Prioritized Training on Worth-Learning Samples with Your Pretrained Model

- **Supervisor**: Yihong Liu
- **Examiner**: Prof. Hinrich Schütze
- **Open**: MSc
- **Summary**: The currently advanced pipeline is to pretrain a language model first and then fine-tune it on task-specific data. Although the fine-tuning step consumes much less time or resources compared with the pretraining step, we might be able to further save resources and accelerate the fine-tuning by carefully designing the sampling strategy, since not all data points are equally important to the fine-tuning (the model can already well-learn some samples in the pretraining phase, or some samples are noisy or not learnable at all) [1, 2]. In addition, the order of instances being sampled for training is also important, as examples vary greatly in difficulty [3]. To this end, this project aims to design a tractable and wise sampling strategy for fine-tuning pretrained language models (PLMs) on task-specific data, similar to Curriculum Learning [4]. The expected outcome is that the proposed strategy could (1) accelerate the training and (2) achieve better performance, compared with the standard uniformly random sampling.
-
- **Prerequisites**: enthusiasm, good programing background (preferably python), good knowledge of NLP, a good command of DL framework (preferably PyTorch)

[1] https://proceedings.mlr.press/v162/mindermann22a/mindermann22a.pdf
[2] https://aclanthology.org/2023.emnlp-main.125.pdf
[3] https://aclanthology.org/2020.acl-main.542.pdf
[4] https://arxiv.org/pdf/2101.10382.pdf