TECHNISCHE UNIVERSITÄT DRESDEN

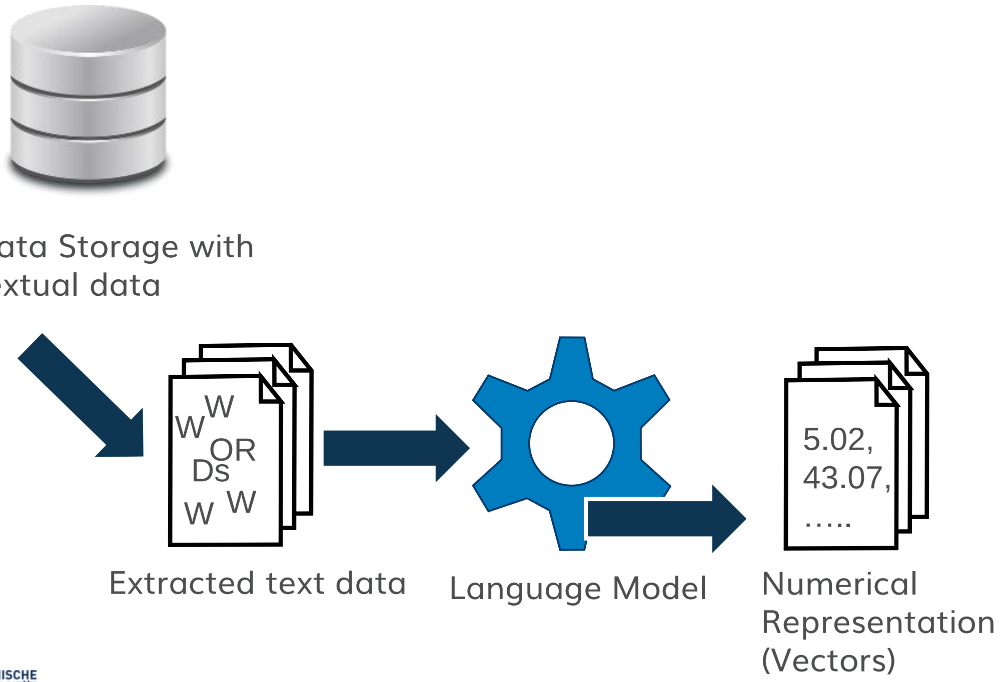
Dresden Database Systems Group

FacetE: Exploiting Web Tables for Domain-Specific Word Embedding Evaluation

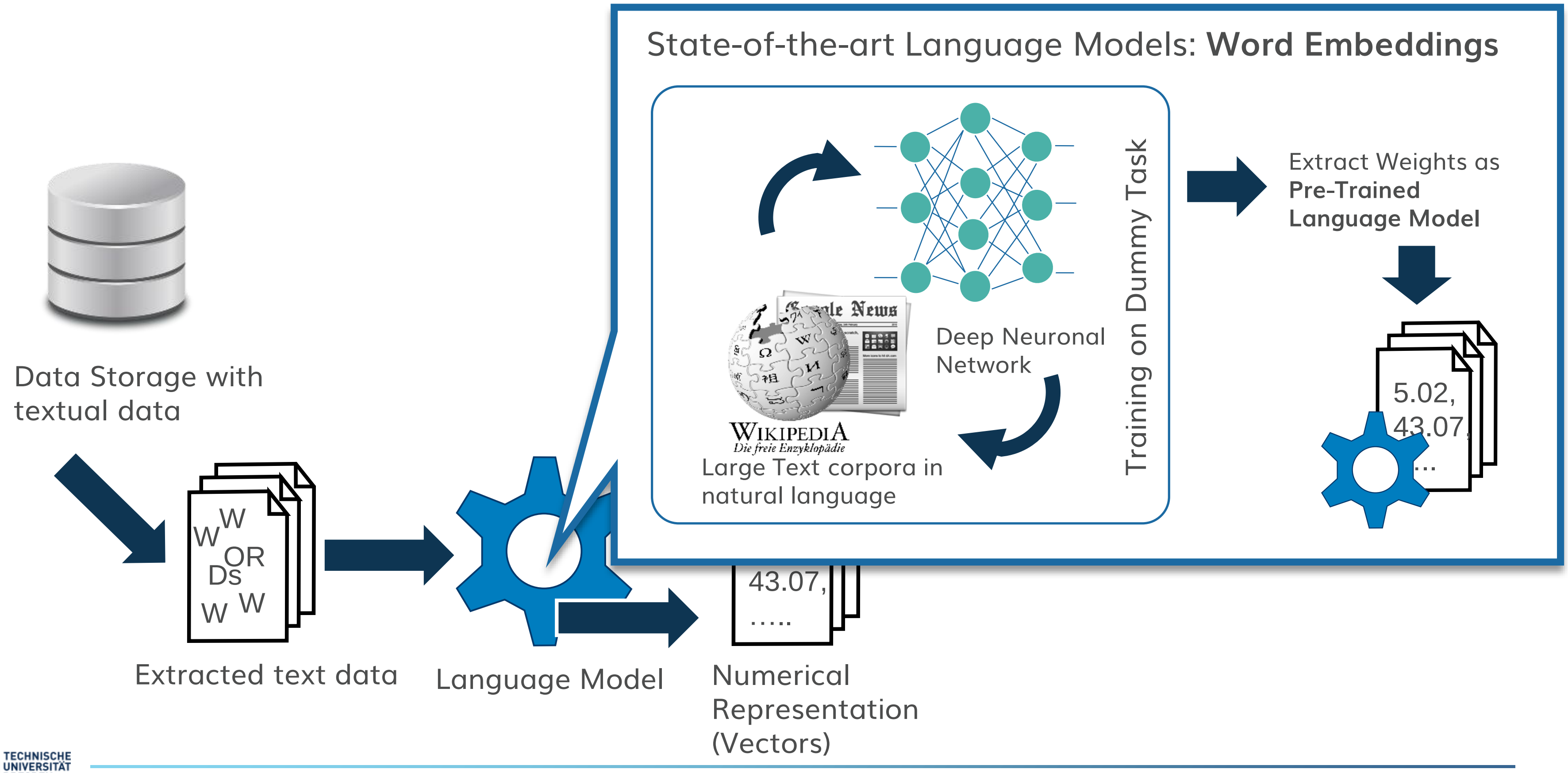
Michael Günther, Paul Sikorski, Maik Thiele, and Wolfgang Lehner

