Chapter Segmentation In German Literary Texts

In this thesis, you will separate German novels into their chapters. The segmentation will be performed given a novel‘s text and (optionally) the number of chapters. While chapter separation has already been performed on English data [1], neither convenient German datasets nor ready-made implementations exist.

Your task is to first create a dataset and, in a second step, evaluate at first a baseline model and later a more advanced model-based, for example, on a transformer architecture.

Approach:

- Build a chapter dataset based on the d-Prose dataset [2]
  - Extract gold chapter separations based on projekt-gutenberg [3] and other data sources
  - Remove such surface-level identifications as chapter headings and whitespace
- Build baseline model, e.g. based on word co-occurrence [4]
- Evaluate further approaches e.g. BERT-based

Suggested links:
- [1] https://www.aclweb.org/anthology/2020.emnlp-main.672/
- [2] https://zenodo.org/record/4315209
- [3] https://www.projekt-gutenberg.org/
- [4] https://arxiv.org/abs/cmp-lg/9406017
- https://github.com/licsth/Gutenberg-Projekt

Suitable for a Bachelor‘s Thesis (but could be adapted to a Master’s Thesis as well)

Contact Person: Hans Ole Hatzel