeeze 2.4)

Wird diese Stapelklassen-Eigenschaft nicht konfiguriert greift automatisch die lokale OCR-Engine `|ocrmypdf|`.

Je nach Spezifikation und Lizenzierung ihres Squeeze-Systems können folgende Optionen für die OCREngine verwendet werden:

Squeeze Version	Optionen
ab 2.4.0	<code> default </code>

ab 2.4.0	ai-ocr
ab 2.5.0	maxocr
ab 2.6.0	proxy-ocr

Voraussetzungen:

- default:
 - keine
- ai-ocr:
 - um die Remote-AI-OCR zu verwenden ist es notwendig dass eine Internetverbindung auf dem System existiert und dass die Anmeldedaten von Ihrem Squeeze Berater konfiguriert werden.
- maxocr
 - die konfigurierte Mandanten-Konfiguration/Server-Konfiguration für die [Dexpro Platform Integration](#).
 - die [MaxOCR Konfiguration](#).
- proxy-ocr
 - die [KI Proxy Konfiguration](#)

Stapelklassen-Eigenschaften für die lokale OCR Engine

OCRForce

Im Standard wird bei digitalen PDF's der Textlayer genutzt und die Felderkennung darauf angewendet (|false|). Um aber eine OCR zu erzwingen ist dieser Schalter auf |true| zu setzen.

OCRLanguage

Im Standard werden die Sprachpakete Deutsch und Englisch verwendet. Für die deutsche Detektion wird der Wert |deu| eingetragen und für die englische Detektion der Wert |eng| eingetragen.

Hier können projektspezifisch auch weitere Sprachen oder abgewandelte Sprachpaket-Varianten angegeben werden, bei denen die OCR schneller/langsamer bzw. mit niedriger/höherer Qualität Ergebnisse liefert. Im folgenden eine Übersicht über die im Standard enthaltenen Sprachpakete:

Squeeze Version	Optionen
-----------------	----------

vor 2.4.0	<ul style="list-style-type: none"> • deu • eng
ab 2.4.0	<ul style="list-style-type: none"> • deu, deu_best, deu_fast, deu_std • lat_best, lat_fast, lat_std • eng • osd

OCRPageLimit

Anzahl der auszulesenden Seiten im Dokument. Syntax n-m

Beispiel für Auslesung der ersten 3 Seiten: 1-3

PDFA-Conversion

Es wird ein PDFA kompatibles Dokument erzeugt. Eingabe 1|0 (`|true|`||`false|`)

PDFProcessor

Hier gilt `PDFBox` als Standard. `PDFMiner` ist die Alternative .

PSM-Modes

Im Project bietet es sich an, die Modi 3, 4, 6 und 11 zu verwenden. Dabei gilt 3 als Standard.

3	Standardeinstellung liefert gute Ergebnisse.
4	Wortweise Segmentierung. Es wird nicht nach Zeilen geschaut sondern Worten. (verfügbar ab Version 2.0)
6	Gut für Positionsdaten. Hat aber Probleme bei Linien die sehr dicht am Text sind.
11	Gut bei vielen Grafiken auf den Dokumenten.

OCRRotationThreshold

Mit dieser Eigenschaft können Sie beeinflussen wie *agressiv* Seiten in der OCR gedreht werden. Nutzen Sie diesen Wert, wenn Dokumente falsch gedreht werden.

Geringe Werte führen dazu, dass mehr Dokumente gedreht werden. Die Software muss sich also nicht sehr sicher sein, dass eine Seite rotiert werden muss.

Hohe Werte führen dazu, dass Dokumente seltener gedreht werden, also nur wenn sich die Software sehr sicher ist, dass eine Seite rotiert werden muss.

Im Standard ist dieser Wert `9.0`

Stapelklassen-Eigenschaften für die Remote-AI-OCR/MaxOCR/KI-Proxy Engine

Aktuell gibt es keine Möglichkeiten die Remote-AI-OCR zu beeinflussen.

Fragen und Antworten?

1. Ich habe die `ai-ocr/maxocr` als OCREngine Eigenschaft ausgewählt, jedoch funktioniert die Texterkennung nicht mehr ?
 - Gehen Sie bitte Sicher das Ihr Squeeze Berater die notwendigen Anmeldedaten zur Aktivierung der Remote OCR hinterlegt hat.
2. Ich habe mit der Remote-AI-OCR ein Dokument verarbeitet, mehrere Dokumente liefen erfolgreich durch, jedoch bleibt dieses Dokument hängen.
 - Aufgrund der begrenzten Ressourcen kann die AI-Remote-OCR maximal 100 Seiten pro Dokument verarbeiten. Überprüfen Sie daher die Anzahl der Seiten und nutzen bei nicht erfolgreicher Verarbeitung die lokale OCR.
3. Ich nutze die Remote-AI-OCR und mein Dokument hat mehrere Seiten jedoch nicht mehr als 100 Seiten trotzdem hängt das Dokument in der Verarbeitungskette fest.
 - Squeeze wartet insgesamt 3 Minuten auf die Verarbeitung des Dokumentes. Konnte der entfernte Dienst innerhalb dieser 3 Minuten das Dokument nicht verarbeiten, wird Squeeze eine Fehlermeldung mit einem Timeout Hinweis liefern. Schieben Sie das Dokument erneut über die technische Warteschlange in den Schritt "Texterkennung" Squeeze prüft in dem Fall ob das bereits hochgeladene Dokument verarbeitet wurde.

Revision #36

Created 28 February 2023 12:25:53 by Vahdettin Balum

Updated 21 August 2024 09:47:50 by Fabian Terstegen