DAIMLER

Advanced Job Analytics @ Daimler

Julian Leweling, Daimler AG

Agenda

From Job Ads to Knowledge: Advanced Job Analytics @ Daimler

- About Daimler AG
- Why KNIME?
- Our Inspiration
- Use Case
- KNIME Walkthrough
- Application
- Next steps

DAIMLER

Who is...
Daimler AG

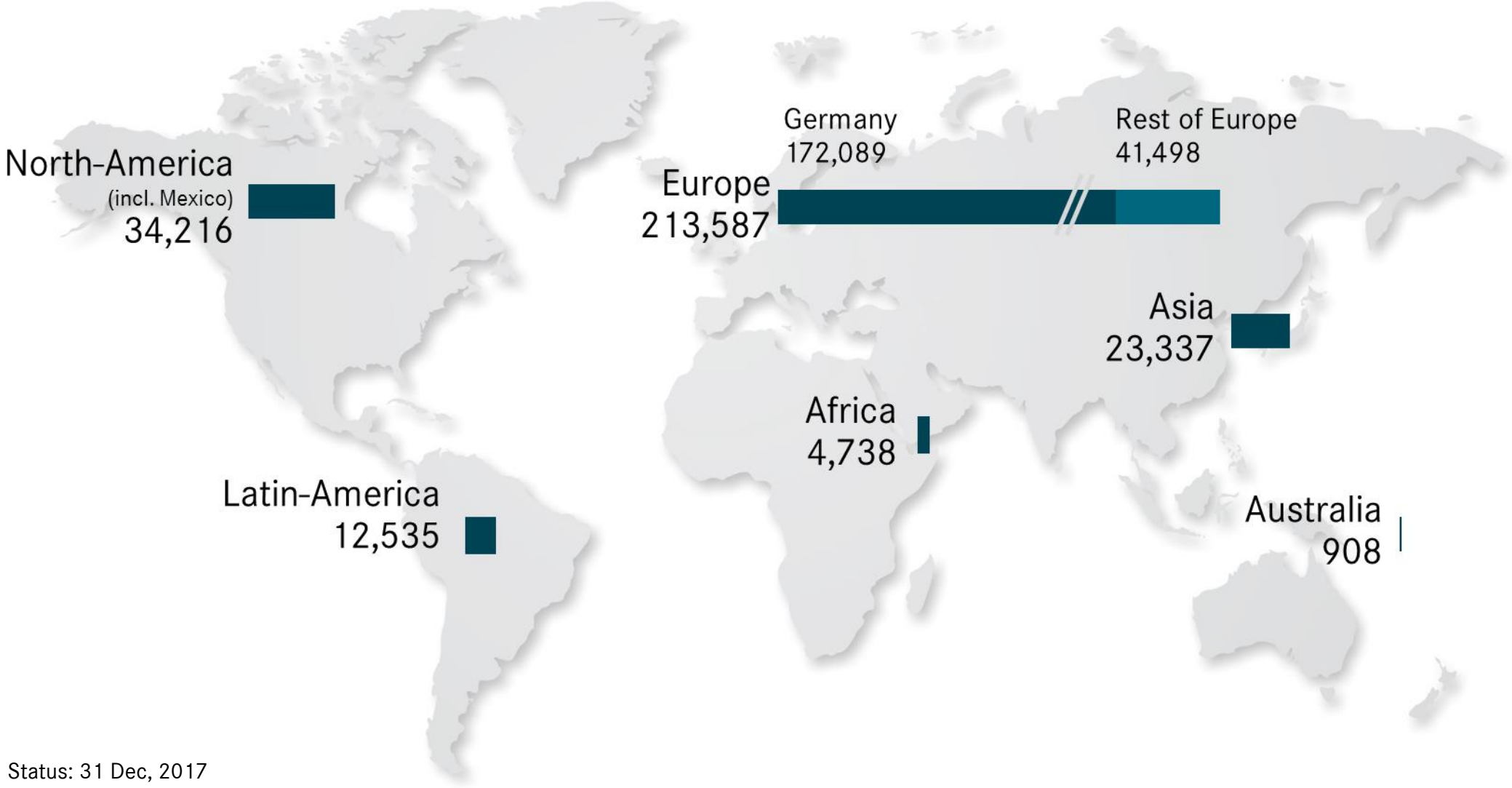
Daimler consists of five divisions

	Mercedes-Benz Cars	Mercedes-Benz Vans	Daimler Trucks	Daimler Buses	Daimler Financial Services
					
2017					
Revenues	€ 94.7 bn	€ 13.2 bn	€ 35.7 bn	€ 4.4 bn	€ 23.8 bn
Employees	142,666	25,255	79,483	18,292	13,012

MAYBACH, smart, EQ, Mercedes me, AMG, FUSO, BHARATBENZ, WESTERN STAR, FREIGHTLINER, Thomas BUILT BUSES, SETRA, moovel, CAR2GO, mytaxi, Mercedes-Benz Financial Services, Daimler Truck Financial, Mercedes-Benz Bank

Note: 2017 Revenue Group 164,330 million €, Employees: 289,321 , thereof corporate-wide functions & services 10.613

Daimler has about 289,300 employees worldwide
Regional distribution of workforce



Status: 31 Dec, 2017

Why KNIME?