DBTest ‘20 Workshop at SIGMOD 2020 19.06.2020

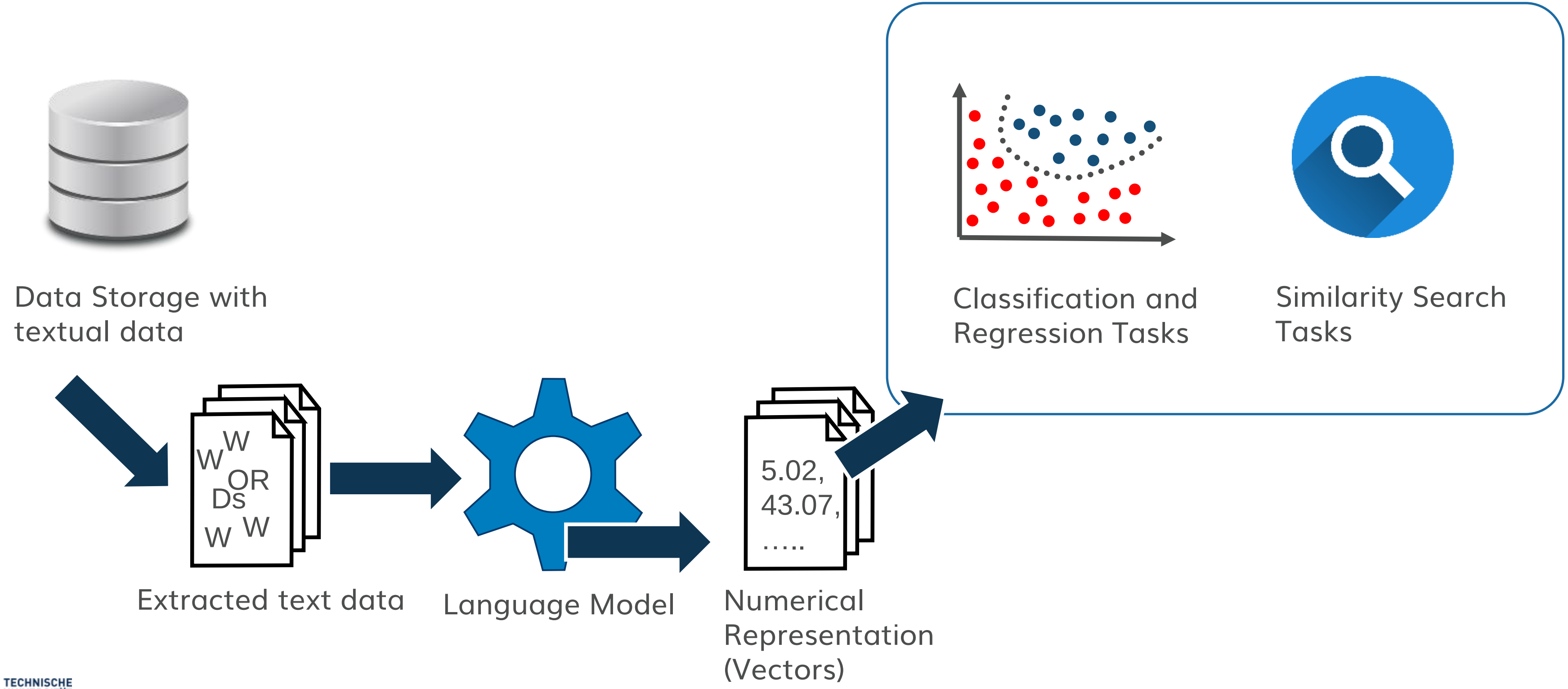
NLP Systems Workflow



NLP Systems Workflow



NLP Systems Workflow



Word Embedding for Systems

ML Systems	*Database Systems*	*Information Retrieval Systems*
		
▪ Utilize implicitly encoded knowledge from large text corpora ▪ Capture sematic similarities of text values	▪ Semantic text similarity queries ▪ Data exploration ▪ Data integration	▪ Semantic search ▪ Query Expansion ▪ Multi-lingual search

Choice of the word embedding model is crucial for the performance!

Evaluation of Word Embedding Models

Dresden Database Systems Group

Word Similarity

- Similar Words by cosine similarity of word vectors

$$sim_{cos}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

- Example: most similar to "king"?

$\rightarrow$ prince, man, and queen

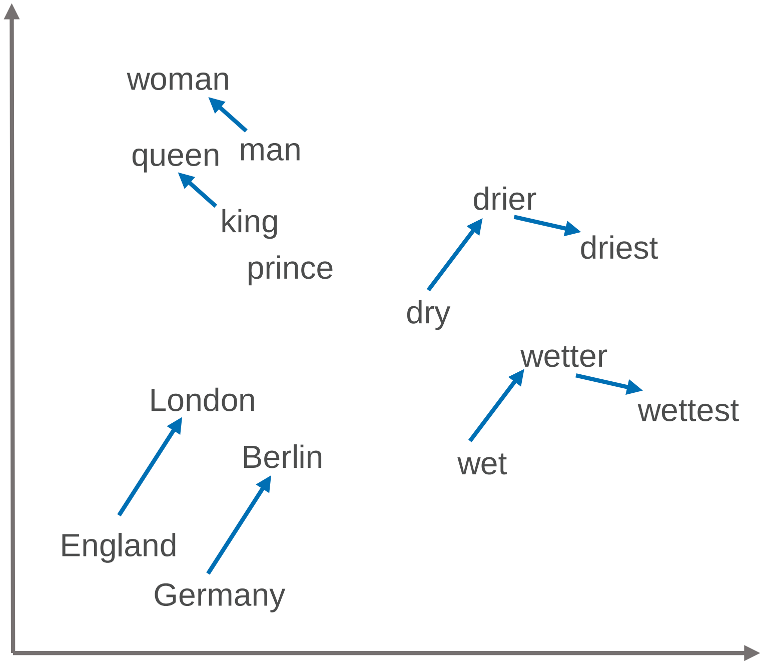
Analogy Queries

- Retrieve Similar Relations

$$a - b \approx c - ?$$

<u>**3CosAdd**</u>: $\underset{d \,\epsilon V \;\{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\}}{\arg\max}\, sim_{cos}(\boldsymbol{d}, \boldsymbol{c} - \boldsymbol{a} + \boldsymbol{b})$

- Example: man – woman ≈ king - ?
  $\rightarrow$ queen



Schematic Representation of Word Vectors

