



Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim Geisenheim

WIKIFICATION

Machine Learning

VORGETRAGEN VON TIRAS ZEMICAEL,
MARTIN ZAKARIAN KHENGI

Hochschule **RheinMain** | 24.02.2017

1. Einleitung
2. Datenmenge
3. Link Detection
4. Commonness, Relatedness & Contextquality
5. Link Disambiguation
6. Categorizer
7. Vorführung
8. Fazit

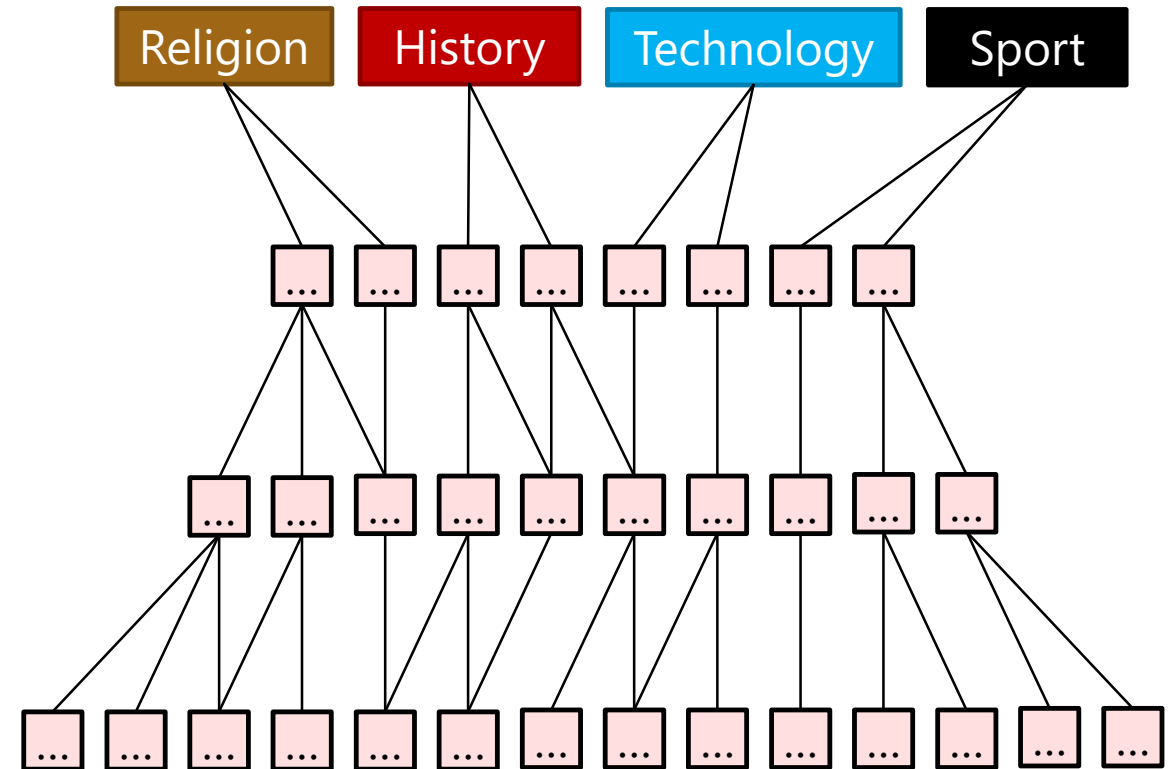
- Autoren Kategorisieren Artikel und erstellen Verweise
- Automatisiertes Verweisungsverfahren von Artikeln
- Verweisen von Schlüsselwörter in Artikeln
- Kategorisieren von Artikeln