- Fast & Versatile
- Easily joining different data sources
- Full transparency and reproducibility
- No more errors due to manual data editing
- Advanced analytic features

DAIMLER

From Job Ads to Knowledge:
Advanced Job Analytics @ Daimler

Our Inspiration

Use Case

- Semantic analysis of 3.800 positions
- Similarities and differences between jobs?
- Which qualifications are important?
- Clustering positions enhances transparency and facilitates active HR development

Qualifications

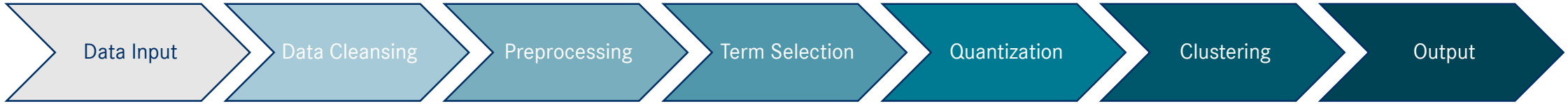
BA/BS in Computer Science, Math, Physics, Engineering, Statistics or other relevant technical field. Advanced degrees preferred.
Demonstrable programming experience with at least two of the following languages: Python, Java, Scala, R, Ruby, MATLAB, SQL.
Solid knowledge and experience with a scientific computing platform (e.g. scikit learn, Weka, MATLAB)
Hands-on experience working with common DBMS (SQL, NoSQL), as well as distributed application platforms (Hadoop).
Strong knowledge of statistical data analysis and machine learning techniques (e.g. SVM, regression, classification, clustering, time series, deep learning).
Hands-on experience with visualization tools (e.g. D3.js, Tableau) and an acute ability to prepare and present data in a visually appealing and easy to understand manner.
A strong voice for data integrity and reporting quality utilizing best-practices and industry standards
Excellent critical thinking, problem solving and analytical skills.
Excellent communication skills, and the ability to work effectively with others.
Ability to work with Linux-based systems and command-line tools.
Previous experience working with geospatial data is a plus.
Automotive experience is a plus.

Qualifications

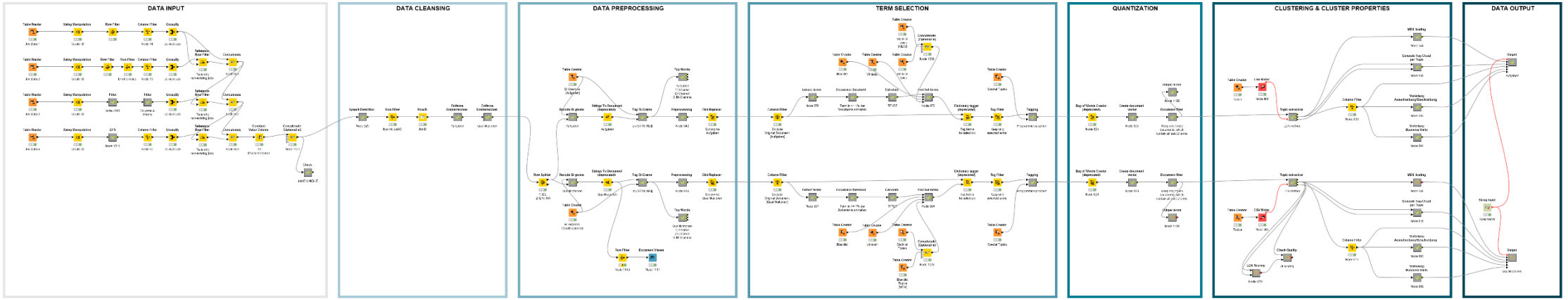
BA/BS in Computer Science, Math, Physics, Engineering, Statistics or other relevant technical field. Advanced degrees preferred.
Demonstrable programming experience with at least two of the following languages: Python, Java, Scala, R, Ruby, MATLAB, SQL.
Solid knowledge and experience with a scientific computing platform (e.g. scikit learn, Weka, MATLAB)
Hands-on experience working with common DBMS (SQL, NoSQL), as well as distributed application platforms (Hadoop).
Strong knowledge of statistical data analysis and machine learning techniques (e.g. SVM, regression, classification, clustering, time series, deep learning).
Hands-on experience with visualization tools (e.g. D3.js, Tableau) and an acute ability to prepare and present data in a visually appealing and easy to understand manner.
A strong voice for data integrity and reporting quality utilizing best-practices and industry standards
Excellent critical thinking, problem solving and analytical skills.
Excellent communication skills, and the ability to work effectively with others.
Ability to work with Linux-based systems and command-line tools.
Previous experience working with geospatial data is a plus.
Automotive experience is a plus.

