

# Research Knowledge Discovery

---

## Detektion von Experten und Aufbau eines Recommender-Systems für die TH Köln

### Projektgruppe 2 - DIS - 18

#### Teilnehmer

- Pia Störmer
- Matteo Meier
- Jüri Keller
- Martin Bilko
- Sascha Gharib
- Verena Pawlas
- Michelle Reiners
- Saskia Brech
- Fabian Ax
- Constantin Krah
- Leon Munz
- Andreas Kruff
- Fabian Gitzler
- Jonas Dudda
- Annika Füssel

# Agenda

## 1. Projektplanung - Constantin

- 1.1. Überblick
- 1.2. Projektmanagement & Organisation

## 2. Methoden und Projektpipeline - Saskia

- 2.1. Browsing Ansatz
- 2.2. LDA
- 2.3. Zusammenhang der Teilsysteme

## 3. Erstellung des Datensatzes - Pia & Fabian

- 3.1. BA Crawler
- 3.2. Interne Quellen (Ansatz TH-Suche, Name Matching)

# Agenda

## 4. Vorverarbeitung des Datensatzes - Sascha, Andreas, Leon

- 4.1 Datenbank
- 4.2 Pre Processing Pipeline
  - 4.2.1 Designentscheidungen
  - 4.2.2 Problematik Implementierung

## 5. Extraktion der Topics aus den Dokumenten - Michelle

- 5.1 LDA-Pipeline
- 5.2 Korpusauswahl
- 5.3 Optimale Anzahl der Topics bestimmen
- 5.4 LDA-Pipeline
- 5.5 Topicverteilung (500 Topics)
- 5.6 Topic-Dokument-Relation
- 5.7 Herausforderungen, Probleme + Zwischenergebnisse

# Agenda

## 6. Intellektuelle Erschließung der Topics - Fabian G

### 6.1 Intellektuell

## 7. Experten Empfehlung - Matteo und Martin

### 7.1 Expert Ranking

### 7.2 Das Frontend-System

#### 7.2.1 Probleme und zukünftige Arbeitspakete

## 8. Ergebnisse & Ausblick - Annika

### 8.1 Ergebnisse & Ausblick

# 01 —

## Projektplanung

# 1.1 Überblick

## Datenakquise und Konzeption (Semester 1)

### Kick-Off & Phase 1

- Erarbeitung möglicher Umsetzungsideen
- Erarbeitung konkreter User-Stories & Arbeitspakete
- Klein-Gruppenaufteilung
- Literaturrecherche

### Phase 2 & 3

- Präsentation der ersten Ergebnisse
- Datenspeicherung, Externe Quellen, Optimierung Prototyp
- Vorbereitung der Zwischenpräsentation

### Phase 4 - 5 / Next Steps

- Diskussion und Integration des Feedbacks aus Zwischenpräsentation
- Neuauswahl an User-Stories aus dem Ideenpool
- Neu-Aufteilung der Arbeitsgruppen
- Abschluss der Datenakquise & Konzeption

## Technische Umsetzung (Semester 2)

### Schritt 0 - Scraper Optimierung

- Weiterentwicklung Scraper (Intern & Extern)



### Schritt 1 & 2 - Topic Modeling TH Topics & Dokumente

- Methoden der Erschließung
- Topic Modeling
- Preprocessing Pipeline



### Schritt 3 & 4 - Verknüpfung der Topic Generierung & Personen, sowie Dokumente

- Zuordnung der Personen zu den erstellten Topics Modelle & Entwurf



### Schritt 5 - Basis Entwurf Frontend

- Model Entscheidung -> Browsing
- HTML Prototyp erstellen
- Script für Generierung der Index.html

## 1.2 Projektmanagement & Organisation

# GitHub



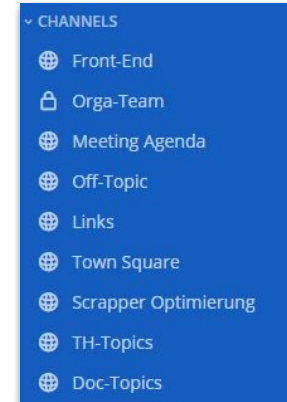
- Archiv für Files
- Technische Dokumentation