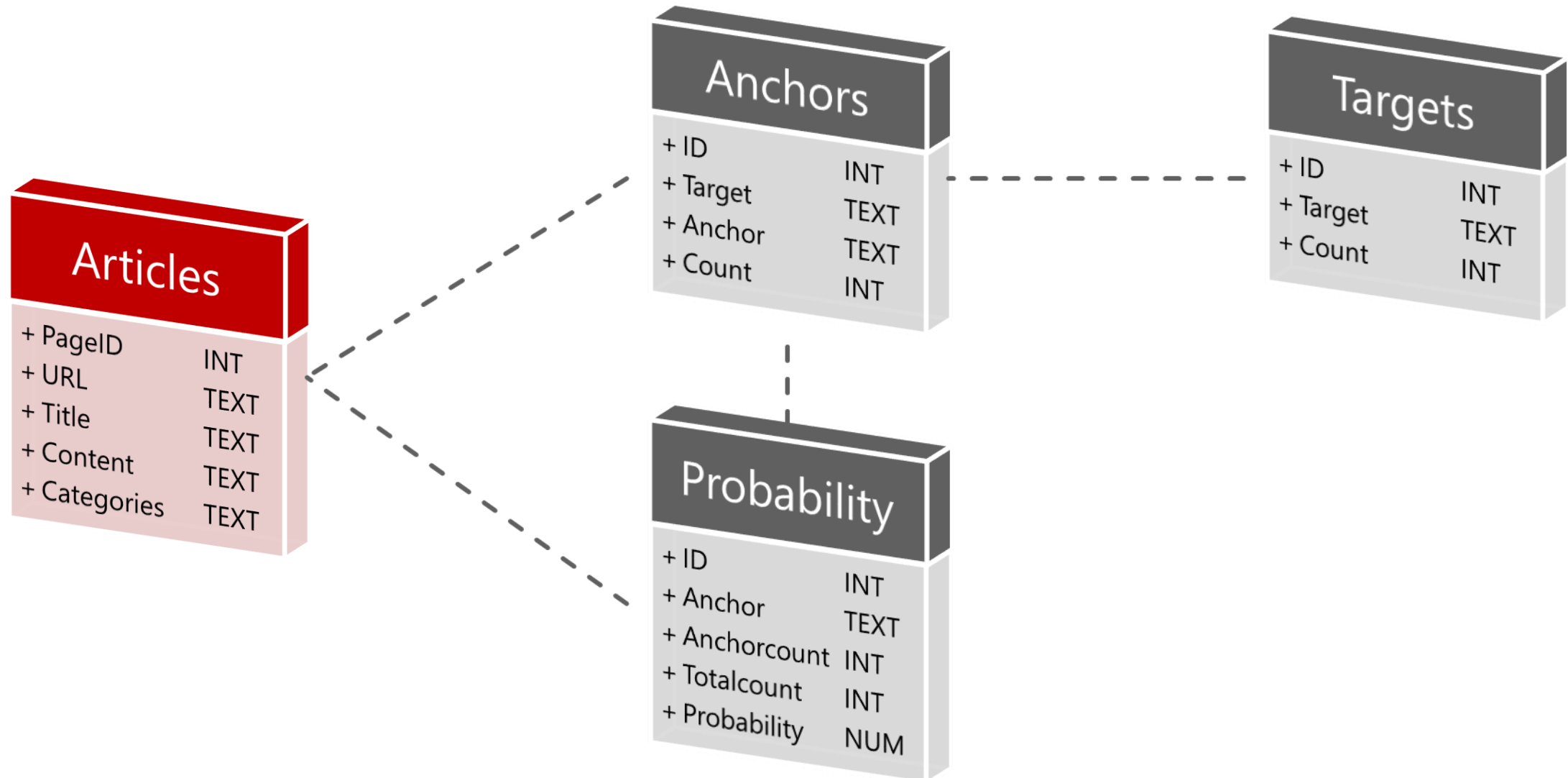


- Entgegennahme von Dokumenten
- Analyse des Dokuments
- Verweisungen erstellen von Schlüsselwörtern
- Dokument Kategorisieren
- Resultierendes Dokument Ausgeben

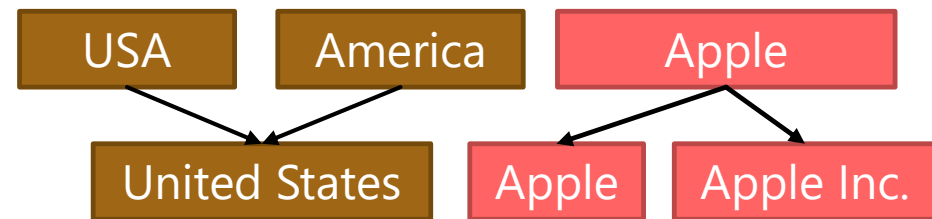


- PetScan von Magnus Manske
 - „Main topic classifications“
 - Tiefe von 3
 - Mindestens 50 Links
 - Keine Listen oder Glossare
- 14.193 Wikipedia Artikel
- 44.441 eindeutige Kategorien
- Python Wikipedia API
- SQLite Datenbank