Overview

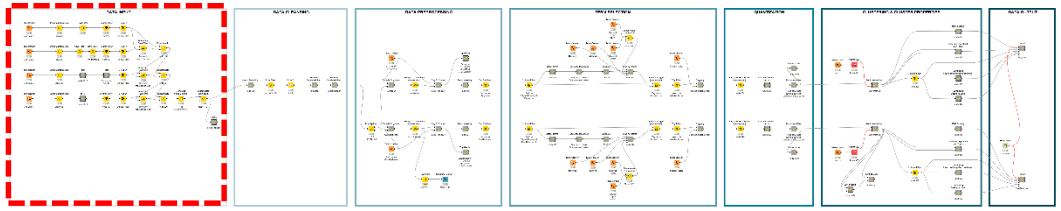
Processing Stream



KNIME Workflow



KNIME Walkthrough

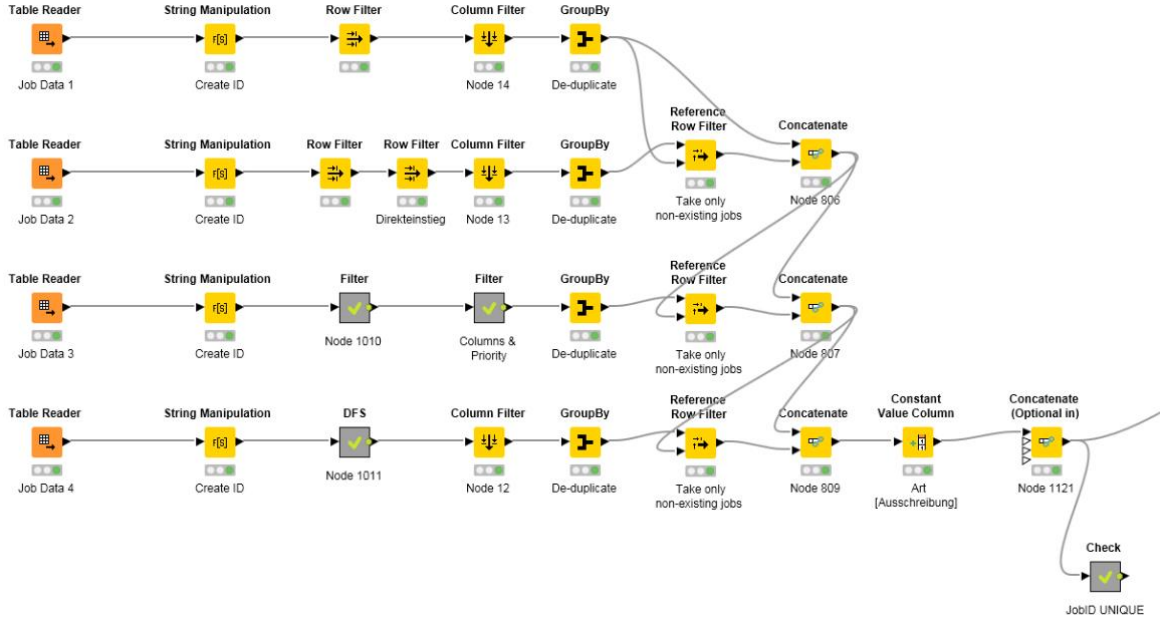


Data input

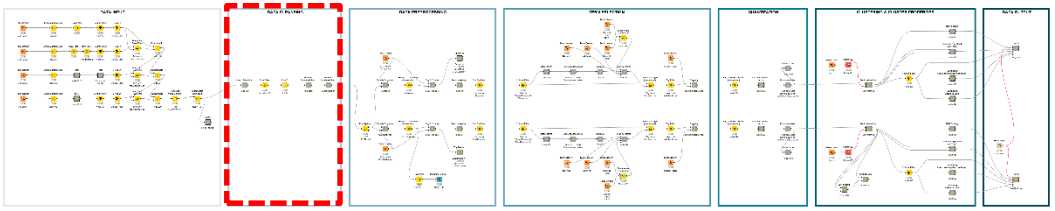
- Job advertisements
- Job descriptions

Selection of data from a specific division

- IT department, Finance & Controlling, etc.
- Relevant for extracting domain specific knowledge

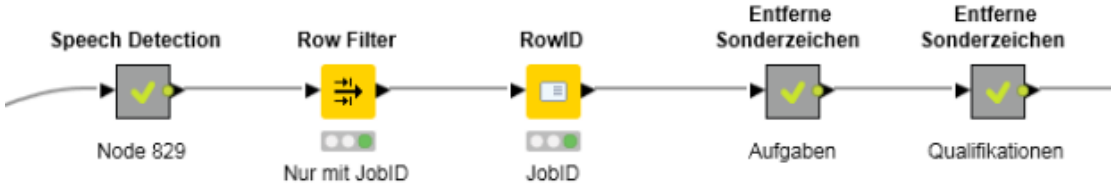


KNIME Walkthrough



Data cleansing

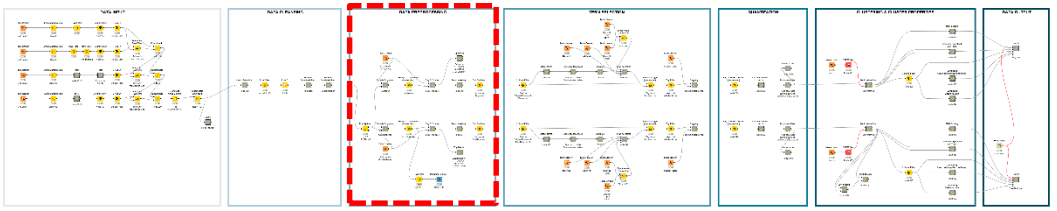
- Speech detection
  - Most job descriptions are German or English
  - Language-specific preprocessing needed



Removal of special characters

- Structuring signs [•, –, :]
- Multiple whitespace, line breaks, etc.