Evaluation of Word Embedding Models

Dresden Database Systems Group

Common Similarity Datasets

- **WS-353** 353 word pairs of **general domain knowledge** quantifying semantic **relatedness**
- **SimLex-999** 999 word pairs of **general domain knowledge** quantifying semantic **similarity**

Depend on human notion of similarity
→ Require human labeling effort

Common Analogy Query Datasets

- **Google Analogy** 550 semantic and syntactic relations, **mostly city-country relations**
- **MSR** 8,000 analogies of 800 **syntactic relations**

Facts of general domain knowledge
→ Automatic extraction possible

Similarity Eval*

Embedding Model	WS353	RW	…
CBOW	57.2	32.5	…
SkipGram	62.8	37.2	…
…	…	…	…

Analogy Eval*

Embedding Model	Semantic	Syntactic	Total
CBOW	57.3	68.9	63.7
SkipGram	66.1	65.1	65.6
…	…	…	…

* Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Limitations:

- Only small datasets
- Return a single value only
- Only general domain

Evaluation of Word Embedding Models

Dresden Database Systems Group

Common Similarity Datasets

- **WS-353** 353 word pairs of **general domain knowledge** quantifying semantic **relatedness**
- **SimLex-999** 999 word pairs of **general domain knowledge** quantifying semantic **similarity**

