Adapt multilingual PLMs to new languages? Scripts can matter.

- **Supervisor**: Yihong Liu
- **Examiner**: Prof. Schütze
- **Msc, Open**: MSc
- **Summary**: Pretrained language models (PLMs), e.g., BERT, have boosted the development of various NLP applications. However, many of those NLP applications are only restricted to high-resource languages like English and German. There are more than 7,000 languages in the world, so then multilingual PLMs become a rather popular topic after PLMs are introduced. Compared with PLMs for a single language, multilingual PLMs, like mBERT, mBART, XLM-R are trained on monolingual corpora of multiple languages. Nevertheless, the languages covered by those multilingual PLMs are still far behind the number of languages used in the world. Therefore, transfer learning, i.e., adapting the knowledge encoded in multilingual PLMs to unseen languages, can often reach good performance in those unseen languages. As languages can be written in different scripts or writing systems (there can be multiple scripts commonly used by some languages,), a natural question would be: can the script (if there are multiple of them available) of the unseen languages influence the performance of this transfer learning? If the answer is yes, can we further ask a question: which script is the most helpful if we are able to transliterate a language into multiple scripts? And most importantly, could the process of choosing the best script be done automatically?

- **Prerequisites**: enthusiasm (publishing the results of the thesis at a conference/workshop), good programing background (preferably python), good knowledge of NLP, a good command of DL framework (preferably PyTorch)