KNIME Walkthrough

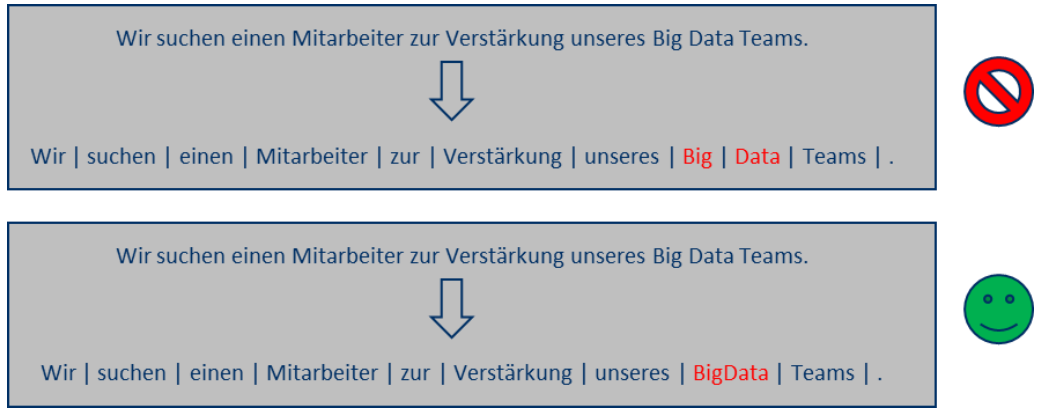


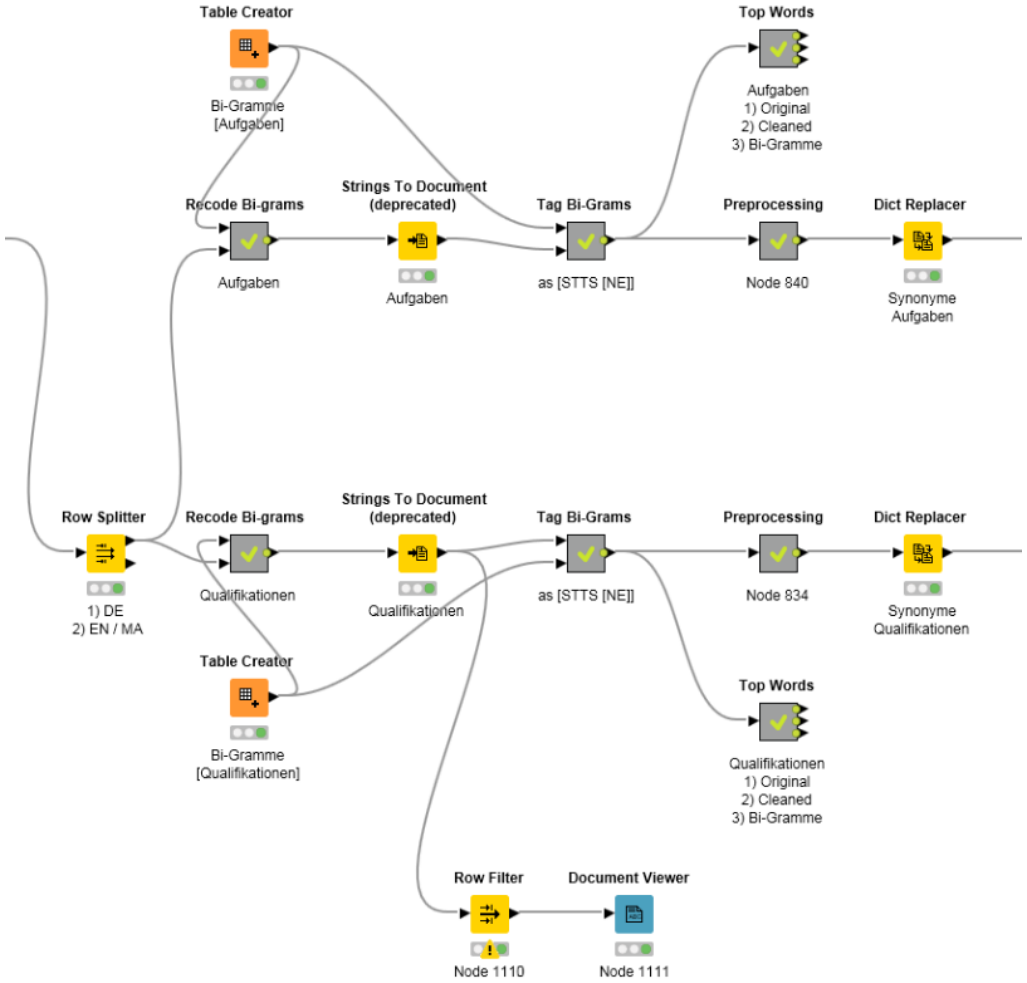
Data preprocessing #1

- Identification of relevant Bi-grams with a strong semantic link

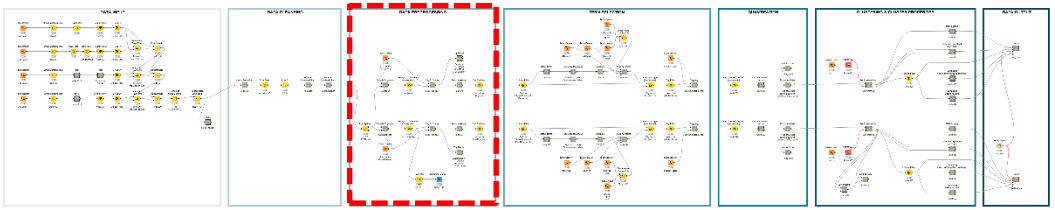
  => “Big Data”, “MS Office”, etc.