Depend on human notion of similarity
→ Require human labeling effort

Common Analogy Query Datasets

- **Google Analogy** 550 semantic and syntactic relations, **mostly city-country relations**
- **MSR** 8,000 analogies of 800 **syntactic relations**

Facts of general domain knowledge
→ Automatic extraction possible

Limitations:

- Only small datasets
- Return a single value only
- Only general domain

Design Goals:

- Large number of relations
- Flexible structure
- Multiple categories

Design Strategies:

- Extraction from millions of web tables
- Organization in facets
- Definition of categories

Dataset Design

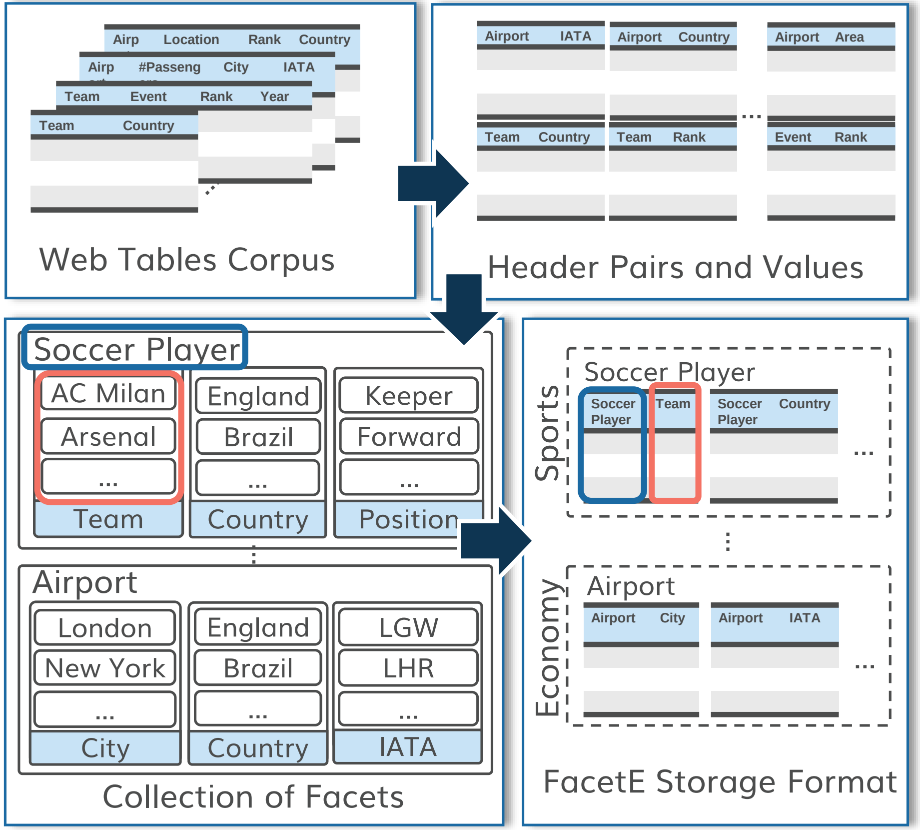
Dresden Database Systems Group

Data Source: Web Tables

- Large amount of knowledge
- General enough to be expected in pre-trained word embedding models
- Redundancy allows to exclude temporary facts (e.g. time dependent facts like home soccer team to visiting team)

Target Design: Facets

- Each Facet $F:\boxed{O}\rightarrow\boxed{V}$ assigns objects (e.g. Soccer Player) to values (e.g. Teams)
- Allows flexible construction of application specific evaluation datasets
- More flexible then hierarchical categorization



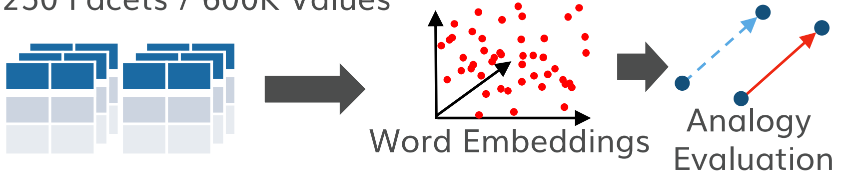
Extraction Pipeline

Dresden Database Systems Group

DWTC Dresden Web Table Corpus — 125M Web Tables

1. **Pre Filtering:** Frequency and Regex Filter, Facet Creation
2. **Soft Functional Dependencies:** Check contradiction of most frequent relation
3. **Post Filtering:** Filter by Pooling, Blacklist, …
4. **Categorization:** Assign facets to 8 broader categories

