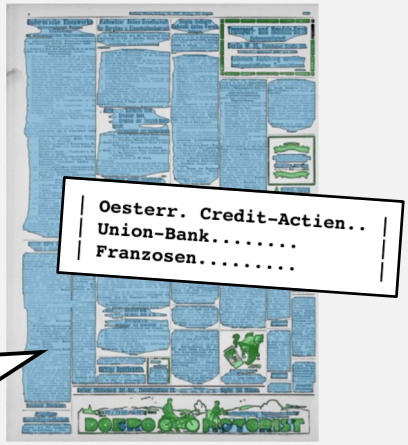
Origami – Eine OCR-Pipeline zur Erschließung historischer Zeitungen

Was wir haben:

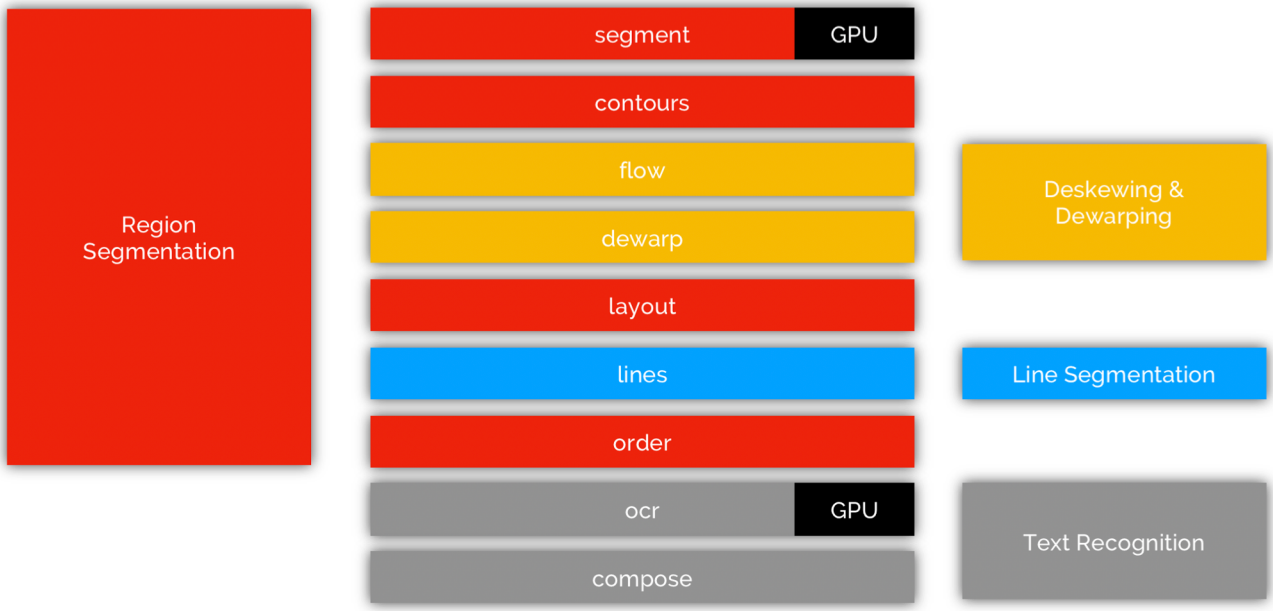
- Berliner Börsen-Zeitung (BBZ)
- Viele unterschiedliche Schriftarten
- Komplexes Layout (Spalten, Tabellen, Werbeanzeigen)

Was wir wollen:



- Segmentierung der Zeitung
- Identifikation von Textspalten
- Transkription der Texte (OCR)

Origami OCR-Pipeline



https://github.com/poke1024/origami

Liebl, B. & Burghardt, M. (2020). From Historical Newspapers to Machine-Readable Data: The Origami OCR Pipeline. Proceedings of the 1st Workshop on Computational Humanities Research (CHR).

Liebl, B. & Burghardt, M. (2020). An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers. 25th International Conference on Pattern Recognition, Mailand. (Preprint https://arxiv.org/abs/2004.07317)

BERNHARD LIEBL, MANUEL BURGHARDT

COMPUTATIONAL HUMANITIES GROUP

UNIVERSITÄT LEIPZIG