- Replaced with concatenated representations



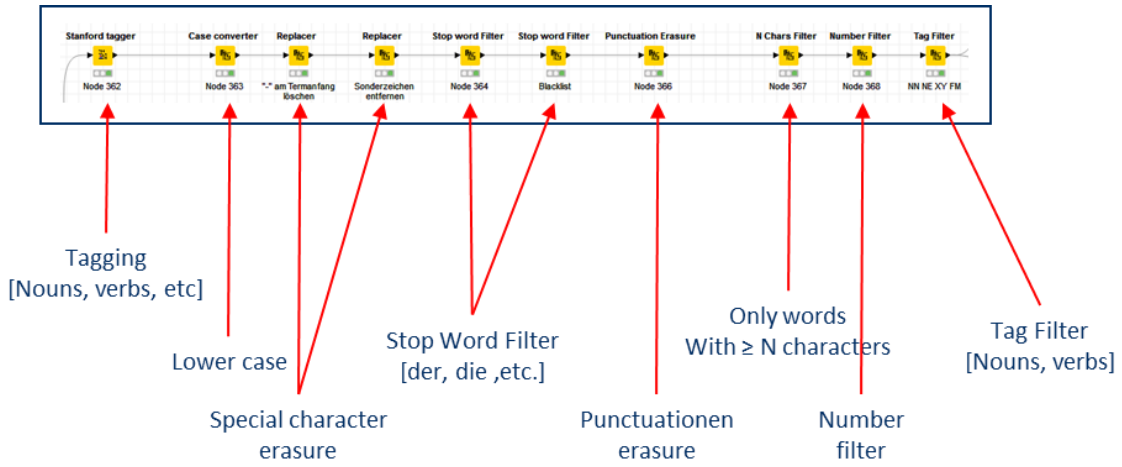


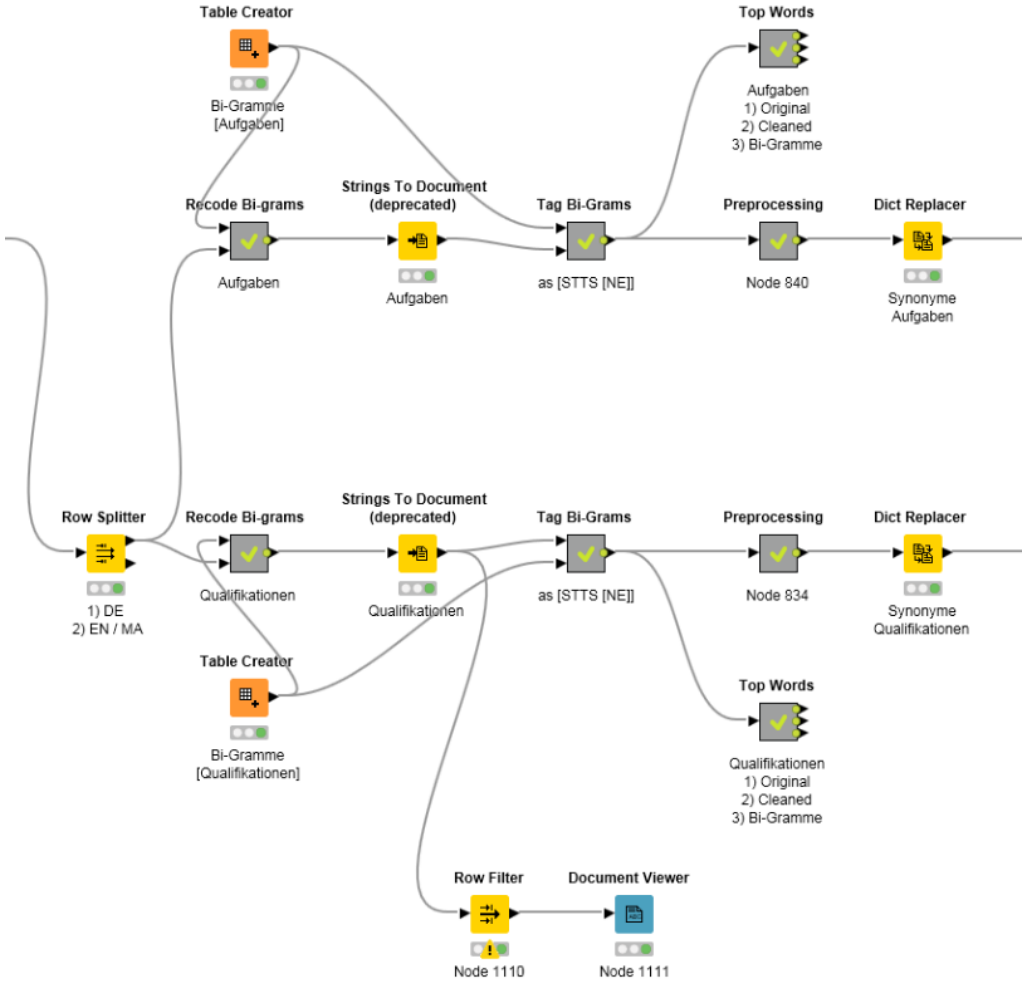
KNIME Walkthrough



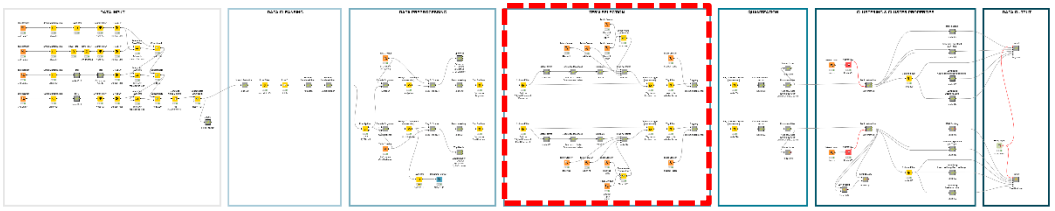
Data preprocessing #2

- String to document
- Typical preprocessing steps
- Replacement of synonyms and abbreviations



