End-to-End Autoregressive Pixel Models of Generation in Text Space

- **Supervisor**: Yihong Liu
- **Examiner**: Prof. Hinrich Schütze
- **Open**: MSc
- **Summary**: Recent work shows that instead of modeling languages in text space, by rendering the text to images, models that operate on pixel representations can also offer attractive performance [1, 2]. Those pixel-based models address the vocabulary bottleneck issue, i.e., a trade-off in balancing input granularity against computational feasibility [1]. However, one major limitation of pixel-based models is that they are not good at generation. Some recent studies apply autoregressive-style training, i.e., next-batch prediction [3, 4], to train a decoder-like pixel-based model that autoregressively generates text in image patches. Nevertheless, the final generated patch has to be transformed into text space by the OCR system, which is inherently limited by the accuracy of text extraction. This research wants to explore the following research question: *can we model the text in pixel space but the generation is done in text space?* The benefit of this setup is that it does not require any tokenizer at the input space and the model is robust to different fonts. For the generation, since the model directly generates output in text space, OCR is not needed.

- **Prerequisites**: enthusiasm, good programing background (preferably python), good knowledge of NLP, a good command of DL framework (preferably PyTorch)

[1] https://arxiv.org/abs/2207.06991
[2] https://arxiv.org/abs/2310.18343
[3] https://arxiv.org/abs/2401.03321
[4] https://arxiv.org/abs/2404.10710