Projektmanagement (Prozessabbildungen & To-Do Listen)



Austausch & Koordination

- Kommunikation
- Terminkoordination



### Was hat gut funktioniert?



- Timeline “im Blick”
- Visualisierungen
- Teilen von Literatur & Terminen

<https://github.com/DIS-KD-Project>  
<https://chat.iim.th-koeln.de/dis18/channels>  
[https://miro.com/app/board/o9J\\_lJz7Ko8=](https://miro.com/app/board/o9J_lJz7Ko8=/)

### Was hat nicht gut funktioniert?



- Qualitätskontrolle
- Faire ausgewogene Aufgabenverteilung
- Proaktivität

# 02 —

## Methoden & Projektpipeline



# 2.1 Browsing Ansatz



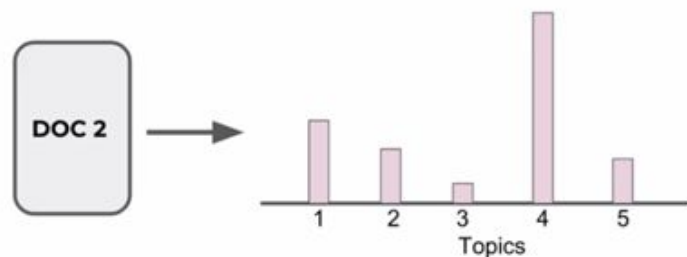
The screenshot shows the official website of the U.S. Department of Labor, Bureau of Labor Statistics. The header includes the agency's name and logo, a navigation bar with links like 'About BLS', 'Jobs in BLS', and 'Economic News Releases', and a row of small photos of diverse people. The main content area is organized into several columns:

- Left Column:** Contains links to major categories such as 'Inflation & Consumer Spending', 'Wages, Earnings, & Benefits', 'Productivity', 'Safety & Health', 'International', and 'Occupations'. Each category has a list of sub-links.
- Center Column:** Features a 'Latest Numbers' section with a yellow header. It displays key indicators like CPI, Unemployment Rate, Payroll Employment, Average Hourly Earnings, PPI, ECI, Productivity, and U.S. Import Price Index, each with a small trend arrow and the most recent data point.
- Right Column:** Includes sections for 'Employment & Unemployment', 'At a Glance Tables', 'Publications', 'Research', and 'Industries', each with a list of relevant links.

