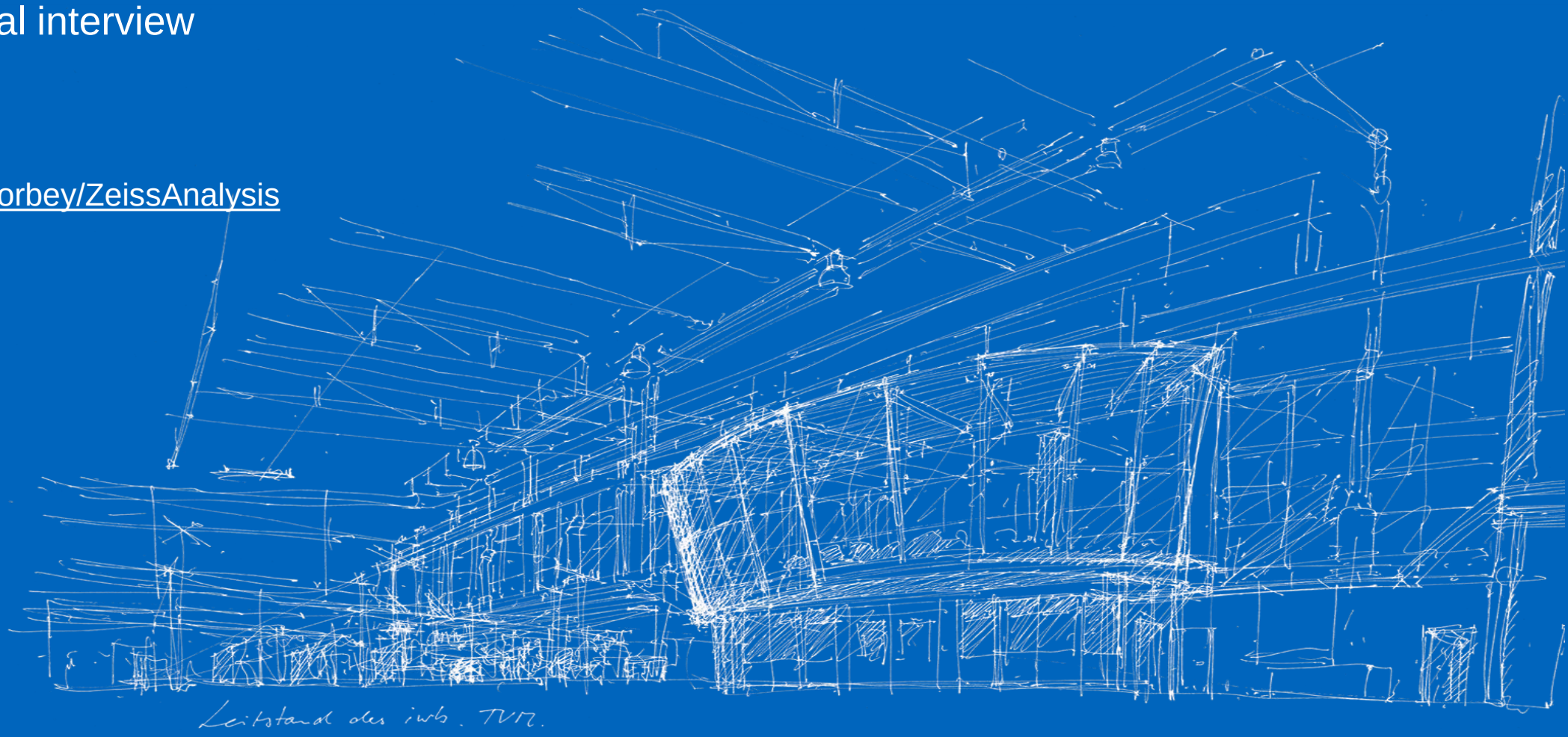
Institut für Werkzeugmaschinen und Betriebswissenschaften
TUM School of Engineering and Design
Technische Universität München