250 Facets / 600K Values

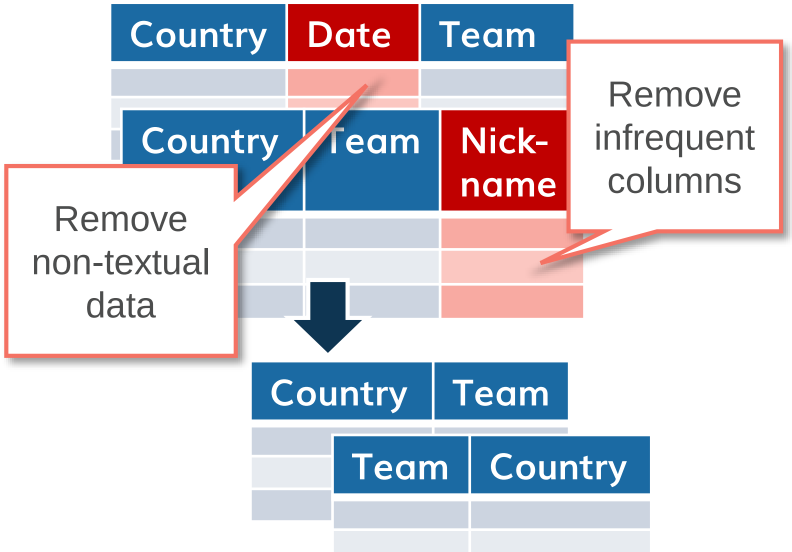


Extraction Pipeline

Dresden Database Systems Group

1) Pre-Filtering

- Filters infrequent and non-textual data of English tables

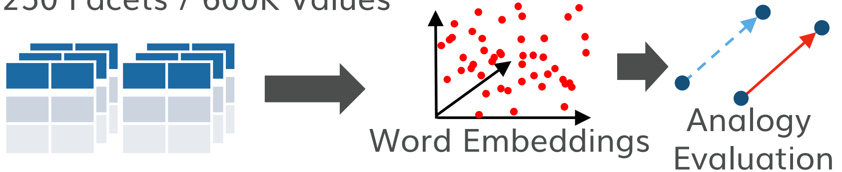


Column-Tuples
→ Basis for Facets

DWTC Dresden Web Table Corpus — 125M Web Tables

1. **Pre Filtering:** Frequency and Regex Filter, Facet Creation
2. **Soft Functional Dependencies:** Check contradiction of most frequent relation
3. **Post Filtering:** Filter by Pooling, Blacklist, ...
4. **Categorization:** Assign facets to 8 broader categories

250 Facets / 600K Values



Extraction Pipeline

Dresden Database Systems Group

2) Soft-Functional Dependencies

- Determine static facts

1) Determine most frequent relation pairs

2) Check on contradictions

$$SFD(o,v) = \frac{count(o,v)}{\sum_{v':(o,v')} count(o,v')}$$

$$SFD(Arsenal, England) = \frac{2}{3}$$

Team	Country
Arsenal	England
AC Milan	Italy
Juventus	Italy

Most frequent for "Arsenal"

Team	Country
Arsenal	United Kingdom
AC Milan	Italy

One Contradiction

Team	Country
AC Milan	Italy
Juventus	Italy
Arsenal	England

125M Web Tables

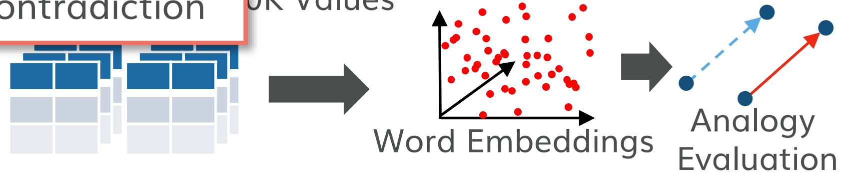
1 **Pre Filtering:** Frequency and Regex Filter, Facet Creation

2 **Soft Functional Dependencies:** Check contradiction of most frequent relation

3 **Post Filtering:** Filter by Pooling, Blacklist, …

4 **Categorization:** Assign facets to 8 broader categories

0K Values

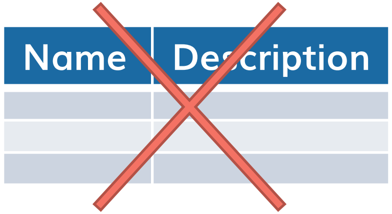


Extraction Pipeline

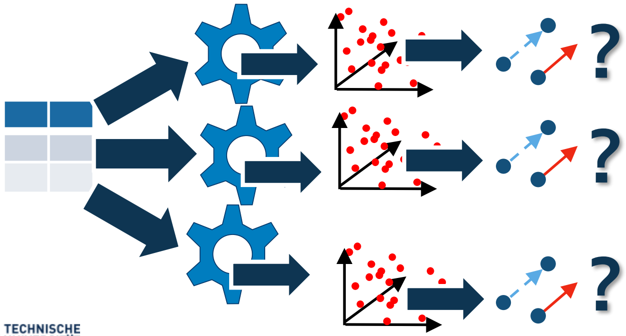
Dresden Database Systems Group

3) Post-Filtering

- **Blacklists**
  Remove too generic facets



- **Word Embedding Pooling**
  Retain only facets modeled by at least one word embedding model

