Beiträge zur Automatischen Texterkennung Universitätsbibliothek Mannheim

Weiterentwicklung.
Bereitstellung.
Unterstützung.

Bericht Pilotbibliothek Universitätsbibliothek Mannheim

Weiterentwicklung.

DFG

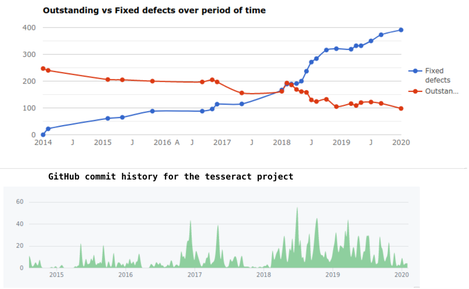
Im Rahmen des DFG-geförderten Modulprojektes *Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow* wurde die Texterkennungssoftware Tesseract verbessert und weiterentwickelt. Neben der Bereitstellung von Schnittstellen für andere Modulprojekte (Zeichenalternativen) sind Optimierungen in den Bereichen Performance, Codequalität und genauere Zeichenkoordinaten erzielt worden.

Tesseract findet in acht OCR-D-Modulen Anwendung:
ocrd-tesserocr-deskew (Neigungskorrektur)
ocrd-tesserocr-crop (Freistellen von Regionen)
ocrd-tesserocr-binarize (Binarisierung)
ocrd-tesserocr-segment-region (Segmentierung)
ocrd-tesserocr-segment-table (Segmentierung)
ocrd-tesserocr-segment-line (Segmentierung)
ocrd-tesserocr-segment-word (Segmentierung)
ocrd-tesserocr-recognize (Texterkennung)

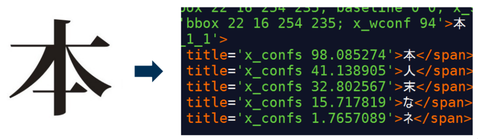
Die UB Mannheim hat diese Module nicht selbst entwickelt, aber kleinere Beiträge dazu geleistet.

Tesseract im Modulprojekt

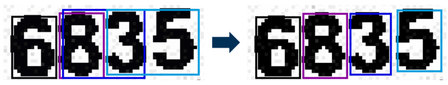
- **Verbesserung der Performance**
- **Steigerung der Codequalität**



- **Ausgabe von Zeichenalternativen**



- **Genauere Zeichenkoordinaten**



Bereitstellung.

Neue Modelle für Alte Drucke

Die an der Universitätsbibliothek Mannheim neu trainierten Modelle für Alte Drucke erkennen deutlich mehr Text fehlerfrei als bisher. Darüber hinaus haben wir neue Erkenntnisse gewonnen, die uns helfen, das Training noch weiter zu optimieren und Empfehlungen für andere Nutzer zu formulieren.
Mit dem *Tesseract Installer for Windows* der UB Mannheim können Windows-Anwender direkt eigene Versuche starten.

Unterstützung.

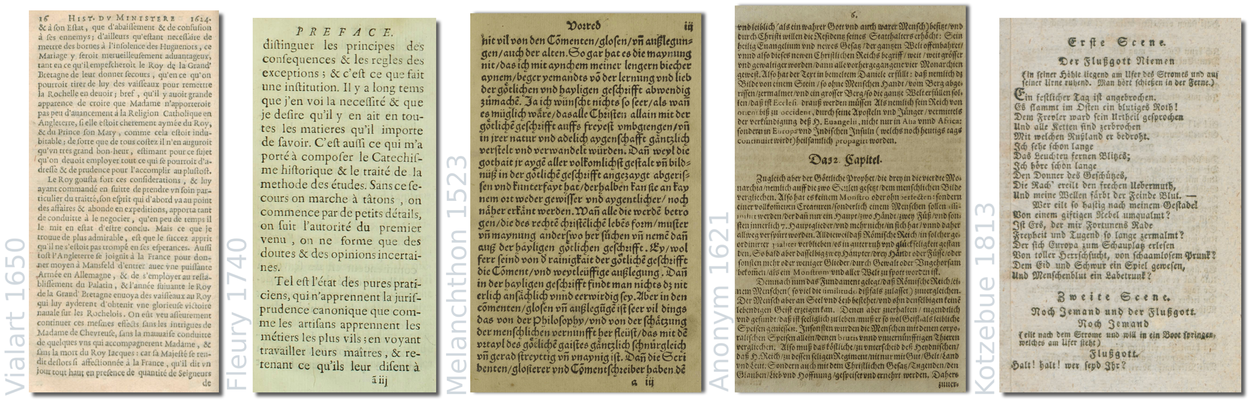
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

