Technische Universität München

Marlene Knigge, Felix Willnecker,
Tamas Neumer, Prof. Dr. Helmut Krcmar

Chair for Information Systems

Final Project Report: Applying Text Mining on Job Offers and Curricula Vitae Using SAP HANA: Analyzing Skills and Competencies for Industry 4.0

Situation

Industrie 4.0/Industry 4.0/the Industrial Internet: combining known technologies in a new way: disruptive solutions

- The Internet of Things
- Cyber Physical Systems (CPS)
- Smart Factories
- Embedded Systems

(Kagermann et al. 2013, Industrie 4.0 2016)

Complication

The way we live and work will change significantly

- New ways of business value creation
- Changing business models and strategies
- Adjusted business processes (Kagermann et al. 2013)
- Requirements for employees change, employees have to be prepared for new tasks and a new working environment
- There is still a lack of a qualification agenda for Industry 4.0

Resolution

Apply text analysis and text mining offered by SAP HANA and the SAP Predictive Analysis Library (PAL) on **job offers** collected from German online job portals to extract skill and competency requirements for Industry 4.0:

- Build on technical experience from former projects
- Apply technical knowledge to a new area
- Try to discover competency profiles from analyzing job offers

Extract skills and competencies from **curricula vitae (CVs)** from IT professionals using the same application for later comparison with results from job offer analysis

- Dataset is handed over once (no need for extraction of data on a regular base)
- Technical adjustments to the analysis application
- Evaluation of quality of analysis, preparation for combining results with analysis of job offers

System

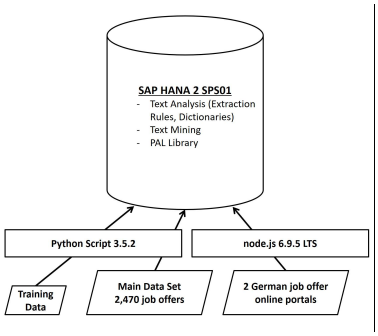
- SAP HANA 2 SPS01, 1 TB RAM, 32 Cores (CPU)
- SAP HANA Studio, version 2.3.10
- PAL library
- Python, version 3.5.2
- Node.js 6.9.5 LTS

Data Sets

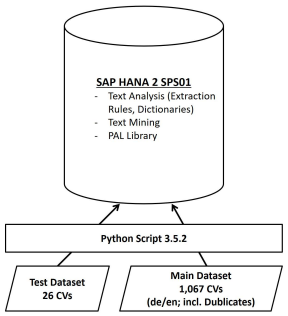
- Manually collected German **job offers** from two job portals, most tests done on collection of Nov – Apr. (2,470 job offers)
- Test data sets: T1: 15 job offers, T2: 50 job offers, manually classified

- 1,067 **CVs** from IT professionals from a project partner, containing different file-formats (mainly pdf), German and English language, and duplicates
- Test data set: T3: 26 CVs in German language in pdf-format

Project Architectures (Job Offers/CVs)



Source: Own illustration.



Source: Own illustration.

Results

- **Skill and competency requirement extraction**
  - Custom dictionaries and extraction rules lead to good results:
    - 68% of skills and competency requirements extracted
    - Precision: 92,8%, sensitivity: 77,5%, F1-score: 84,5%
  - Recommendations for further improvement
    - Convert further parts of custom dictionary into custom extraction rules
    - Apply the Grammatical Role Analysis to reduce false positives (only available in English so far)
  - Frequency analysis on discovered skill and competency requirements
    - Top 5: languages (64%), flexibility (40%), capacity for teamwork (36%), communication skills (29%), and self-responsibility (25%)
  - Association analysis: Weak results, difficulties with the attempt to discover competency profiles
  - Clustering: Good approach for deriving competency profiles from clusters

- **Web Crawler**
  - Implemented as stand alone application using node.js
  - Working with one German online job offer portal so far

- **CV Analysis**
  - Technically working
  - Structure of the CV documents is not extracted and therefore not taken into account in analysis → misinterpretations possible

Outlook (Ideas)

- Integrate web crawler into SAP HANA application
- Analysis of CVs with regards to metadata (e.g., education section, project description, …)
- Compare skill and competency requirements (job offers) with what is offered by today's IT professionals (CVs)
- Analysis of skill and competency requirements with regards to metadata (e.g., job title, company, region) (cp. results from parallel work of Neumer 2018)
- Implementation of a regular analysis to discover changes over time