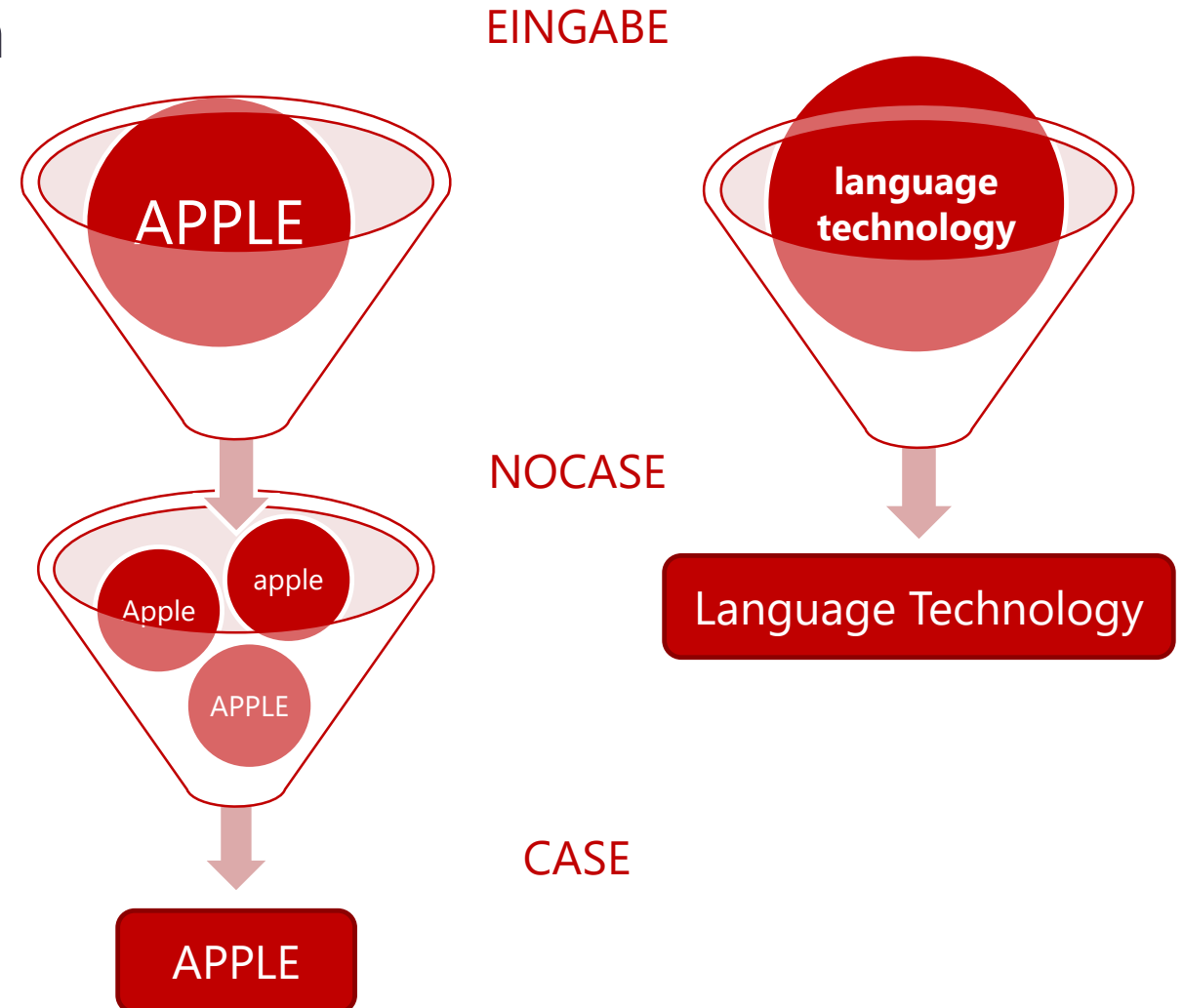


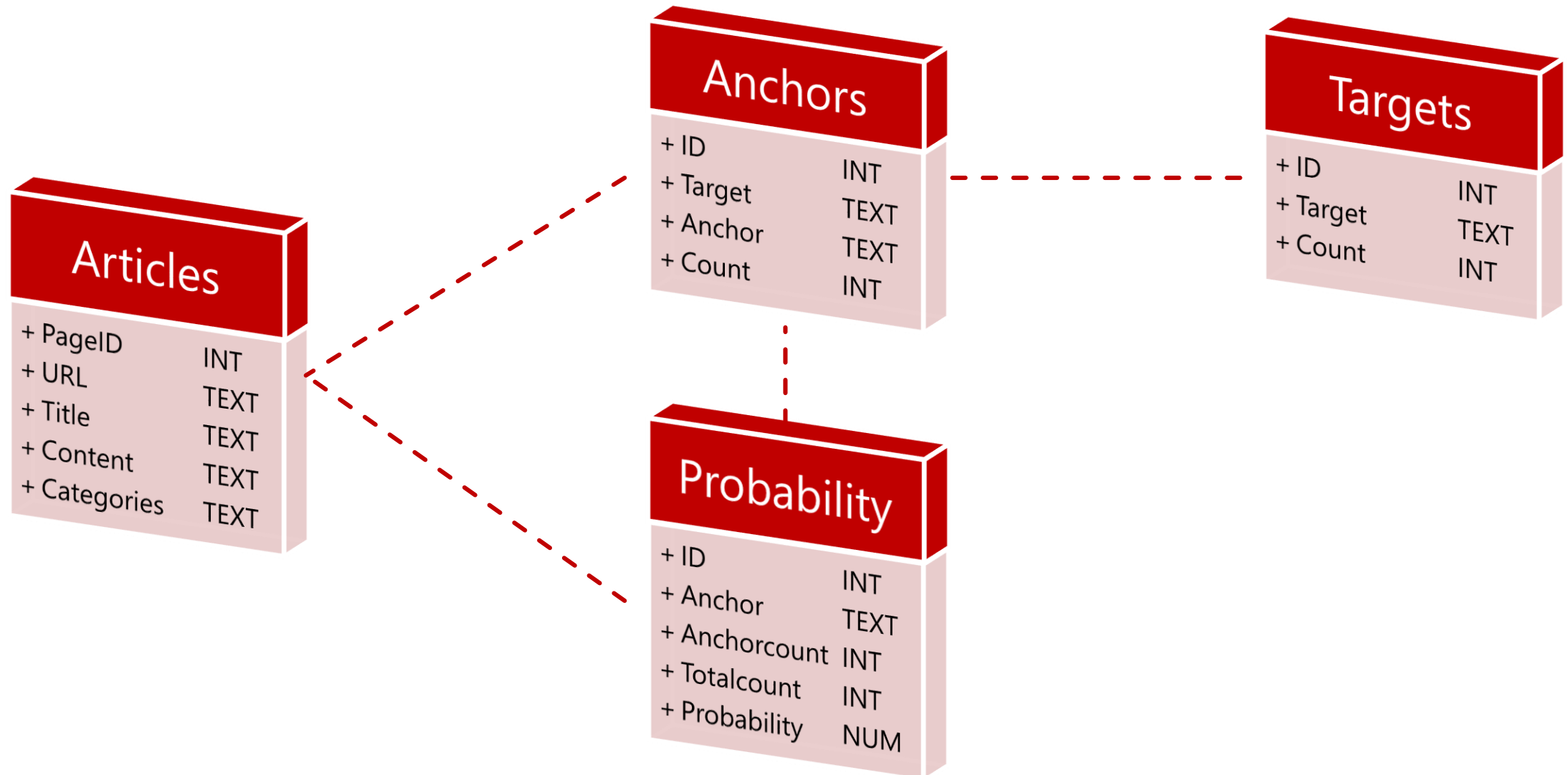


- Verlinkte Terme in Datenmenge
- Link Tags in HTML
- Unterscheidung zwischen Link und Term
- Link Probability
 - $\frac{\text{Anzahl der Linkvorkommen}}{\text{Anzahl der Vorkommen}}$
- Fester Schellenwert
 - 40% beim trainieren
 - 30% beim anwenden
- Word Frequency List



- Korrektur von Nutzereingaben
- Recall von 55%
- ✓ 1 Kandidat NOCASE
- ✓ Alle Kandidaten CASE





Commonness

- Wikimedia API Backlinks
- Anzahl der Verweise auf eine Wikipedia Seite

Relatedness

- TfidfVectorizer
- Termgewichtung über die Datenmenge
- On-The-Fly Berechnung

Contextquality

- Prozentualer Anteil der Häufigkeit von Linkkandidaten innerhalb des Dokuments
- Mittelwert aller **Relatedness** von allen eindeutigen Links zum Text des Nutzers

Für mehrdeutigen Term „News“		
Kandidaten	Commonness	Relatedness
News	74.0432	43.8893
Broadcast journalism	43.2349	40.2546
ITV News	16.5801	37.7289

- Aufteilung in Eindeutige und Mehrdeutige Verweise
- Eindeutige Verweise werden für die Disambiguation nicht betrachtet
- Mehrdeutige Verweise werden für Disambiguation vorbereitet

- Mehrdeutige Verweise bekommen eine Tabelle mit ihren jeweiligen Verweisen
- Commonness und Relatedness werden errechnet
- Qualität des Dokuments wird ermittelt