OCR für Alle!.. GLAMs im Land

Im MWK-geförderten Projekt *OCR-BW* unterstützen die Universitätsbibliotheken Mannheim und Tübingen Archive, Bibliotheken und andere Institutionen in Baden-Württemberg bei der Anwendung von Texterkennungs- und Transkriptionssoftware. Dabei liegt der Schwerpunkt der UB Mannheim auf der automatischen Texterkennung von Druckwerken, während die UB Tübingen die Transkription von Handschriften im Fokus hat. Die UB Mannheim setzt dabei auch die im OCR-D-Projekt enwickelten Module ein.

Ausgangsdaten.

Druckwerke in Antiqua und Fraktur aus dem 16. – 19. Jahrhundert



Antiqua 17. Jhd. | **Antiqua 18. Jhd.** | **Fraktur 16. Jhd.** | **Fraktur 17. Jhd.** | **Fraktur 19. Jhd.**

Als Pilotbibliothek hat die UB Mannheim fünf ausgewählte Werke aus ihrem Bestand (16. bis 19. Jahrhundert) mit den bei OCR-D entwickelten Tools bearbeitet und die Ergebnisse mit den bereits vorhandenen Texten verglichen.

Workflow und Ergebnisse.

Der folgende Workflow wurde auf die Druckwerke angewendet, wobei nicht immer alle Schritte verwendet wurden:

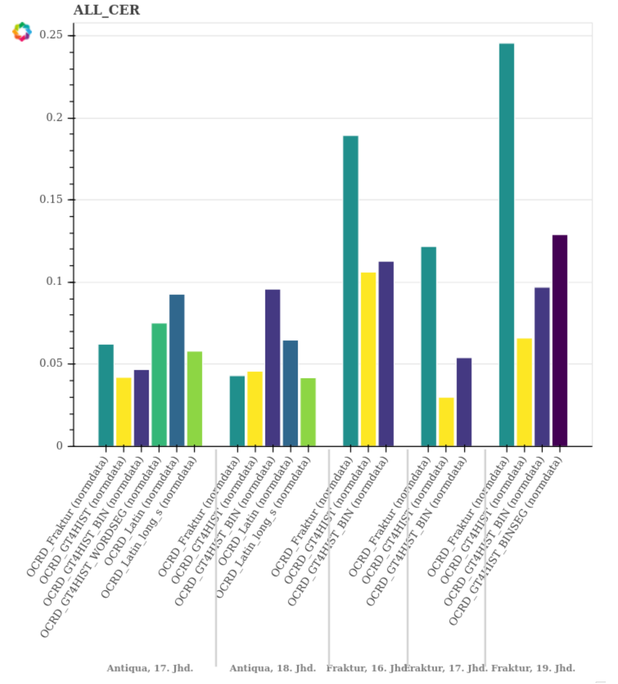
- ocrd-cis-ocropy-binarize (Label: BIN)
- ocrd-tesserocr-segment-region (Ø-Zeitbedarf: 4 s)
- ocrd-tesserocr-segment-line (Ø-Zeitbedarf: 2 s)
- ocrd-tesserocr-segment-word (Label: WORDSEG)
- ocrd-tesserocr-recognize (Ø-Zeitbedarf: 6 s)

Ø-Zeitbedarf: 12 Sekunden pro Seite (ohne BIN und WORDSEG)

Verwendete Texterkennungsmodelle:
- script/Fraktur (Fast)
- GT4HIST (Best)
- script/Latin (Best)

Evaluierung;
- dinglehopper

Die Ergebnisse wurden normalisiert und mit der ebenfalls normalisierten Ground Truth verglichen.



UNIVERSITÄT MANNHEIM

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

DFG