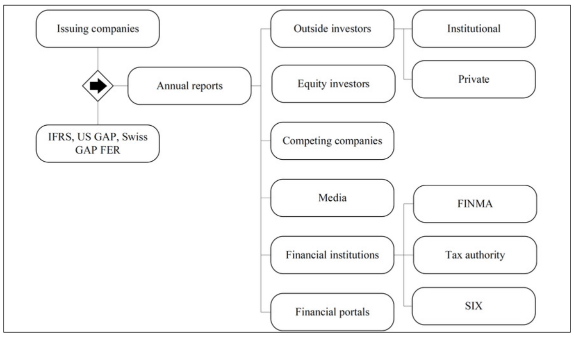
Marc Alexander Willhaus

Graduate Candidate	Marc Alexander Willhaus
Examiner	Prof. Dr. Felix Nyffenegger
Co-Examiner	Thomas Lutz, Intelliact AG, Zürich, ZH
Subject Area	Innovation in Products, Processes and Materials - Business Engineering and Productions

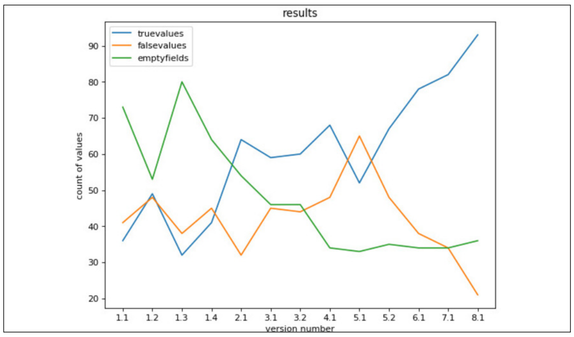
Automated PDF Text Mining

Determination of Predefined Figures by Using a Rule Based-Approach

Illustration of Annual Reports



Stakeholders of Annual Reports



Algorithm Output Results

Introduction: With the thriving development of the internet, the sources in order to choose investing signals for the correct stocks is getting bigger from day to day. Therefore, institutional investors can gain information from financial platforms such as the Bloomberg terminal. These platforms charge a fee for usage which is mostly not bearable for the private investor. He or she could gain the information by searching the annual reports of the company. But this is only possible by investing time into reading those reports and having an underlying knowledge of financial statements to read them. In a first step, the figures, in general, should be looked at and compared to benchmarks and other companies, which can be exhausting and time intense.

Objective: Therefore, the goal of this thesis will be to implement an algorithm which extracts predefined key figures automatically such as the operative cashflow out of annual reports as exact as possible. Those figures could then be presented online and accessed by an Application Programming Interface (API).
Several different key figures will be predefined and examined more closely. For evaluating the algorithm, 30 randomly chosen annual reports are going to build the data. The true values will be picked by hand and compared to the results of the algorithm. The algorithm will be based on a rule-based approach and tries to make use of the structure of the given annual reports in a PDF format.

Result: The results of the algorithm lead to an accuracy of 81,6%, which is sufficient to hold on to the hypothesis one: There is a rule-based algorithm for detecting more than 80% of the figures from an annual report in a PDF document format correctly. Continuing from that basis we a are thriving to get a result towards 100%. Further research could focus on the proposed algorithm, other statistical approaches or a combined method to extract figures out of PDF documents.

HSR
HOCHSCHULE FÜR TECHNIK
RAPPERSWIL
FHO Fachhochschule Ostschweiz