News

- News {...}
- Broadcast journalism {...}
- ITV News {...}

- Trainingsmenge von 9 und Testmenge von 1 Artikeln

Features

```
[  
  {Commoness: 12.4851 , Relatedness: 4.1845 , Contextquality: 15.12}  
  {Commoness: 4.7389 , Relatedness: 43.8893 , Contextquality: 15.12}  
  {Commoness: 74.0432 , Relatedness: 0.4453 , Contextquality: 15.12}  
  {Commoness: 43.2349 , Relatedness: 12.7748 , Contextquality: 15.12}  
  {Commoness: 23.6309 , Relatedness: 2.934 , Contextquality: 23.41}  
  {Commoness: 16.5801 , Relatedness: 72.3942 , Contextquality: 23.41}  
  {Commoness: 64.1129 , Relatedness: 4.2909 , Contextquality: 65.31}  
  {Commoness: 7.7365 , Relatedness: 37.7289 , Contextquality: 65.31}  
]
```

Labels

```
[  
  0  
  0  
  1  
  0  
  0  
  1  
  0  
  1  
]
```

LINK DISAMBIGUATION CLASSIFIER (TEST)

15

- Logistic Regression
- Score: 77,5 %
- Decision Tree
- Score 71,8 %

News

- News: 1
- Broadcast journalism: 1
- ITV News: 0

■ Prediction Probability

News

- News: 0.756 
- Broadcast journalism: 0.654 
- ITV News: 0.348 

- **ZIEL:** Kategorisieren mit Text der Nutzereingabe
- Classifier lernt die Texte Artikel aus der Datenmenge
- Aus Text wird mithilfe eines Tokenizers ein Bag of Words generiert
- Tokenizer Parameter:
 - Min. vorkommen innerhalb Artikel → 10
 - Min. vorkommen in Artikelmenge → 50
- Kategorielliste aus Datenmenge generiert
- Min. 5 vorkommen

■ Datenaufteilung:

- 13.326 Trainingsdaten
- 271 Testdaten

■ FEATURES:

- Dictionary
 - Key → pageID
 - Value → Bag of Words (type: dict)

■ LABELS:

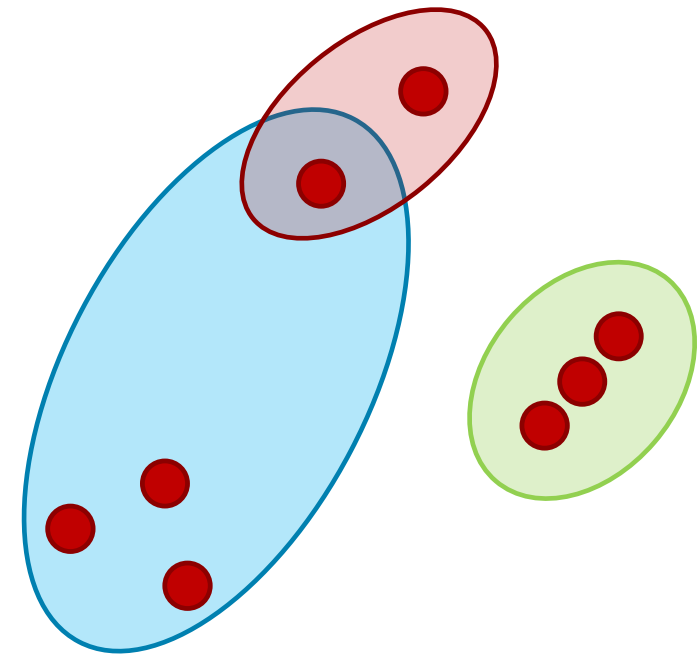
- Dictionary
 - Key → pageID
 - Value → Kategorien (typ: list)

```
{
    '112233':
        {
            'heaven' : 17,
            'world': 3,
            ...
        },
    ...
}
```



```
{
    '112233': ['religion', 'church'],
    ...
}
```

- Multilabel Classification Problem
- OneVsRest Classifier
- Linear Support Vector Classification
- DictVectorizer für BoWs
- MultiLabelBinarizer für Labels
- Präzession von 22%





Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim Geisenheim

VORFÜHRUNG

- DJANGO
- UPLOAD VON DOKUMENTEN
- TEXTEDITOR
- WIKIFICATION
- ERGEBNISSE

VIELEN DANK FÜR IHRE
AUFMERKSAMKEIT

FAZIT
FRAGEN?



EINE PRÄSENTATION IM RAHMEN DER VERANSTALTUNG
LV-NR.: 8581 – VERTIEFUNGSFACH - MACHINE LEARNING
VORGETRAGEN VON TIRAS ZEMICAEL UND MARTIN ZAKARIAN KHENGI