SCRC: SWISS COURT RULING CORPUS

A PREPRINT

Joel Niklaus
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
`joel.niklaus@inf.unibe.ch`

Matthias Stürmer
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
`matthias.stuermer@inf.unibe.ch`

March 23, 2021

ABSTRACT

Researchers and Practitioners in Natural Language Processing (NLP), like in Data Science in general, rely on readily available curated datasets for their work.

While there exist datasets of court decisions in Germany and France, to the best of our knowledge, there are none available for Switzerland yet. Since legal systems differ from country to country, the corresponding legal languages vary considerably as well. Therefore, for legal NLP in Switzerland to be successful, Swiss legal datasets are crucial. Whilst platforms providing access for lawyers to legal text are available, they are not suitable for text mining applications because a) they normally do not offer bulk data access and b) because they normally offer their data in pdf or html and not in raw text format.

The corpus presented in this work offers a solution to the problems mentioned above, in that it provides fast bulk access to cleaned text files of over 500'000 Swiss court decisions. This dataset is the first step in legal NLP research in Switzerland. (too ballsy?) This dataset can be used for a wide variety of applications: Modern pretrained language models rely on huge amounts of raw text (preferably inside the domain of application). This corpus is ideally suitable for pretraining these language models to boost performance on Swiss legal applications. Text classification is an NLP task with a large number of impactful applications like e.g. spam filtering and sentiment analysis in general or e.g. legal judgement prediction (predict outcome of a case based on description of case's facts) and legal area prediction in the legal domain. The rich metadata of this corpus (containing e.g. date, canton, court and chamber) with the additional possibility to extract information from the text (e.g. judge, lawyers, legal area, ruling outcome and cited laws) provide natural labels for many text classification tasks.

The data has been downloaded from entscheidsuche.ch, a Swiss association committed to provide easy and free access to Swiss court decisions for lawyers. It has been cleaned by removing uninteresting artifacts such as headers and footers, page numbers and characters left over from automatic PDF-to-text conversion. The entire codebase is available on GitHub[1] for easy reproduction and re-running to get the current data from entscheidsuche.ch if needed. The German part of the data is available on Kaggle as a dataset[2] and the rest of the data is to be released soon. Additionally, we organized a kaggle competition recently, featuring court chamber classification based on the court decision text[3].

In this work we presented a valuable curated dataset of Swiss court decisions which can be used for a myriad of applications and with the potential to kick-start NLP research on legal data in Switzerland.

Keywords Artificial Intelligence (AI) · Natural Language Processing (NLP) · Law · Legal Tech · Dataset · Corpus

---

[1] `https://github.com/JoelNiklaus/SwissCourtRulingCorpus`

[2] `https://www.kaggle.com/joelniklaus/swiss-german-court-rulings`

[3] `https://www.kaggle.com/c/swiss-german-court-rulings`