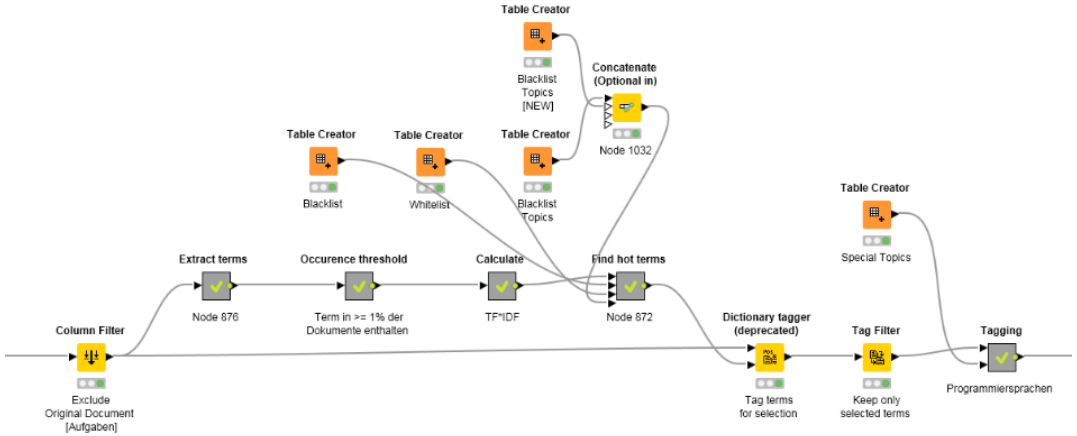


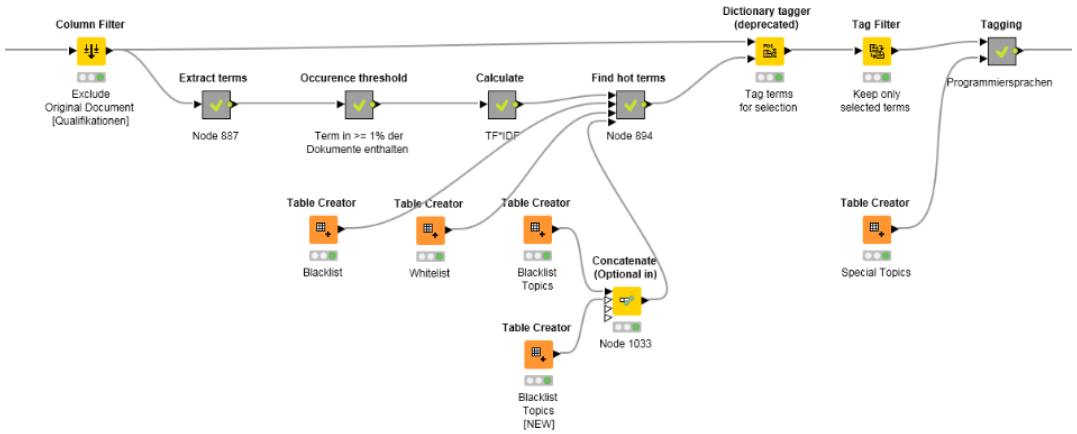
KNIME Walkthrough



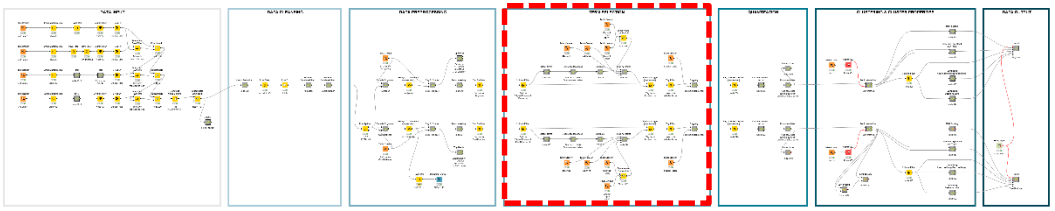
Term selection

- Motivation
  - Reduce the number of distinct terms and keep only those, which really matter
- Goal
  - Performance
  - Extracting the essentials
  - Filter out noise





KNIME Walkthrough



Term selection

1. Identification of relevant terms
   - Occurrence threshold across documents ≥ 1.0%
   - Filter out unusual wordings, special cases, etc.
2. Calculation of TF*IDF measure
   - Selection of Top X terms
3. Black- & Whitelists
   - Manually created, imputing knowledge from domain experts

Relative term frequency (TF)
Frequency of term occurrence within a specific document

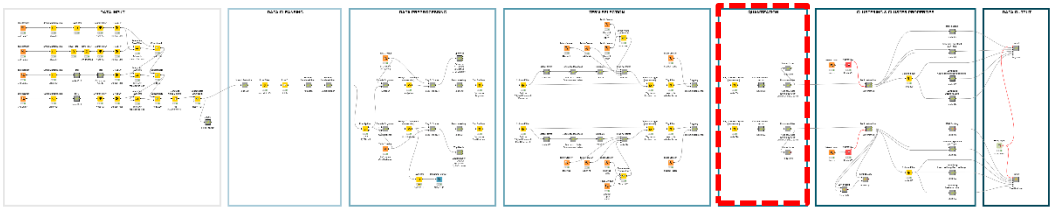
➢ The more often a term occurs in a document, the more relevant it is for this document

Inverse document frequency (IDF)
Log-ratio of „Nb. of documents with Term X“ to „Nb. of all documents“

➢ The more often a term occurs across documents, the less relevant it is in general

The term „Daimler“ for example is relevant for documents describing different automotive manufactures, but not when screening job advertisements within the Daimler AG.

