How Multilingual are Existing Multilingual Benchmark Datasets? A Systematic Investigation

- **Supervisor**: Yihong Liu
- **Examiner**: Prof. Hinrich Schütze
- **Open**: BSc/MSc
- **Summary**: Multilingual benchmark datasets such as XNLI [1], Tatoeba [2], and mLAMA enables the evaluation of crosslinguality of multilingual pretrained language models (mPLMs). However, most of the benchmark datasets only support a limited number of languages, which cannot well represent the whole picture of the languages spoken in the world (more than 7,000). In addition, a thorough analysis and comparison of existing benchmark datasets are missing in the community. To this end, in this project, we want to conduct a systematic investigation of the multilingual benchmark datasets commonly used in the research community. We are especially interested in the metadata of those datasets, e.g., size of the dataset, type of the evaluation, number of languages covered, number of language families covered, and so on. In this way, we can provide systematic analysis and comparison from different dimensions, hoping to shed light on a better understanding of the multilinguality when selecting specific benchmark datasets for evaluation.
- **Prerequisites**: enthusiasm, good ability to conduct a literature review, good knowledge of NLP, good programming background (preferably Python)

[1] https://arxiv.org/pdf/1809.05053.pdf
[2] https://tatoeba.org/
[3] https://arxiv.org/pdf/2102.00894.pdf