DFG-Projekt OCR-D

Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow

Neue Frakturmodelle für Tesseract

Stefan Weil
Universitätsbibliothek Mannheim

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

18.11.2019

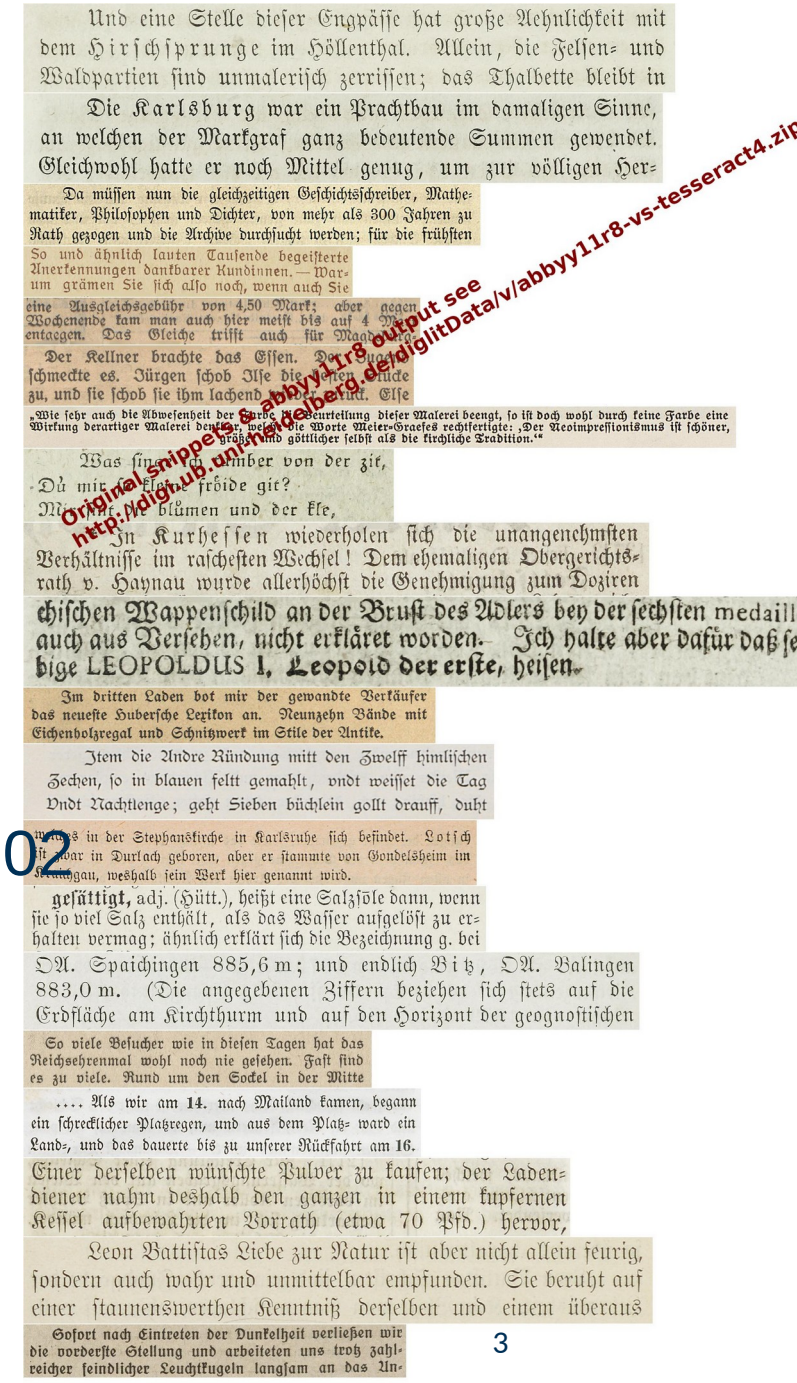
Tesseract OCR

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source
- Komplettlösung „All-in-1“
- Mehr als 100 Sprachen / mehr als 30 Schriften
- Liest Bilder in allen gängigen Formaten (nicht PDF!)
- Erzeugt Text, PDF, hOCR, ALTO, TSV
- Große, weltweite Anwender-Community
- Technologisch aktuell (Texterkennung mit neuronalem Netz)
- Aktive Weiterentwicklung u. a. im DFG-Projekt OCR-D
- OCR-D-Module für Binarisierung, Segmentierung, Texterkennung und weitere

Tesseract Fraktur OCR 2018

- Drei Modelle zur Auswahl: deu_frak, frk, Fraktur
- 14. Juni 2018: Tesseract schneidet im Frakturtest von Jochen Barth (UB Heidelberg) im Vergleich zu ABBYY FineReader schlecht ab: https://github.com/tesseract-ocr/tessdata/issues/102
- Ligaturen wie ch und ck werden häufig als Kleiner- bzw. Größerzeichen erkannt und weitere systematische Erkennungsfehler