KNIME Walkthrough

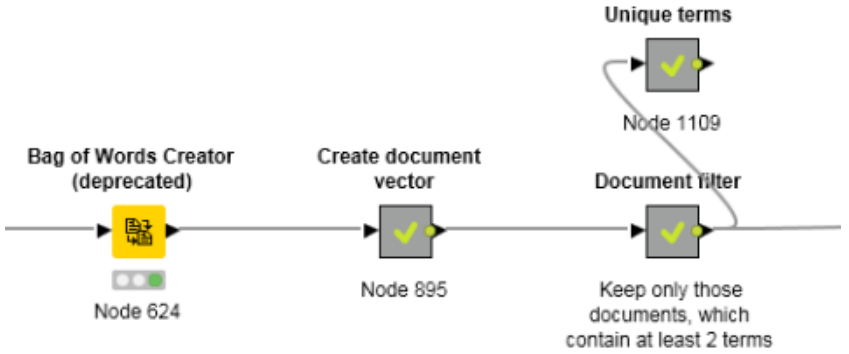


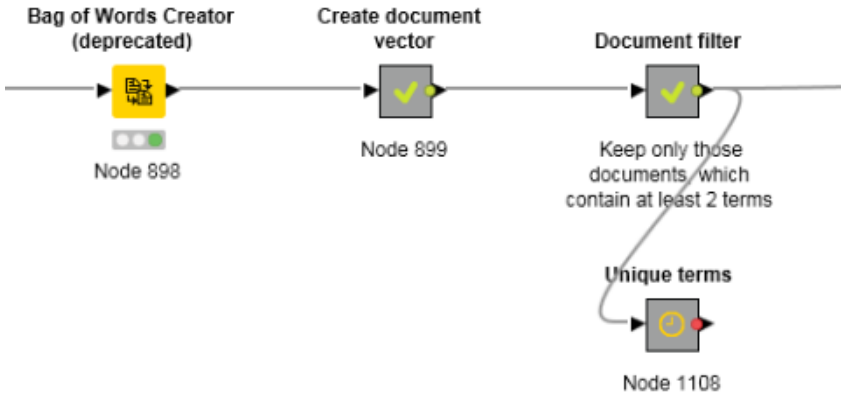
Quantization

- Transformation of qualitative data into quantitative data

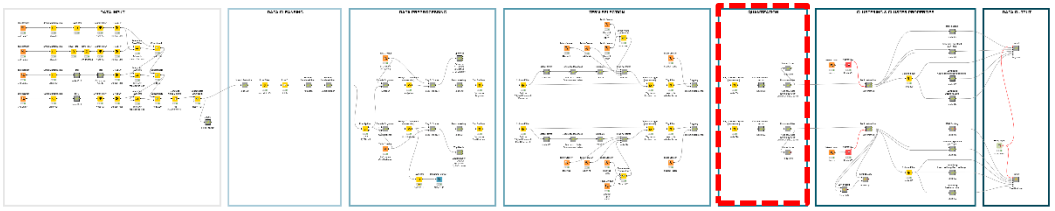
- Steps:

1. “Bag-of-words” creation
   - Think of it as a group-by on [Document, term]
2. “Document vector” creation
   - Transformation of the BoW table into a (bit-)matrix





KNIME Walkthrough



Quantization

- Transformation of qualitative data into quantitative data

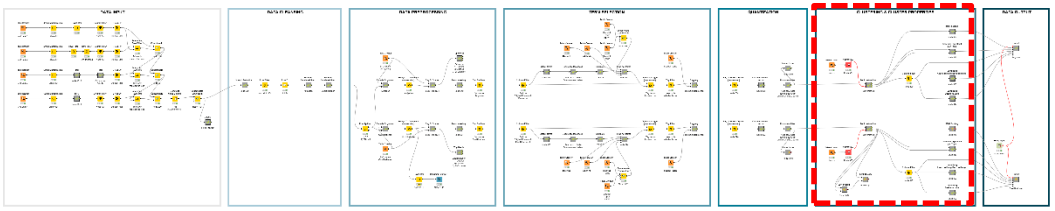
- Steps:

1. “Bag-of-words” creation
   - Think of it as a group-by on [Document, term]
2. “Document vector” creation
   - Transformation of the BoW table into a (bit-)matrix

Document vector

Row ID	D projekte	D bericht...	D regionen	D organis...	D risk
900020	1	1	1	1	0
900058	0	0	0	0	1
GQPAF96YU	0	0	0	0	0
MZ_166498	0	0	0	0	0
MZ_166712	0	0	0	0	0
MZ_169095	0	0	0	0	0
MZ_171360	0	0	0	0	0
MZ_172058	0	0	0	0	0
MZ_172485	0	0	0	0	0
MZ_172511	0	0	0	0	0
MZ_172556	0	0	0	0	0
MZ_172783	0	0	0	0	1
MZ_173065	0	0	0	1	0
MZ_173067	1	0	0	1	0
MZ_173135	0	0	0	0	0
MZ_173137	0	0	0	0	0
MZ_173138	0	0	0	0	0
MZ_173215	0	0	0	1	0

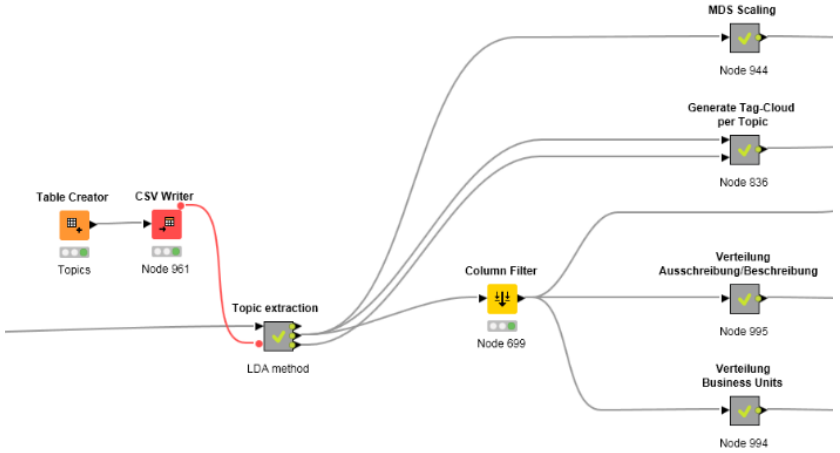
KNIME Walkthrough

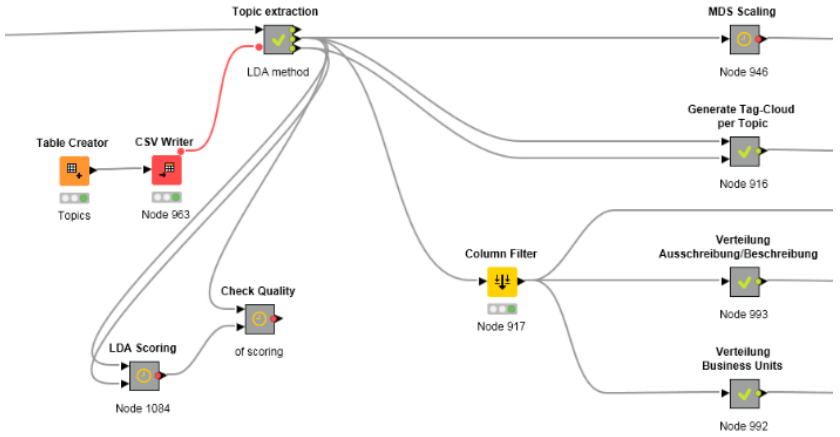


Clustering

- Grouping of jobs into clusters of comparable
  - Job tasks
  - Job qualifications

