Semi-assisted Module Handbook Content Extraction for the Application of Curriculum Analytics

Rene Roepke¹, Maximilian Nell² and Ulrik Schroeder¹

Abstract: Alongside examination regulations, module handbooks provide overview of a study program, including information like workload, learning goals, examinations. They provide guidance to students, but can also be a valuable information source to curriculum analytics, e.g., the identification of trends and patterns across modules, the assessment of course content coherence, and data-driven decision-making regarding curriculum design and revision. This paper introduces a tool for semi-assisted module handbook content extraction, which uses natural language processing and text mining techniques to extract all properties and relevant details from module handbooks, allowing instructors and curriculum designers to efficiently identify key information. As module handbooks between institutions may look very different, fully automated extraction is difficult and error-prone. Here, a user-assisted approach for module handbook content extraction could support the process and introduce manual data correction cycles to improve data quality. This demonstration paper presents a first prototype of a content extraction tool³ using natural language processing (NLP) and text mining techniques to extract all properties and relevant details from module handbooks and prepare them for curriculum analytics.

Motivation and Prototype Workflow

For the analysis of study programs and curricula, the contents of module handbooks and examination regulations as well as student activity data are considered valuable input. Covered under the umbrella term curriculum analytics [Hi22], a subdomain of learning analytics, the beforementioned data can provide meaningful insights to detecting patterns, trends and evidence to guide data-driven decision-making regarding curriculum design and revision. A crucial step in curriculum analytics is the data collection and pre-processing. Here, module handbooks provide an overview of a study program, including information like workload, learning goals, examinations. As module handbooks between institutions may look very different, fully automated information retrieval is difficult and error-prone. Here, a user-assisted approach for module handbook content extraction could support the process and introduce manual data correction cycles to improve data quality. This demonstration paper presents a first prototype of a content extraction tool³ using natural language processing (NLP) and text mining techniques to extract all properties and relevant details from module handbooks and prepare them for curriculum analytics.

While the content seems unstructured at first, a module handbook consists of a list of modules described by a set of properties. Each property can be described using a key-

¹ RWTH Aachen University, Learning Technologies, {roepke, schroeder}@cs.rwth-aachen.de, https://orcid.org/0000-0003-0250-8521, 0000-0002-5178-8497
² RWTH Aachen University, maximilian.nell@rwth-aachen.de
³ https://git.rwth-aachen.de/learntech-lufgi9/mhb-extractor, last accessed 15.06.2023

©©©© doi: 10.18420/delfi2023-45
value pair, where the key is the abstract property name and the value is the assigned content, e.g. a property for module duration called “Moduldauer” and its value “1 Semester”. These characteristics of the document structure can be exploited with the help of an NLP model [Ta20] and text mining [FT21]: On the one hand via a simple frequency analysis and on the other hand via the language model itself, recognizing certain strings in the document. In the first step, according to the motto divide and conquer, a term or string always appearing at the beginning of a module description is searched for, e.g., "Modulname". If not detectable by means of the model, it can also be recognized over a customizable Regular expression. Next, each module start is marked and the document is automatically split into one PDF per module. Now, for the recognition of the property keywords and their values, a predefined set of synonyms for module properties is evaluated and marked in the PDF using the NLP model. In addition, we extract a list of bold strings in the PDF and compare them with the keywords of the model. Under the assumption that keywords are printed in bold in many module handbooks, this heuristic is particularly well suited to supplementing the model. In a semi-assisted manner, the user then checks all marked key-value pairs for one module in order to correct markings and add missing properties and values. By allowing users to verify and correct extraction results, higher accuracy and reliability can be achieved. Based on similarities in the module description structure and repetitively used keywords, checking one marked module allows for adapting the model to detect properties in all. Here, corrected and verified key-value pairs may serve as training data for model improvement.

Overall, while the semi-assisted approach still needs user input and is therefore costly compared to fully automated approaches, the quality of the extraction results is crucial to its further use in curriculum analytics. With the development of a module handbook content extraction tool, we are able to retrieve comparable information from different module handbooks of different institutions and provide rich input data to the application of curriculum analytics. Potential analyses include the identification of trend topics for similar study programs or the generation of competency profiles, depending on students’ module choices. Of particular interest is module handbook data for study path planning, e.g., by constructing a knowledge graph and recommend modules with connected topics or to foster specific competencies. Since NLP is already used in the extraction process, another interesting direction would be using curriculum data for mentoring and guidance tools, e.g., a generative AI-based chatbot for choosing a study program or elective courses.

Bibliography