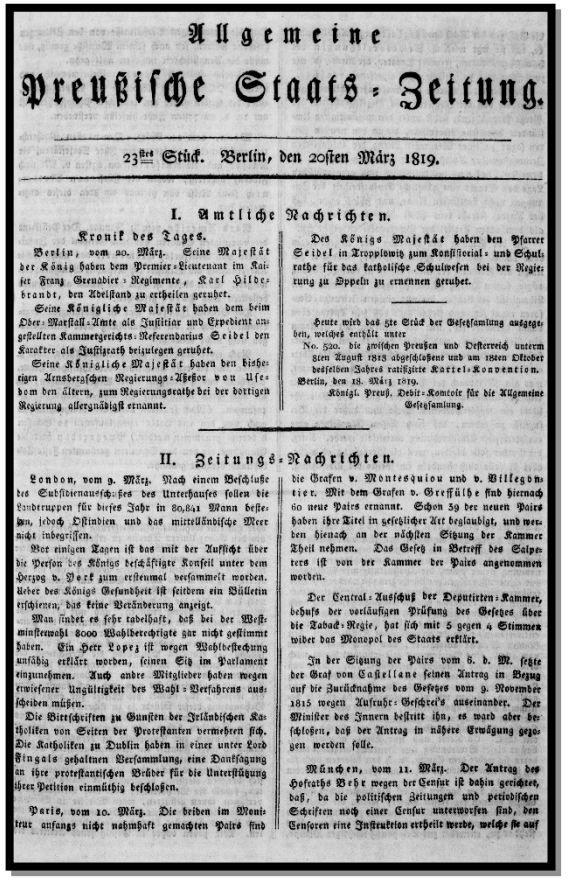


Tesseract Fraktur OCR 2019

- 16. November 2019: Tesseract (grün) schlägt ABBYY (rot) in Jochen Barths neuem Test deutlich (trotz noch vorhandener Schwächen in der Zeilenerkennung): https://digi.ub.uni-heidelberg.de/diglitData/v/abby12r3-OldGerman-vs-Fraktur5000000_0.466_II.png
- Wichtigste Änderung ist dabei ein neues Frakturmodell.

Tesseract Fraktur-Modelle der UB Mannheim

- UB Mannheim benötigt Fraktur OCR u. a. für die historische Zeitung *Deutscher Reichsanzeiger und Preußischer Staatsanzeiger* (ursprünglich *Allgemeine Preußische Staatszeitung*)
- Training neuer Fraktur-Modelle aus publizierten Ground-Truth-Daten seit September 2019:
  - GT4HistOCR (Springmann et al.)
  - Austrian Newspapers (Mühlberger et al.)
- Zeichenerkennung besser als 97,5 %, Worterkennung bei 93 % für Austrian Newspapers



Allgemeine
Preußische Staats-Zeitung.

23stes Stück. Berlin, den 20sten März 1819.

I. Amtliche Nachrichten.

Kronik des Tages.

Berlin, vom 20. März. Seine Majestät der König haben dem Premier-Lieutenant im Kaiser Franz Grenadier-Regimente, Karl Hildebrandt, den Adelstand zu ertheilen geruhet.

Seine Königliche Majestät haben dem beim Ober-Marschall-Amte als Justitiar und Expedient angestellten Kammergerichts-Referendarius Seidel den Karakter als Justizrath beizulegen geruhet.

Seine Königliche Majestät haben den bisherigen Arnsbergschen Regierungs-Assessor von Usedom den Ältern, zum Regierungsrathe bei der dortigen Regierung allergnädigst ernannt.

Des Königs Majestät haben den Pfarrer Seidel in Tropplowitz zum Konsistorial- und Schulrathe für das katholische Schulwesen bei der Regierung zu Oppeln zu ernennen geruhet.

Heute wird das 22te Stück der Gesetzsammlung ausgegeben, welches enthält unter
No. 520. die zwischen Preußen und Oesterreich unterm 21ten August 1818 abgeschlossene und am 18ten Oktober desselben Jahres ratifizirte Kartel-Konvention.
Berlin, den 18. März 1819.
Königl. Preuß. Debit-Komtoir für die Allgemeine Gesetzsammlung.

II. Zeitungs-Nachrichten.

London, vom 9. März. Nach einem Beschluße des Subsidienausschusses des Unterhauses sollen die Landtruppen für dieses Jahr in 80,841 Mann bestehen, jedoch Ostindien und das mittelländische Meer nicht inbegriffen.

Vor einigen Tagen ist das mit der Aufsicht über die Person des Königs beschäftigte Konseil unter dem Herzog v. York zum erstenmal versammelt worden. Ueber des Königs Gesundheit ist seitdem ein Bülletin erschienen, das keine Veränderung anzeigt.

Man findet es sehr tadelhaft, daß bei der Westminsterwahl 8000 Wahlberechtigte gar nicht gestimmt haben. Ein Herr Lopez ist wegen Wahlbestechung unfähig erklärt worden, seinen Sitz im Parlament einzunehmen. Auch andre Mitglieder haben wegen erwiesener Ungültigkeit des Wahl-Verfahrens ausscheiden müssen.