Quelle: Kick-Off Folien SoSe21

## 2.2 LDA

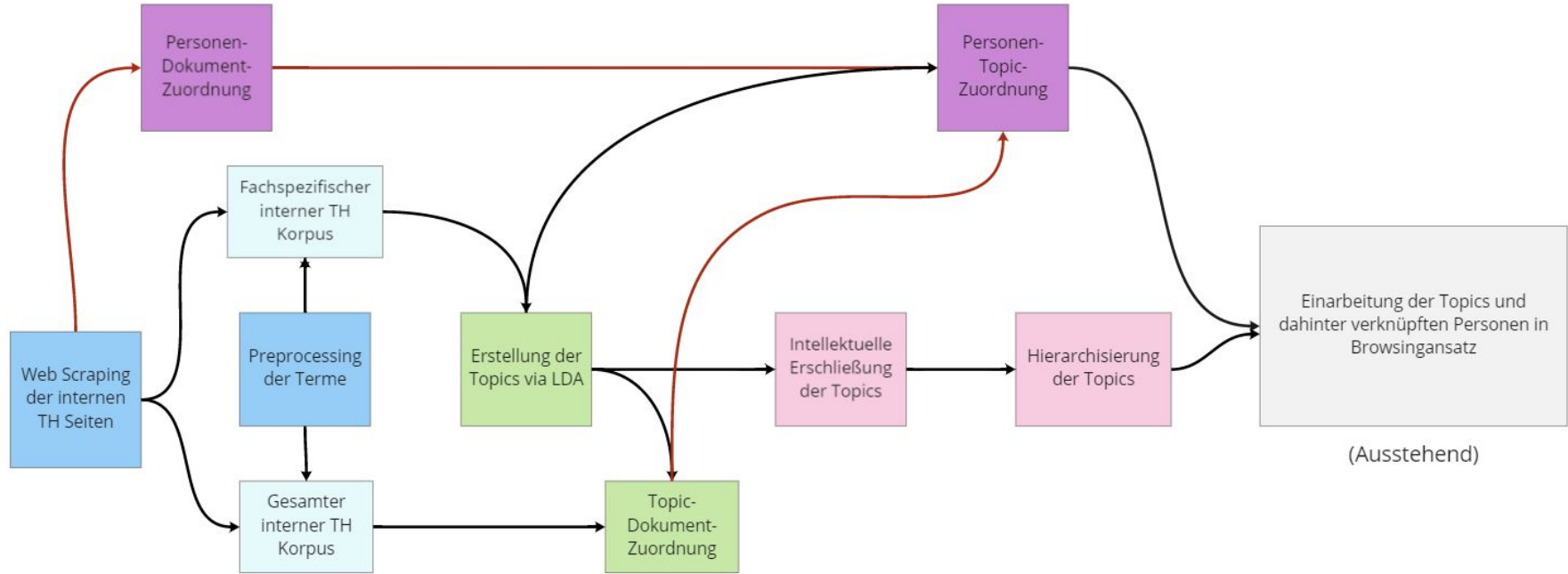
- Dokumente sind eine Mischung aus verborgenen Topics (latent topics)
- Jedes Topic ist eine Mischung von Termen
- Documents are probability distributions over latent topics.



Quelle:

<https://www.udemy.com/course/nlp-natural-language-processing-with-python/>

## 2.3 Zusammenhang der Teilsysteme



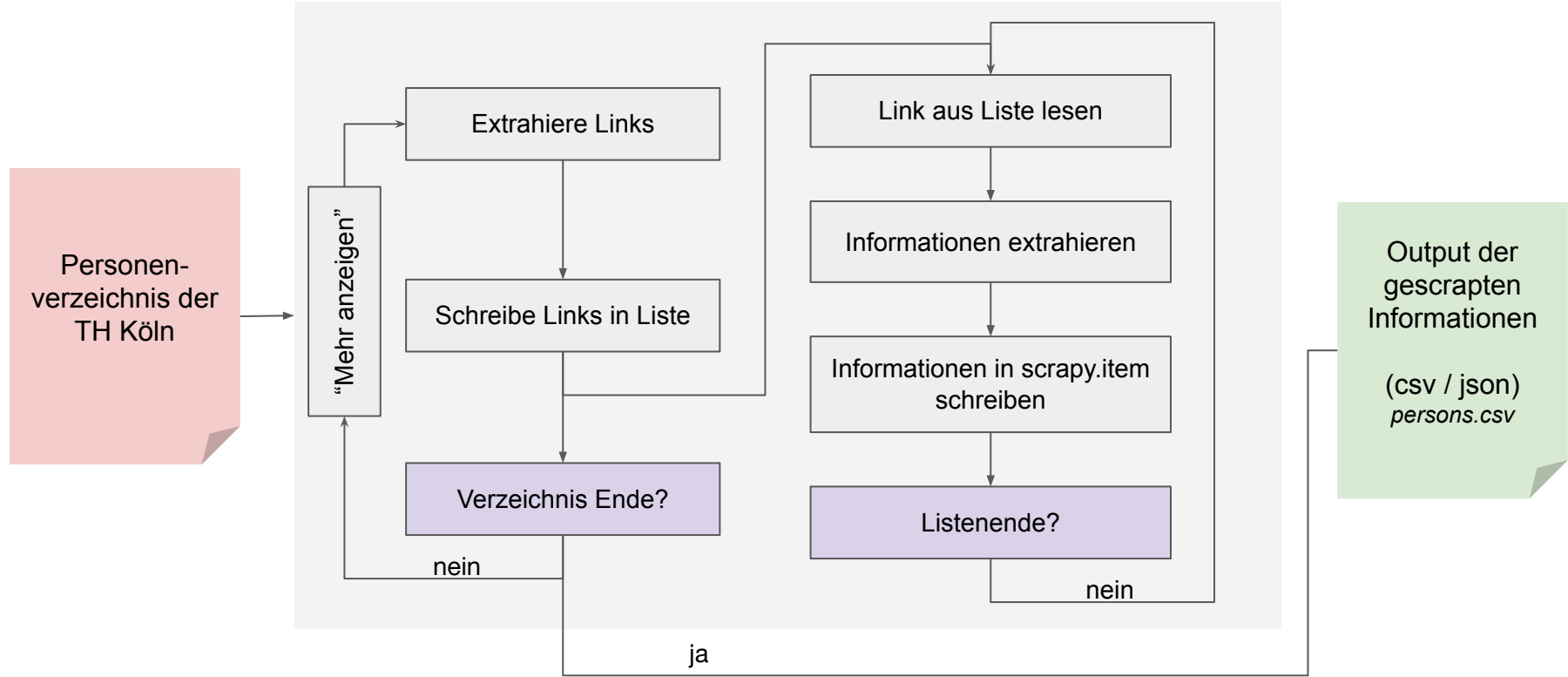
(Ausstehend)