Natural language processing for plagiarism detection

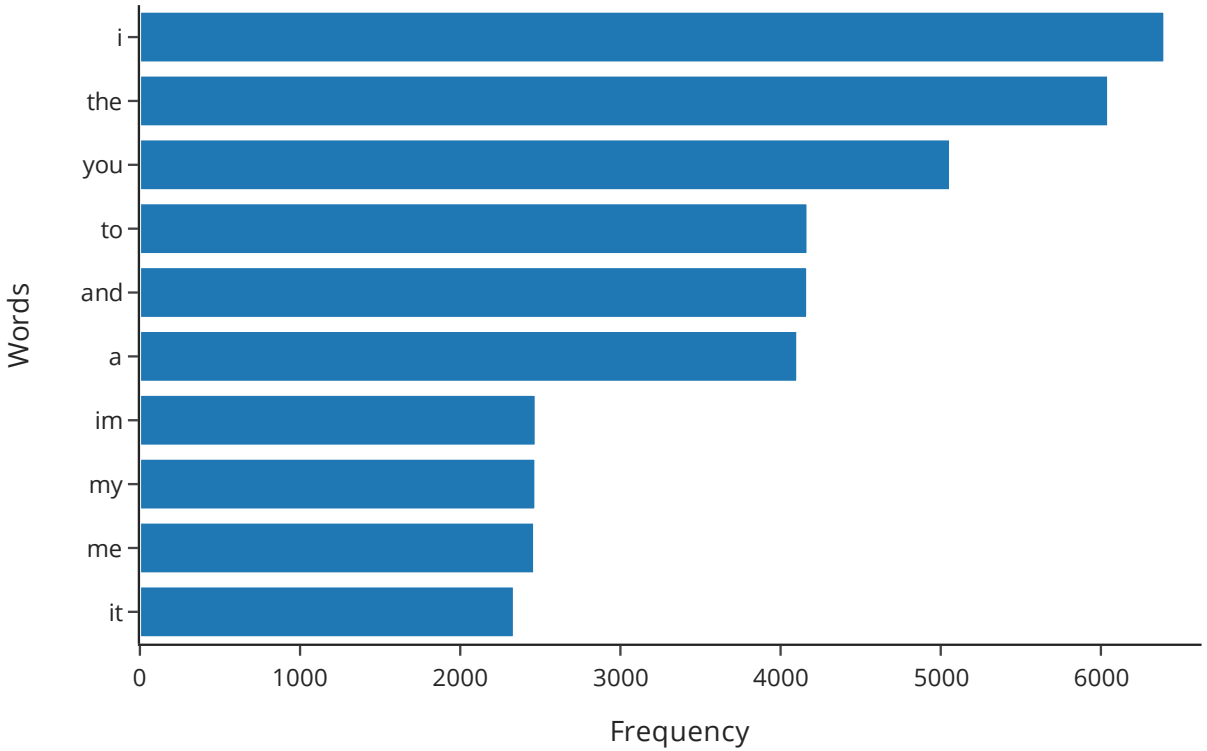
Zeiss – Technical interview

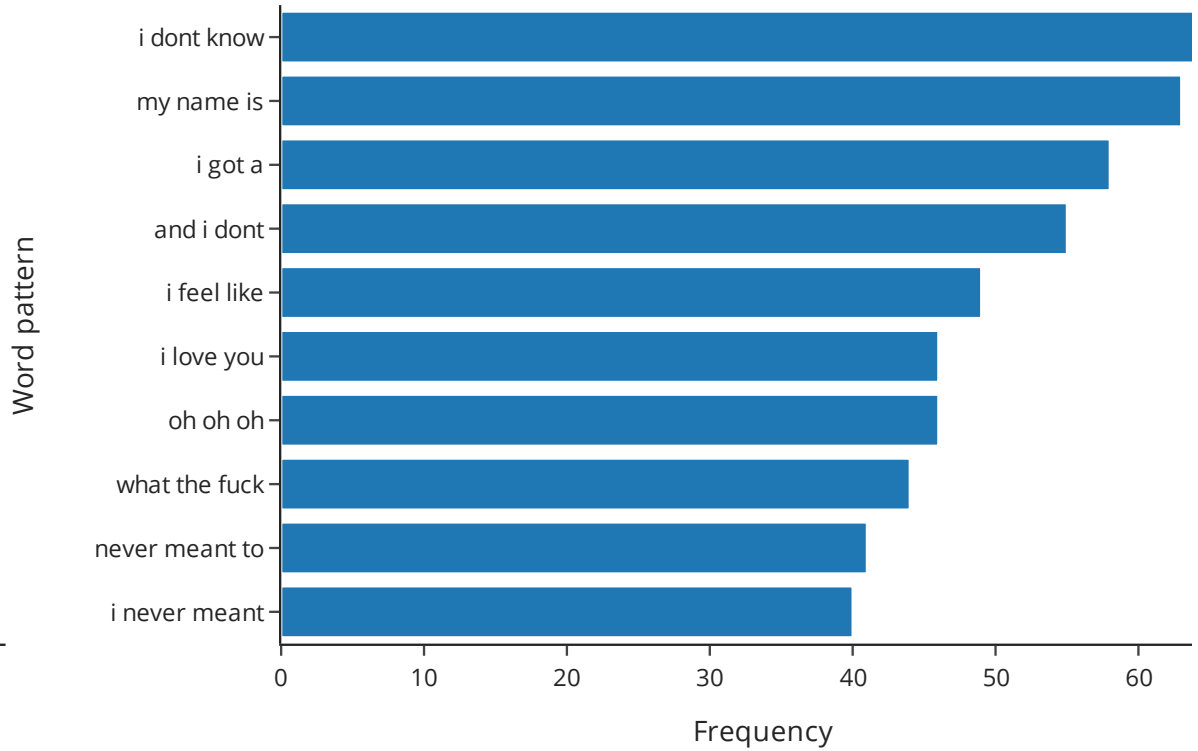
C. Nentwich

https://github.com/xorbey/ZeissAnalysis

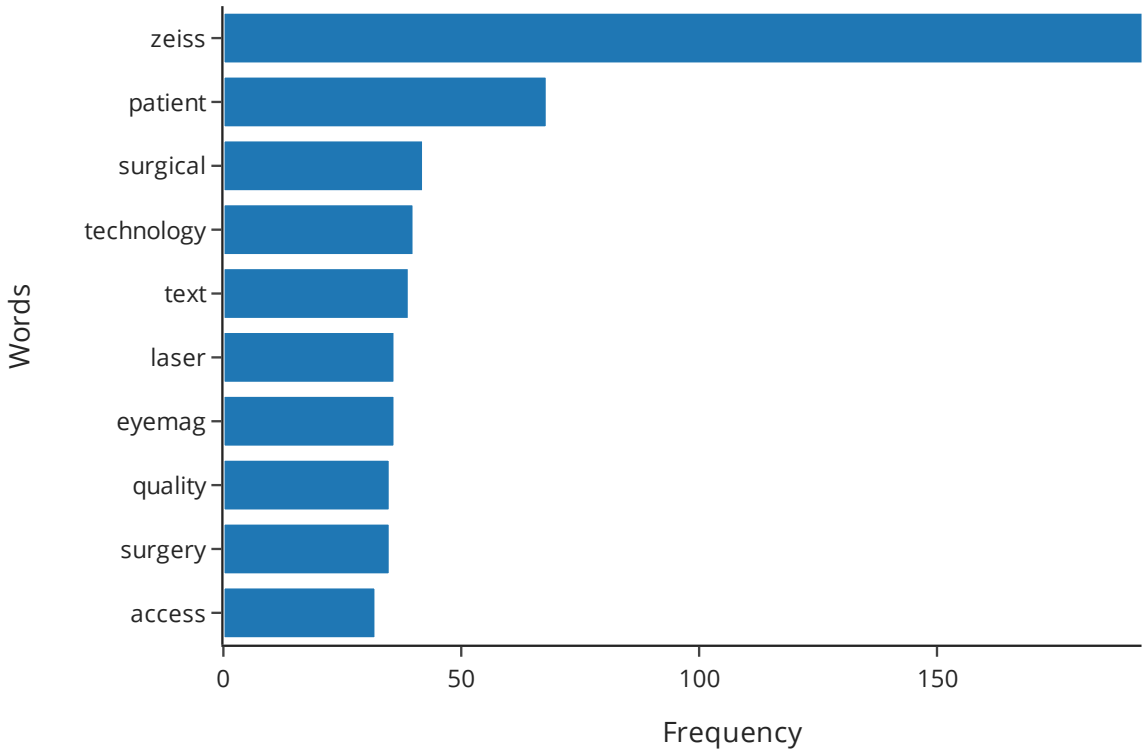


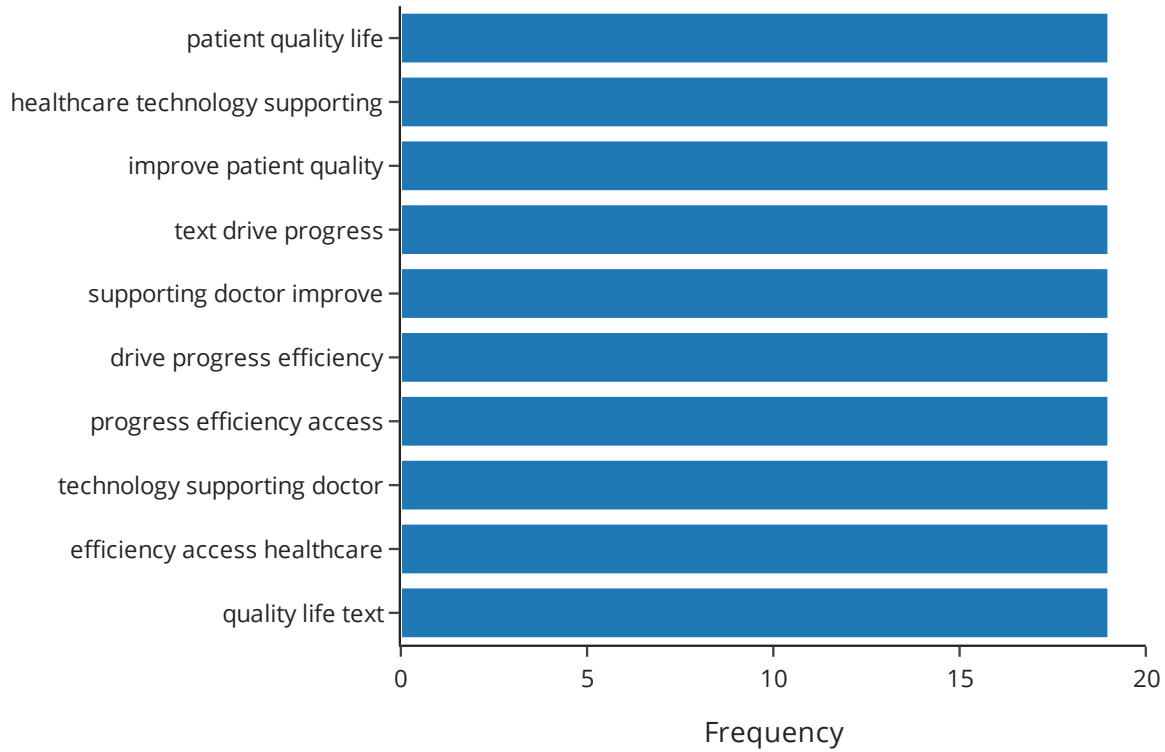
