Utilizing AI/ML methods for measuring data quality

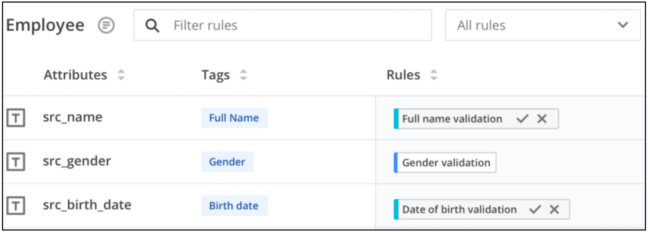
Bc. Michael Mikuš

Knowledge Engineering
Department of Applied Mathematics
Faculty of Information Technology
Czech Technical University in Prague

September 01, 2020

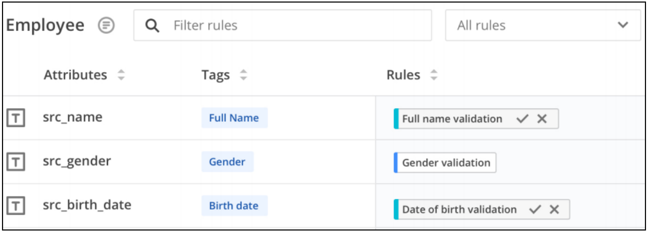
Supervisor: Ing. Tomáš Pajurek

Motivation



- *Current data quality measuring approaches:*
  - **expensive**, **expert** & **time-consuming work**
  - **manual effort** → **prone to error**
  - **DQ issues we know that exist**

MOTIVATION



- *Current data quality measuring approaches:*
  - **expensive**, **expert** & **time-consuming work**
  - **manual effort** $\rightarrow$ **prone to error**
  - **DQ issues we know that exist**

"What are innovative ways to measure data quality?"

GOALS OF THE THESIS

- **Theoretical framework (Data, Data quality & tools)**
- **Proposal Data Quality Measurement (DQM) – AI**
- **Conduct experiments**
- **Propose directions – AI in DQ**

THEORY - FUNDAMENTAL OBSERVATIONS

- **Data diversity & complexity**:
  - Abundant data types & complex data structures
  - Makes automation of DQM methods challenging

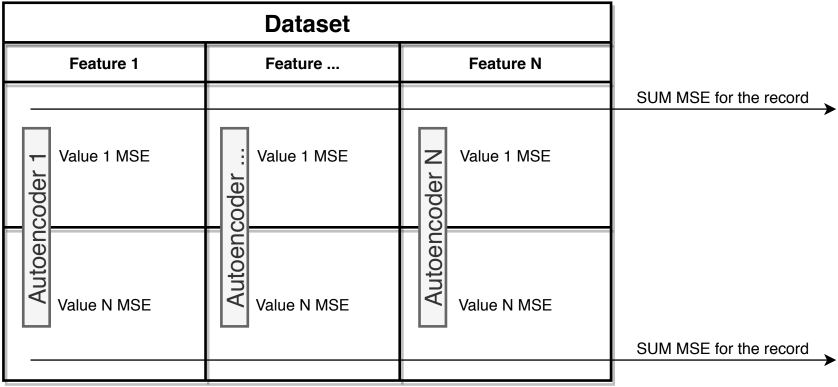
THEORY - FUNDAMENTAL OBSERVATIONS

- **Data diversity & complexity**:
  - Abundant data types & complex data structures
  - Makes automation of DQM methods challenging
- **Review #21 DQ tools:**
  - Do not take full advantage of AI
  - New categorization proposed (target group of users):
    - Regular users
    - Data engineering teams (+ AI $\approx$ promising potential)

Experiments

- Autoencoders
  - ”Universal approach to measure DQ?”
- Association Rule Mining
  - ”Why Association Rule Mining is not widely supported by general-purpose DQ tools?”

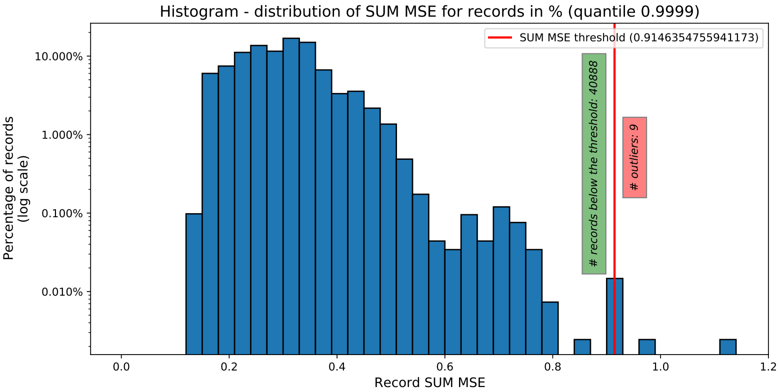
Experiment - Autoencoder - Design



- **Input:** tokenized text → numerical representation (Scaled)

