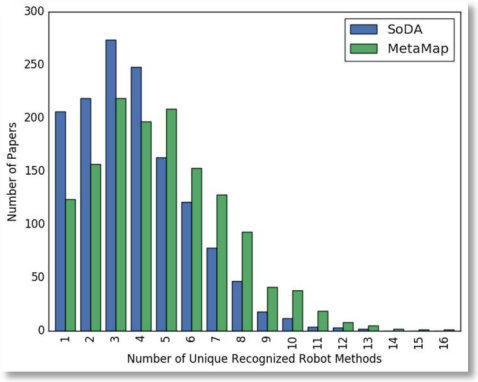
Improving Literature Research using Big Data and Natural Language Processing

Literature research is a crucial part of science: Only if the current state of the art is known, I can build my own research on it, find gaps to fill and thus advance the scientific community. However, due to the massive amount of publications published each day, it is sometimes difficult to keep track of all relevant literature. While finding and reading a few key papers manually is possible, it is often wishful to obtain more insights by relying on statistics gathered on the basis of thousands of papers.

In this project, a framework for big-data based literature research is to be developed and implemented. While the focus of our research is laboratory automation, the objective is to create and showcase a versatile tool that can be used in different research areas. This encompasses among others the following subtasks:



Data extracted using NLP in [1]

- Literature review regarding state of the art (probably manually so far - I'm sorry!)
- Creation of a reusable workflow to obtain texts revolving around a specific topic ("corpus")
- Implementation of a suitable data infrastructure to host and search the data, e.g. using Elasticsearch
- Implementation and comparison of NLP methods to analyze the corpus
- Application to actual questions from our group with focus on lab automation

This project targets the final assignment (Bachelor / Master) and, as an option, the preparatory "Praktikum". It is particularly suitable for computer science students, but it is open for students of other areas as well. Knowledge of Python is clearly a plus for the implementation. In case you're interested or you have any questions, feel free to reach out to Henning Zwirnmann:

henning.zwirnmann@tum.de
089/289-29417

Exemplary References:

[1] P. Groth and J. Cox. "Indicators for the use of robotic labs in basic biomedical research: a literature analysis" (2017), https://peerj.com/articles/3997/

[2] K. Roper et al. "Testing the reproducibility and robustness of the cancer biology literature by robot" (2022), https://royalsocietypublishing.org/doi/10.1098/rsif.2021.0821