Eminem is good in denial and love…



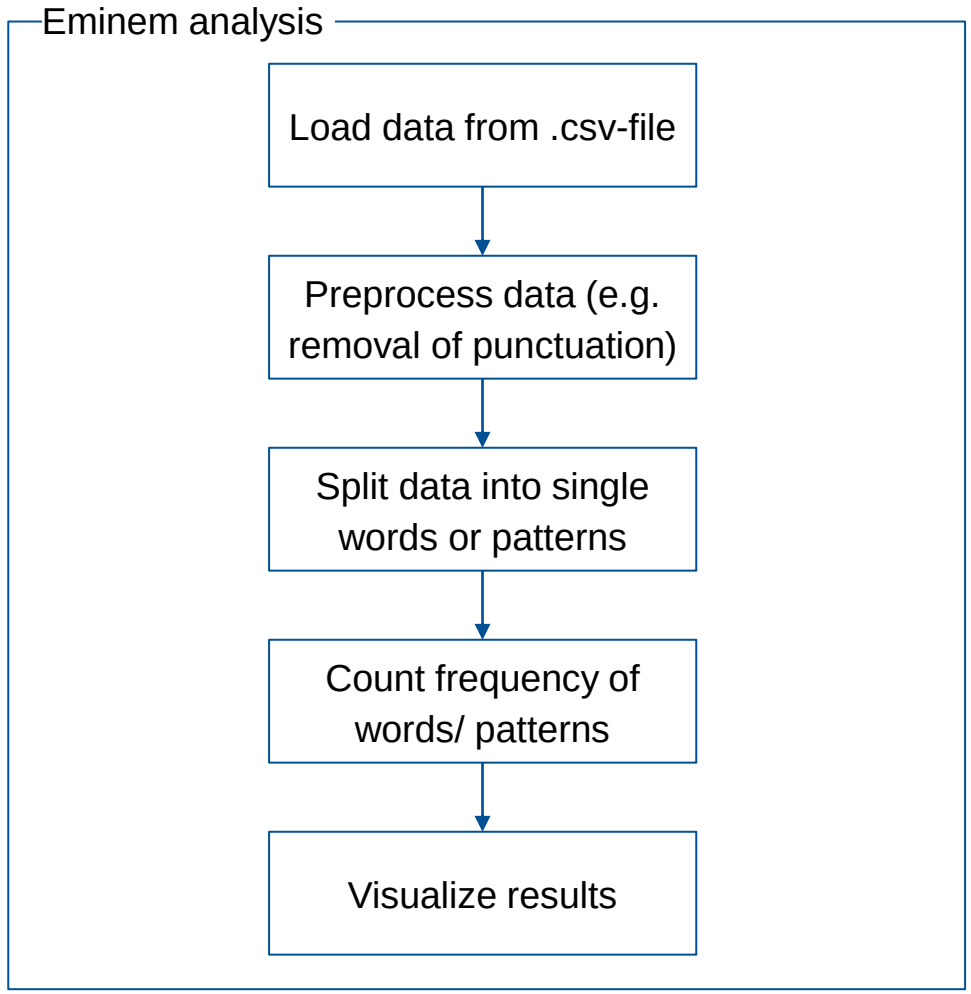


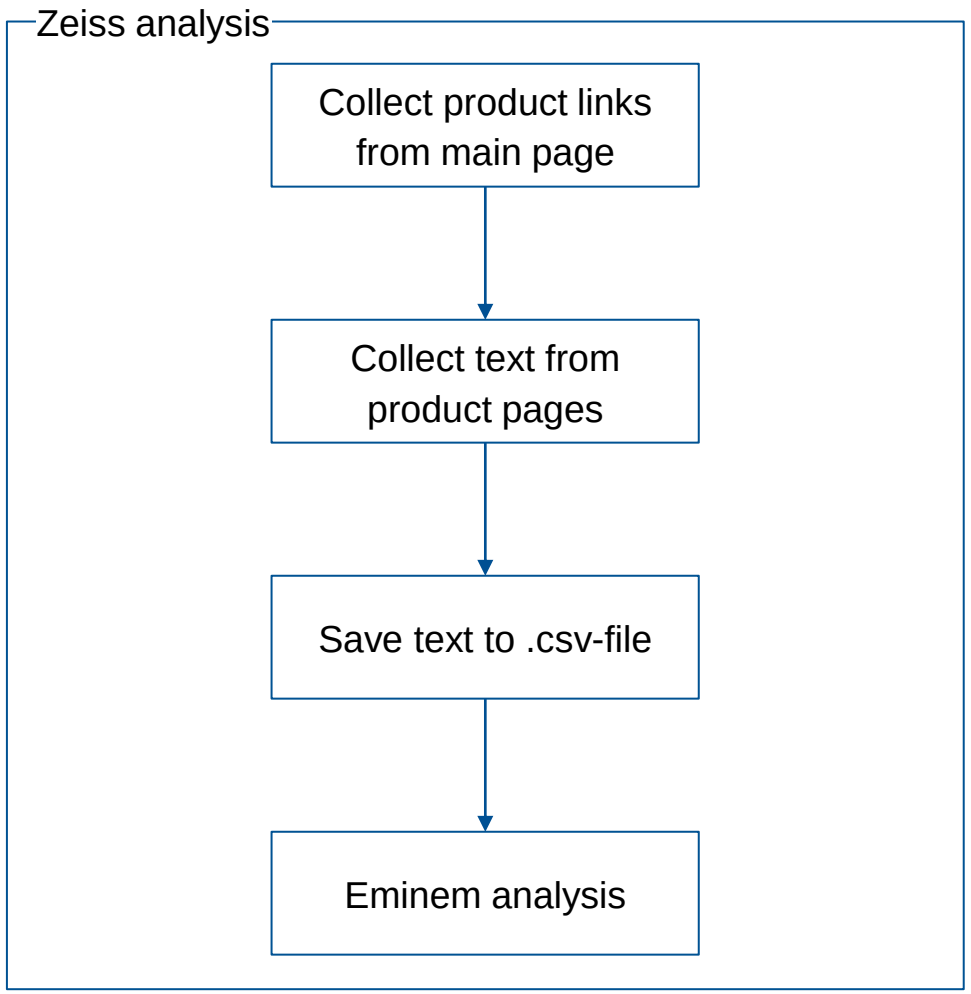
… and Zeiss medical products love to support doctors and patients





The analyses were based on counting the frequency of words and word patterns





Scaling and comparing the analyses requires the consideration of certain details

Scalability: Eminem analysis

Technical aspects:
→ can be scaled using an Apache Spark cluster
+ Reuse of some of the notebook code
→ might involve casting strings into unique numbers to save memory

Data aspects:
→ Preprocessing language dependent

Scalability: Zeiss analysis

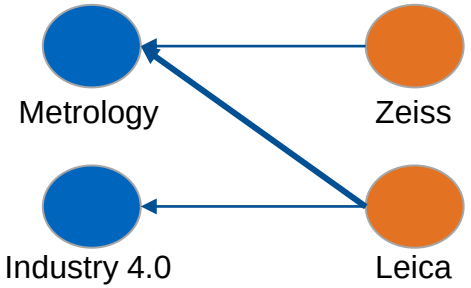
Technical aspects:
→ Web crawling activities can be scaled by multiprocessing/ threading / async functions

Data aspects:
→ Either inform crawling process with webpage structure or improve preprocessing pipeline

Comparison of rappers/ pages

Analysis between different entities could be compared by building and visualizing graphs:

- Nodes: Entity (Rapper/ Company), Word / Pattern
- Edges: „is_using“ with attribute frequency



Backup: Structure of Zeiss web pages