Experiment - Autoencoder - Results

- Dataset (e.g. E-Commerce) + synthetic DQ issues



;

EXPERIMENT - AUTOENCODER - RESULTS

- **Example – Detected DQ issues:**
  - *Wrong currency*
    - '£16.50' vs '$16.50'
  - *Wrong name/category*
    - 'Coca-Cola' vs Names of clothes
  - *Wrong language*
    - German instead of English
  - *Values out of range*
  - *Wrong URL domain*
    - '*.uk*' vs '*.com*'
  - *Wrong datetime*
    - '`2017/07/26 18:27:10`' vs
    - '`2020-01-01T00:51:07Z`'
  - *Wrong data types* (string vs int)
  - *Wrong unit* (gram vs liter)

- **Example – Not detected DQ issues:**

EXPERIMENT - AUTOENCODER - RESULTS

- **Example – Detected DQ issues:**
  - *Wrong currency*
    - '£16.50' vs '$16.50'
  - *Wrong name/category*
    - 'Coca-Cola' vs Names of clothes
  - *Wrong language*
    - German instead of English
  - *Values out of range*
  - *Wrong URL domain*
    - '*.uk*' vs '*.com*'
  - *Wrong datetime*
    - '`2017/07/26 18:27:10`' vs
    - '`2020-01-01T00:51:07Z`'
  - *Wrong data types* (string vs int)
  - *Wrong unit* (gram vs liter)

- **Example – Not detected DQ issues:**
  - Short product description
  - Negative value
  - Wrong color
    - 'blakc' vs 'black'
  - Wrong price format
    - $24, 50 vs $24.50
  - Wrong hour value in datetime
    - '`2019-10-15T72:21:10Z`'
  - Wrong file extension
    - '.jpg' vs '.txt'

Experiment - Autoencoder - Results

- **Advantage:**
  - Promising universal nature of the approach
  - 1 parameter (reconstruction error threshold)
- **Disadvantage:**
  - Does not detect all DQ issues
  - Knowledge of max length of the encoded input in advance
- **Application potential:**
  - Preventive measurement of DQ $\rightarrow$ User notification
- **Alternative non-AI approach:**
  - Regular expression (flawless, task-dependent, expert knowledge, manual effort)

Experiment - Association Rule Mining

> ”Why Association Rule Mining is not widely supported by general-purpose DQ tools?”

Experiment - Association Rule Mining

"Why Association Rule Mining is not widely supported by general-purpose DQ tools?"

- **Apriori algorithm**
- **NLP for data preprocessing** (tokenization, lematization, stemming)

EXPERIMENT - ASSOCIATION RULE MINING

"Why Association Rule Mining is not widely supported by general-purpose DQ tools?"

- **Apriori algorithm**
- **NLP for data preprocessing** (tokenization, lematization, stemming)
- **Result:**
  - Rules were extracted
  - HOWEVER: the approach required significant preprocessing effort & focus on a specific question
  - **Complex data processing**

PROPOSED AI-BASED APPROACHES IN DQM

- **A Deep Learning Approach to Semantic Data Type Detection**
  - e.g. Location, Name, Year
  - Knowledge of feature data type $\rightarrow$ Adequately processed semantic information of values

PROPOSED AI-BASED APPROACHES IN DQM

- **A Deep Learning Approach to Semantic Data Type Detection**
  - e.g. Location, Name, Year
  - Knowledge of feature data type $\rightarrow$ Adequately processed semantic information of values
- **Automatically Generating Regular Expressions via Genetic Programming**
  - Automatic data format check

PROPOSED AI-BASED APPROACHES IN DQM

- **A Deep Learning Approach to Semantic Data Type Detection**
  - e.g. Location, Name, Year
  - Knowledge of feature data type $\rightarrow$ Adequately processed semantic information of values
- **Automatically Generating Regular Expressions via Genetic Programming**
  - Automatic data format check
- **Text Duplicates Detection via a ML Model with NLP Approaches**
  - e.g. Word2Vec, Fuzzy string matching

Conclusion

- **Comprehensive theory insight** – DQ & AI
- **Review of #21 DQ tools**
- **Conducted experiments**:
  - Autoencoder – preventive measurement of DQ
  - Association Rule Mining using NLP – challenging data preprocessing
- **Additional innovative approaches in DQM were proposed**
- **Future work:**
  - Extension of Autoencoder experiment
    - Advanced Autoencoders models (e.g. VAE)
    - Additional row metrics (e.g. inclusion # unique values in feature)
    - Semantic data types detection $\rightarrow$ Appropriately representation of semantic information in input value
  - Experiment with proposed DQM approaches

Thank you for your attention

Michael Mikuš
mikusmi1@fit.cvut.cz