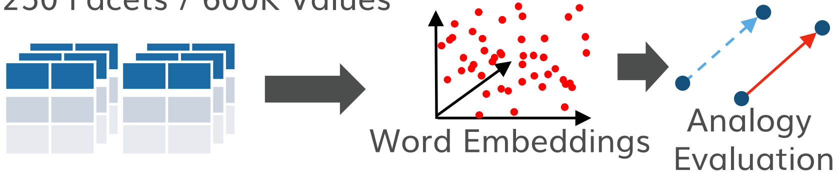


DWTC
Dresden Web Table Corpus

125M Web Tables

1. **Pre Filtering:** Frequency and Regex Filter, Facet Creation
2. **Soft Functional Dependencies:** Check contradiction of most frequent relation
3. **Post Filtering:** Filter by Pooling, Blacklist, …
4. **Categorization:** Assign facets to 8 broader categories

250 Facets / 600K Values

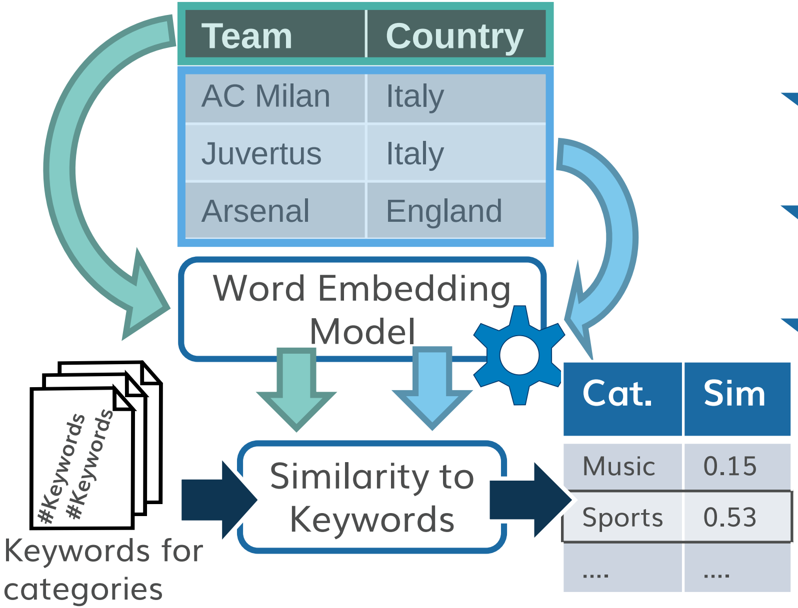


Extraction Pipeline

Dresden Database Systems Group

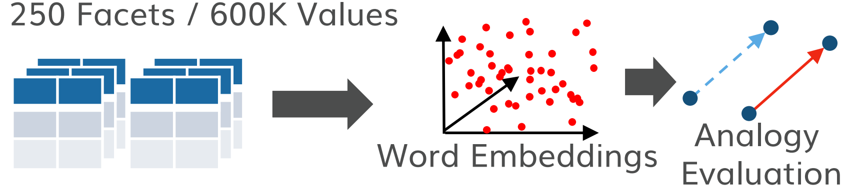
4) Categorization

- Assign each of the 250 facets on of 8 broader categories (e.g. geographic, music, sports, ...)



DWTC Dresden Web Table Corpus — 125M Web Tables

1. **Pre Filtering:** Frequency and Regex Filter, Facet Creation
2. **Soft Functional Dependencies:** Check contradiction of most frequent relation
3. **Post Filtering:** Filter by Pooling, Blacklist, ...
4. **Categorization:** Assign facets to 8 broader categories