● = wesentliche Verknüpfungen

# 03 —

## Erstellung des Datensatzes

## 3.1 BA Crawler

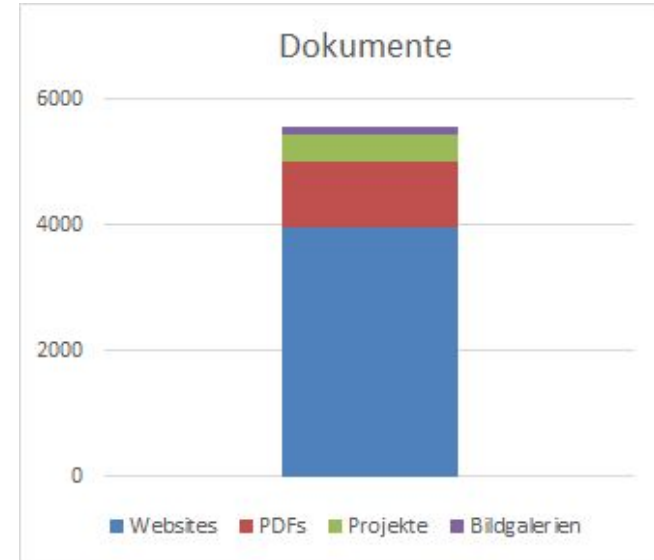


## 3.2 Interne Quellen - Datensatz

- 3950 Webseiten
- 1068 PDFs
- 138 Bildergalerien
- 407 Projekte

Dokumente mit mindestens einer Zuordnung zu einer relevanten Person:

- Professoren
- Lehrbeauftragte
- Lehrkräfte mit besonderen Aufgaben



## 3.2 Interne Quellen - Datensatz: Dokumente

document_id	Eindeutige Dokument ID
url	URL zur Seite
html	Gesamtes HTML
title	Title der Seite
article	Text zwischen <article> - Tags
names_from_left_panel	Alle Namen im linken Panel

### Kontakt



**Prof. Dr. Martin Bonnet**

Raum HO 2-34

☎ +49 221-8275-2649

✉ martin.bonnet@th-koeln.de

» Zur Personenseite



**Prof. Dr. Hans Willi Langenbahn**

Raum HO-02-31

☎ +49 221-8275-2699

✉ hans\_willi.langenbahn@th-koeln.de

» Zur Personenseite



**Prof. Dr. Michael Josef Böhmer**

Raum HO-02-35

☎ +49 221-8275-2164

✉ michael.boehmer@th-koeln.de

» Zur Personenseite

**Prof. Dr. Stefan Benke**

Raum HO 2-25

☎ +49 221-8275-2087

✉ stefan.benke@th-koeln.de

» Zur Personenseite

## 3.2 Interne Quellen - Datensatz: Mapping

document_id	Eindeutige Dokument ID
person_id	Eindeutige Personen ID
found_in_article	True, wenn Name in Text
found_in_left_panel	True, wenn Name im linken Panel



