Project Proposal: Web Scraping for a Database of Court Decision Related Documents

Alec Schürmann

Mai 2021

Abstract

So far, no datasets of external documents related to Swiss federal court rulings have been collected. In order to provide a database of court decision related documents, promising online data sources will be identified. In a second step, the HTML of chosen websites will be analyzed and the documents will be scraped. Furthermore, the texts from the documents will be extracted and structured in a database. Finally the results will be evaluated in the context of the overarching research project "Open Justice vs. Privacy" to automate re-identification of involved people in court decisions.

1 Introduction

To protect the privacy of involved people in Swiss court decisions, documents are anonymized. By linking the rulings with external data, previous research [4] has shown, that it is possible to re-identify companies involved in court decisions. In order to build an automated system for re-identifying people from court rulings, external data is needed. The goal of this project is to create a structured database from Swiss court decisions related documents by scraping related online documents and extracting the data.

2 Related Work

The article "A comparative study on web scraping" from De S Sirisuriya et al [3] shows the background and different techniques of web scraping in general. Additionally, it compares them by evaluating the techniques and software. "Choosing scrapy" from Daniel Myers and James W McGuffee [2] is a paper exploring the viability of Scrapy for undergraduate projects.

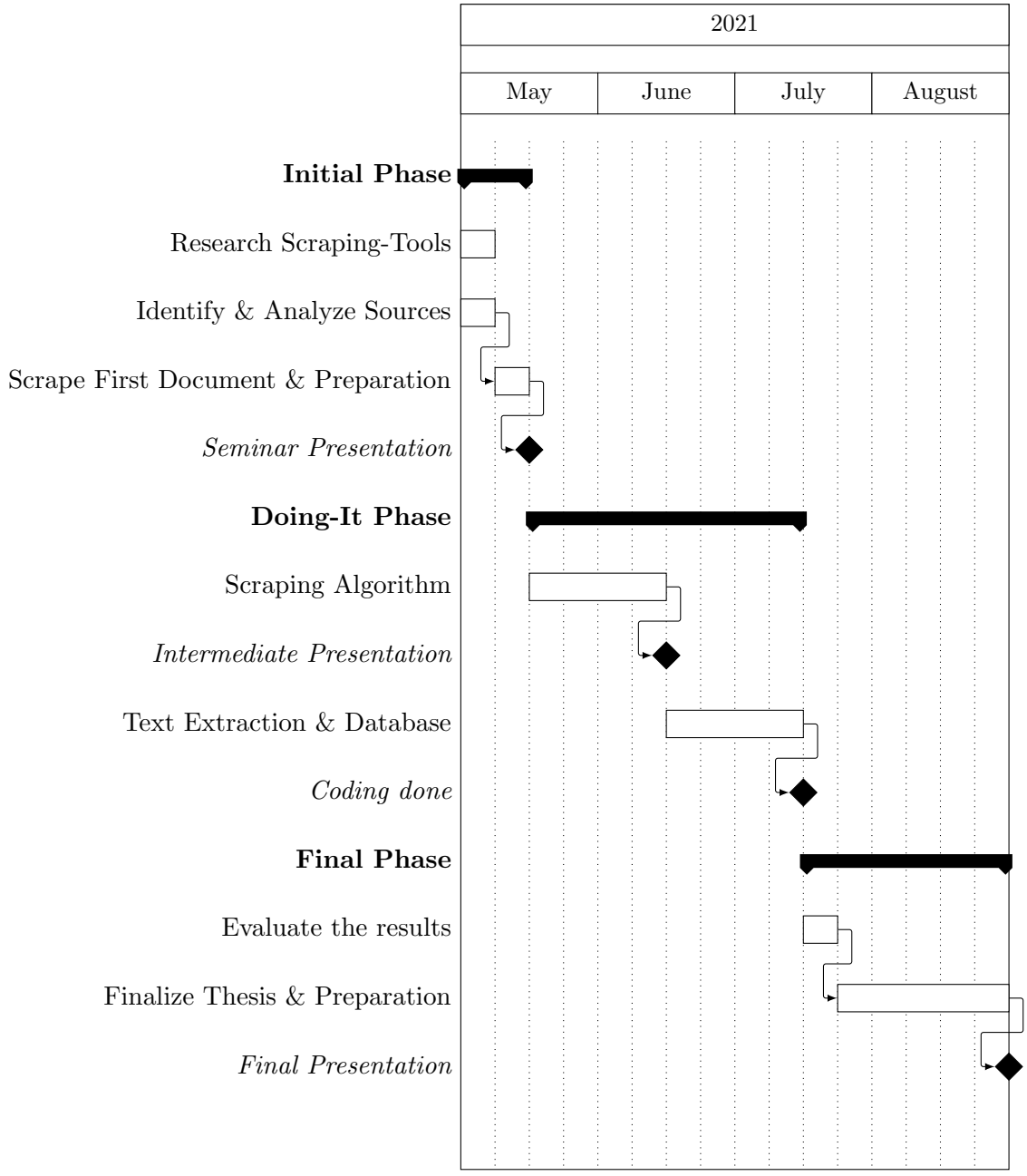
Finally they also explain why it was preferred for their student project. The article "Comparison of web scraping techniques: regular expression, HTML DOM and Xpath" from Gunawan, Rohmat et al [1] explains the importance of web scraping and compares three different web scraping techniques. The comparison is done by testing all the methods when retrieving data from target websites. They use process time, memory usage and data consumption as parameters in the experiment.

3 Tools

In this project, **BeautifulSoup** will be used for scraping the documents from websites. The decision was between either Scrapy or BeautifulSoup. Both of them are based on Python. BeautifulSoup only parses and extracts data from HTML files, whereas Scrapy downloads, processes and saves the data. While Scrapy offers more speed and scalability, BeautifulSoup is more fitting for the scale of this project, as the main use is content parsing. It will also require an additional content downloader to download those HTML files[1].

[1] https://smartproxy.com/blog/scrapy-vs-beautifulsoup

4 Timetable


2021
May
June
July
August
Initial Phase
Research Scraping-Tools
Identify & Analyze Sources
Scrape First Document & Preparation
Seminar Presentation
Doing-It Phase
Scraping Algorithm
Intermediate Presentation
Text Extraction & Database
Coding done
Final Phase
Evaluate the results
Finalize Thesis & Preparation
Final Presentation

References

[1] Rohmat Gunawan, Alam Rahmatulloh, Irfan Darmawan, and Firman Firdaus. Comparison of web scraping techniques: regular expression, html dom and xpath. In *2018 International Conference on Industrial Enterprise and System Engineering (ICoIESE 2018)*, pages 283–287. Atlantis Press, 2019.

[2] Daniel Myers and James W McGuffee. Choosing scrapy. *Journal of Computing Sciences in Colleges*, 31(1):83–89, 2015.

[3] De S Sirisuriya et al. A comparative study on web scraping. *Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka*, 2015.

[4] Kerstin Noëlle Vokinger and Urs Jakob Mühlematter. *Re-Identifikation von Gerichtsurteilen durch" Linkage" von Daten (banken): eine empirische Analyse anhand von Bundesgerichtsbeschwerden gegen (Preisfestsetzungs-) Verfügungen von Arzneimitteln.* 2019.