TECHNISCHE UNIVERSITÄT DRESDEN

Evaluation

Dresden Database Systems Group

Evaluation of Categories

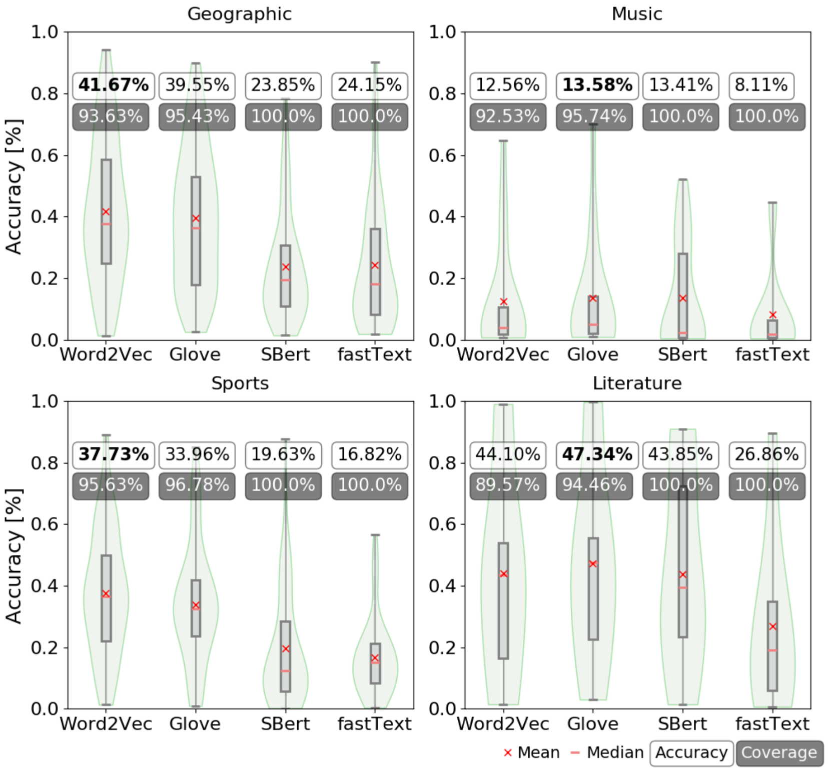
Setup

- 4 Pre-trained word embedding models: GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- Selection of 4 FacetE categories

Calculation

- Select facets $F: O \rightarrow V$ from the categories
- Determine the value V for each object O with 3CosAdd analogy method
- Calculate amount of correctly assigned values
- Calculate average in each category

Coverage: For some text values word embedding models can not determine a vector



Evaluation of 4 Categories

Evaluation

Dresden Database Systems Group

Evaluation of Categories

Setup

- 4 Pre-traine GloVe, Word2 SentenceBert
- Selection o

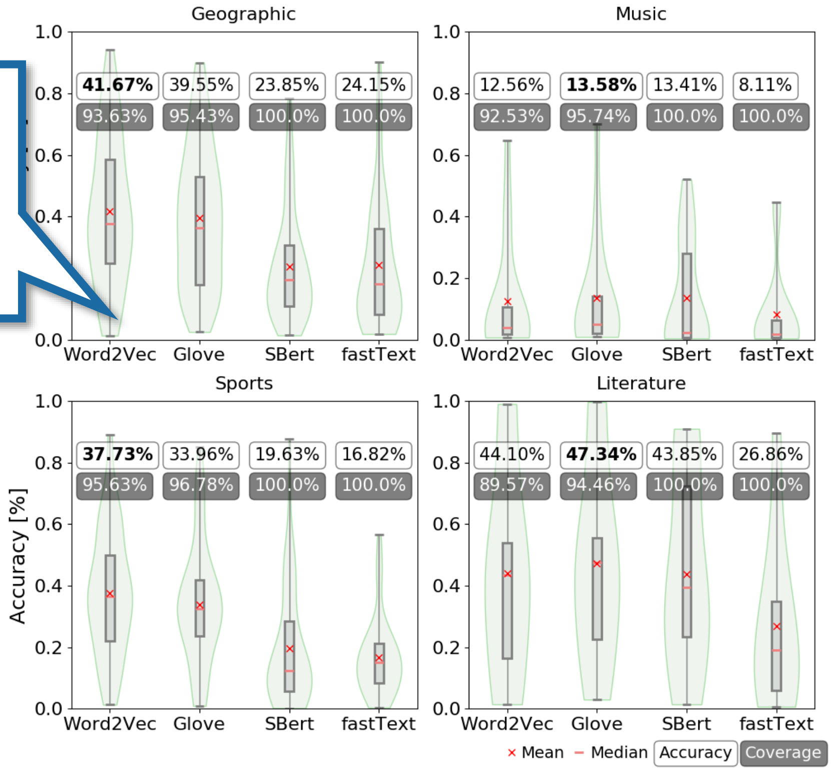
Calculation

- Select facets $F: O \rightarrow V$ from the categories
- Determine the value V for each object O with 3CosAdd analogy method
- Calculate amount of correctly assigned values
- Calculate average in each category