## 3.2 Interne Quellen - Ansatz

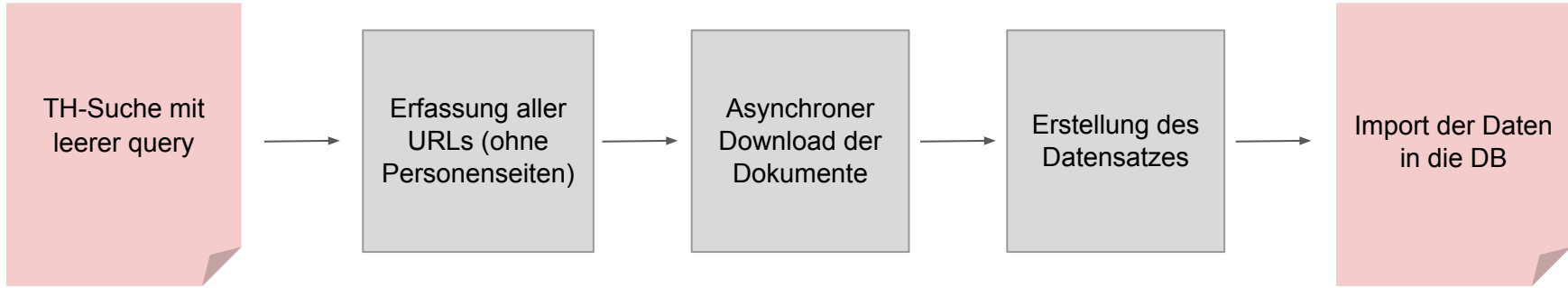
### **Ansatz: TH-Suche**

- + relevante Dokumente sind gebündelt
- + Fast alle Dokumente haben Inhalt
- + braucht nur wenige Minuten
- + ermöglicht Spezifikation des Dokumentformates

#### **Dateiformat**

- ☐ Webseite (10101)
- ☐ PDF (5599)
- ☐ Bildergalerie (232)
- ☐ Word (143)
- ☐ Video (82)