- Cluster Properties
  - Multi-dimensional scaling
  - Tag clouds
  - Intersection of tasks and qualifications



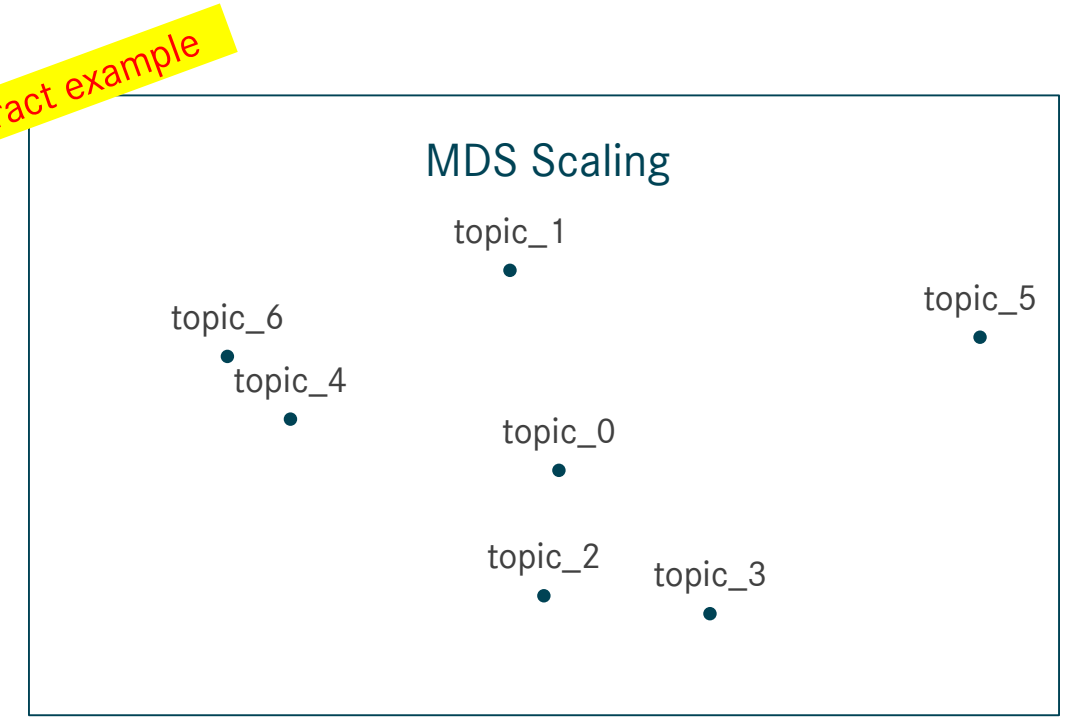


Application

Multi-dimensional scaling

- Visualization of the relative distances between clusters

- Relevance
  - Refinement of clustering
  - Ease of job shifts between clusters
  - …



Application

Word Clouds

- Visualization of important terms per cluster

- Relevance
  - Description of clusters
  - Easy to understand for non Data-Scientists
  - …

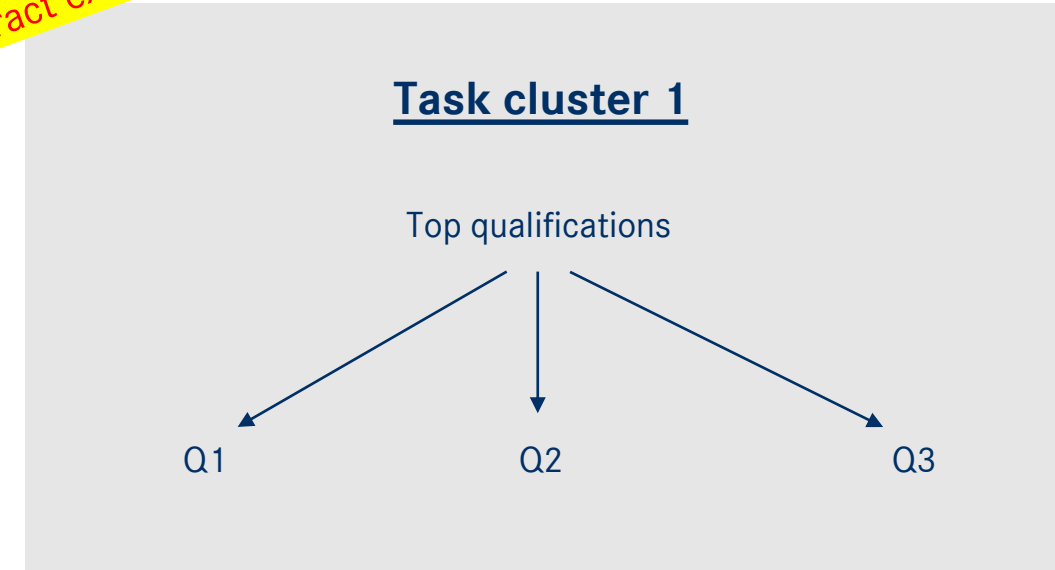
Application

Intersection of tasks and qualifications

- Extract the most relevant…
  - Qualifications per task cluster
  - Tasks per qualification cluster

- Relevance
  - Sharpen job advertisements
  - Identify specialist groups
  - …

Abstract example



Next Steps