Coverage: For some text values word embedding models can not determine a vector

Observation

- No single best model
- High Coverage



Evaluation of 4 Categories

Evaluation

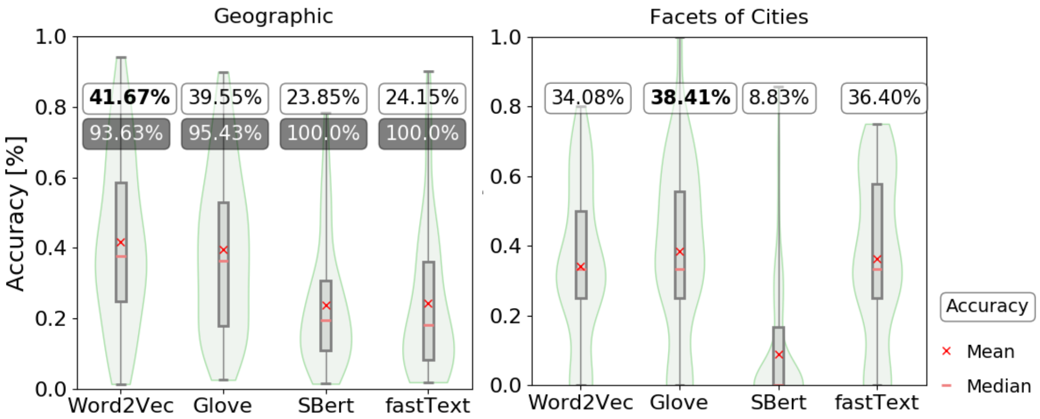
Evaluation of a Single Object Set

Setup

- 4 Pre-trained word embedding models:
  GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- **Selection of all facets for cities**

Calculation

- Determine the value V for each object O with 3CosAdd analogy method
- Calculate amount of **correctly assigned values for each city name**
- Calculate average across all objects



Evaluation of a Single Object Set - Cities

Evaluation

Evaluation of a Single Object Set

Setup

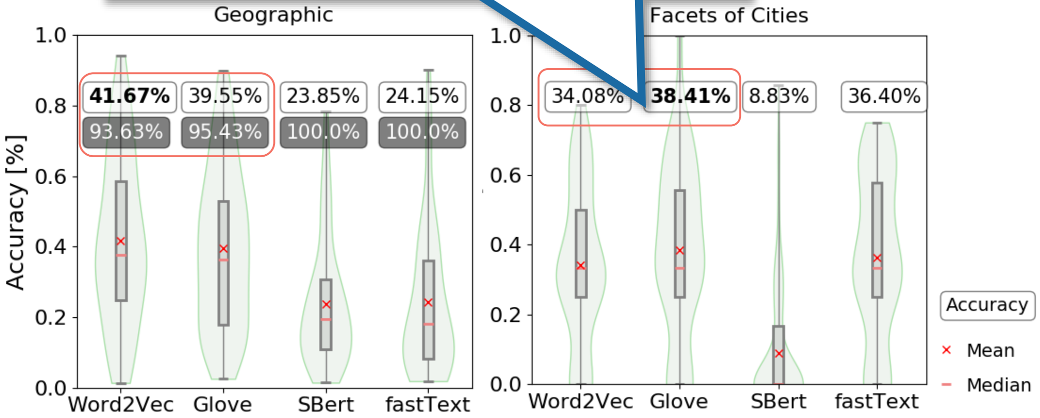
- 4 Pre-trained word embedding models:
  GloVe, Word2Vec-SkipGram, fastText, SentenceBert
- **Selection of all facets for cities**

Calculation

- Determine the value V for each object O with 3CosAdd analogy method
- Calculate amount of **correctly assigned values for each city name**
- Calculate average across all objects

Observation

Word2Vec performs better on geographic data, however GloVe has a better representation of cities



Evaluation of a Single Object Set - Cities

Conclusion

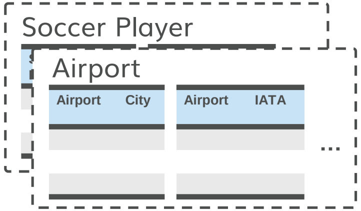
Web Table Extraction Pipeline

- Web Tables are a good resource for structured relations of general common knowledge
- Pipeline is able to process millions of tables

→ Reusable for other table corpora

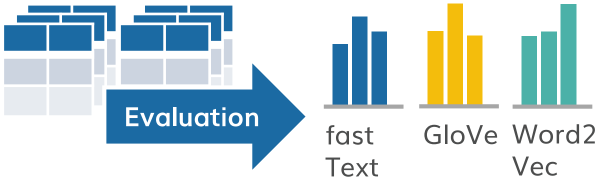
Facet Structure

- Enables flexible construction of evaluation datasets
- Evaluation of different granularity Single Facts (e.g. City → Country), Objects (e.g. Cities) or Domains (e.g. Geographic)



Evaluation of Common Word Embedding Models

- Large differences in accuracy values on different domains
- No best model for all cases



FacetE Dataset: https://www.kaggle.com/guenthermi/facete