Die Bittschriften zu Gunsten der Irländischen Katholiken von Seiten der Protestanten vermehren sich. Die Katholiken zu Dublin haben in einer unter Lord Fingals gehaltnen Versammlung, eine Danksagung an ihre protestantischen Brüder für die Unterstützung ihrer Petition einmüthig beschlossen.

Paris, vom 10. März. Die beiden im Moniteur anfangs nicht nahmhaft gemachten Pairs sind die Grafen v. Montesquiou und v. Villegontier. Mit dem Grafen v. Greffülhe sind hiernach 60 neue Pairs ernannt. Schon 59 der neuen Pairs haben ihre Titel in gesetzlicher Art beglaubigt, und werden hienach an der nächsten Sitzung der Kammer Theil nehmen. Das Gesetz in Betreff des Salpeters ist von der Kammer der Pairs angenommen worden.

Der Central-Ausschuß der Deputirten-Kammer, behufs der vorläufigen Prüfung des Gesetzes über die Tabak-Regie, hat sich mit 5 gegen 4 Stimmen wider das Monopol des Staats erklärt.

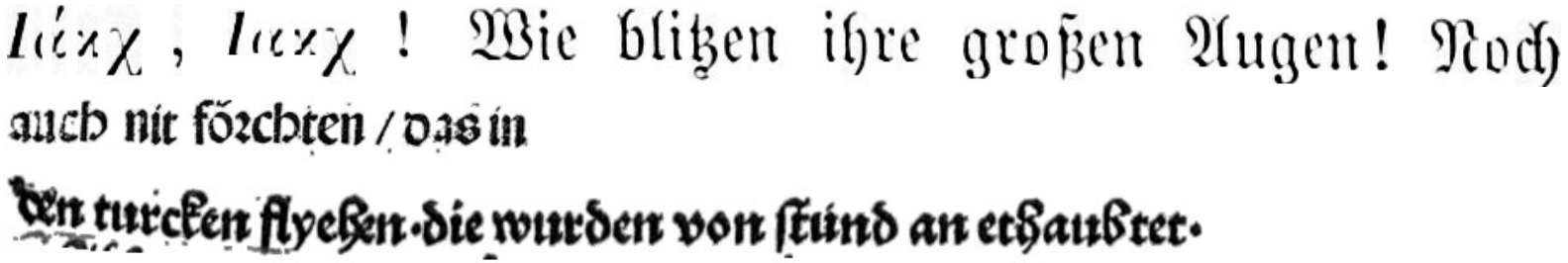
In der Sitzung der Pairs vom 6. d. M. setzte der Graf von Castellane seinen Antrag in Bezug auf die Zurücknahme des Gesetzes vom 9. November 1815 wegen Aufruhr-Geschrei's auseinander. Der Minister des Innern bestritt ihn, es ward aber beschlossen, daß der Antrag in nähere Erwägung gezogen werden solle.

München, vom 11. März. Der Antrag des Hofraths Behr wegen der Censur ist dahin gerichtet, daß, da die politischen Zeitungen und periodischen Schriften noch einer Censur unterworfen sind, den Censoren eine Instruktion ertheilt werde, welche sie auf

Tesseract Fraktur-Training

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Beispiele für Zeilen aus GT4HistOCR



Beispiele für Zeilen aus Austrian Newspapers

Literatur

- Springmann, U., Reul, Ch., Dipper, S., & Baiter, J. (2018). GT4HistOCR: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin (Version 1.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1344132
- Mühlberger, G. & Hackl, G. (2019). NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3387369
- Weil, S. (2019). Training Fraktur. GitHub. https://github.com/tesseract-ocr/tesstrain/wiki
- Metzger, N. & Weil, S. (2019). Optimierter Einsatz von OCR-Verfahren - Tesseract als Komponente im OCR-D-Workflow. https://nbn-resolving.org/urn:nbn:de:bsz:180-madoc-522132
- Weil, S. (2019). Vom Bild zum Text. Automatisierte Texterkennung in historischen Drucken mit der freien Software Tesseract. https://nbn-resolving.org/urn:nbn:de:0290-opus4-163511
- Weil, S., & Zumstein, P. (2016). Mit freier Software Text in Digitalisaten erkennen. https://speakerdeck.com/zuphilip/mit-freier-software-text-in-digitalisaten-erkennen-ocr-praxis-an-der-ub-mannheim

Bildquellen

- Titelseite:
  https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126884/
  https://pixabay.com/de/vectors/flach-design-symbol-icon-www-2126880/
  https://pixabay.com/de/vectors/werkzeug-schraubenschl%C3%BCssel-3456474/
- Abbildungen von Jochen Barth, UB Heidelberg:
  https://digi.ub.uni-heidelberg.de/diglitData/v/abby12r3-OldGerman-vs-Fraktur5000000_0.466_II.png
  https://camo.githubusercontent.com/47ac160cf86bb8f69fe98677112e5597a46da3ed/687474703a2f2f646967692e75622e756e692d68656964656c626572672e64652f6469676c697444617461612f762f6162262279793131723382d76732d7465737365726374342e6a7067
- DFG-Logo: https://www.dfg.de/
- GitHub Logos: https://github.com/logos
- OCR-D Logo: http://www.ocr-d.de/