## 3.2 Interne Quellen - Prozess



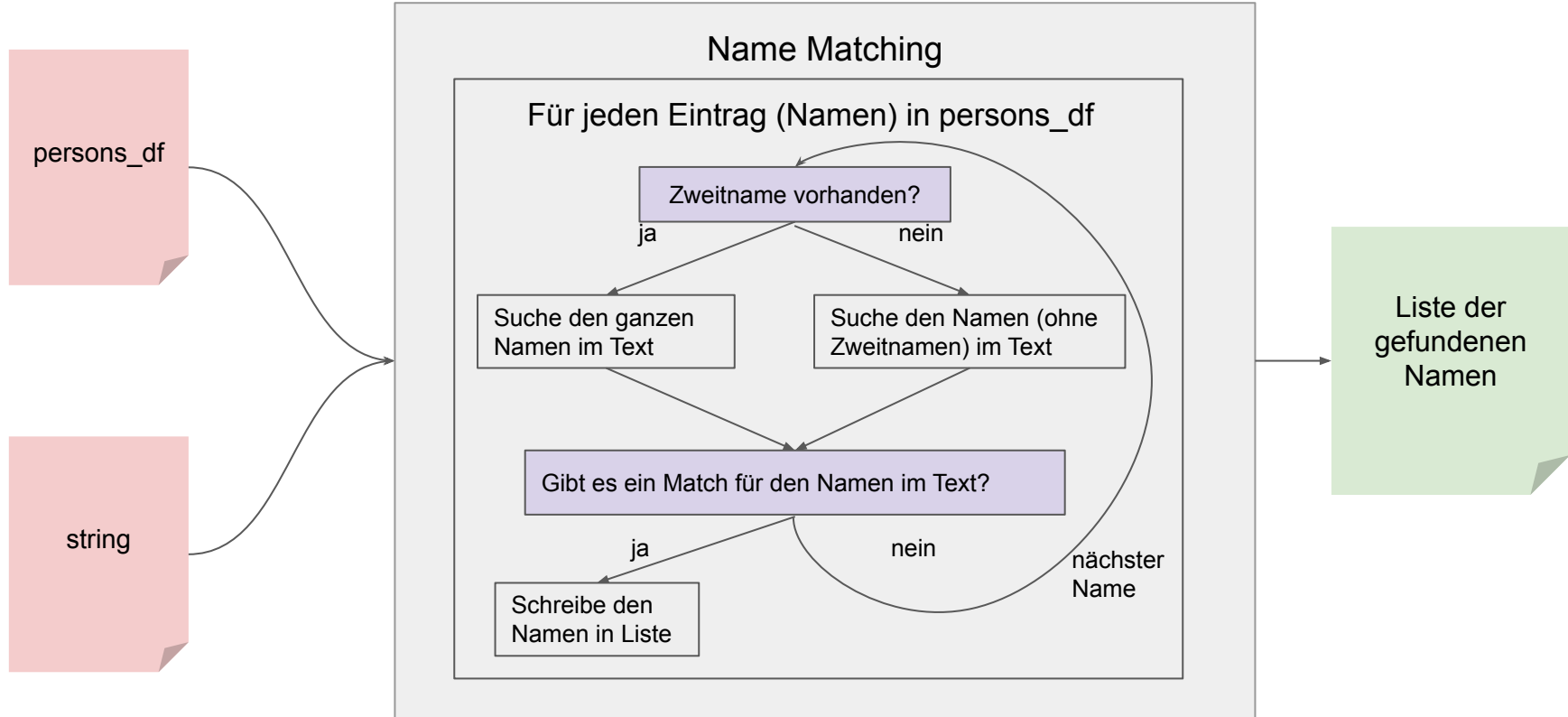
Websites, Bildergalerien, PDFs (*nächste Seite*)

[th-koeln.de/suche/index.php?query=&file\\_type\\_de\[\]=Webseite&page=0](http://th-koeln.de/suche/index.php?query=&file_type_de[]=Webseite&page=0)

Projekte (*mehr Ergebnisse*)

[th-koeln.de/filter\\_list\\_more.php?document\\_type\[\]=Forschungsprojekt&target=%2Fforschung%2Faktuelle-projekte\\_2418.php&resultlayout=standard&sortfield=publish\\_date&sortorder=desc&language=de&keywords=&start=0](http://th-koeln.de/filter_list_more.php?document_type[]=Forschungsprojekt&target=%2Fforschung%2Faktuelle-projekte_2418.php&resultlayout=standard&sortfield=publish_date&sortorder=desc&language=de&keywords=&start=0)

## 3.3 Name Matching



## 3.3 Name Matching

```
names_matched = []
for i, r in persons_df.iterrows(): # iterate through all names
    if type(r.m_name) == list: # check if name includes middle names
        pattern = r'(' + str(r.titles) + ')?((' + str(r.f_name.replace('-', '\\-')) + ')|(' + str(
            r.f_name[0]) + '\\.? ))((' + str(' '.join(r.m_name)).replace('-', '\\-') + ')|(' + str(
            ' '.join([str(name[0] + '\\.?') for name in r.m_name])) + ' ))(' + str(r.l_name) + ')'
    else:
        pattern = r'(' + str(r.titles) + ')?((' + str(r.f_name.replace('-', '\\-')) + ')|(' + str(
            r.f_name[0]) + '\\.? ))(' + str(r.l_name) + ')'
    match = re.search(pattern, text) # perform the matching
    if match != None:
        names_matched.append(r.person_id) # add the person_id of matched name to the result list
return names_matched # return all person_ids of the names matched in this text
```

Titel (*optional*)

Vorname ganz oder nur Initiale

Zweitname(n) ganz oder nur Initiale(n) (*optional*)

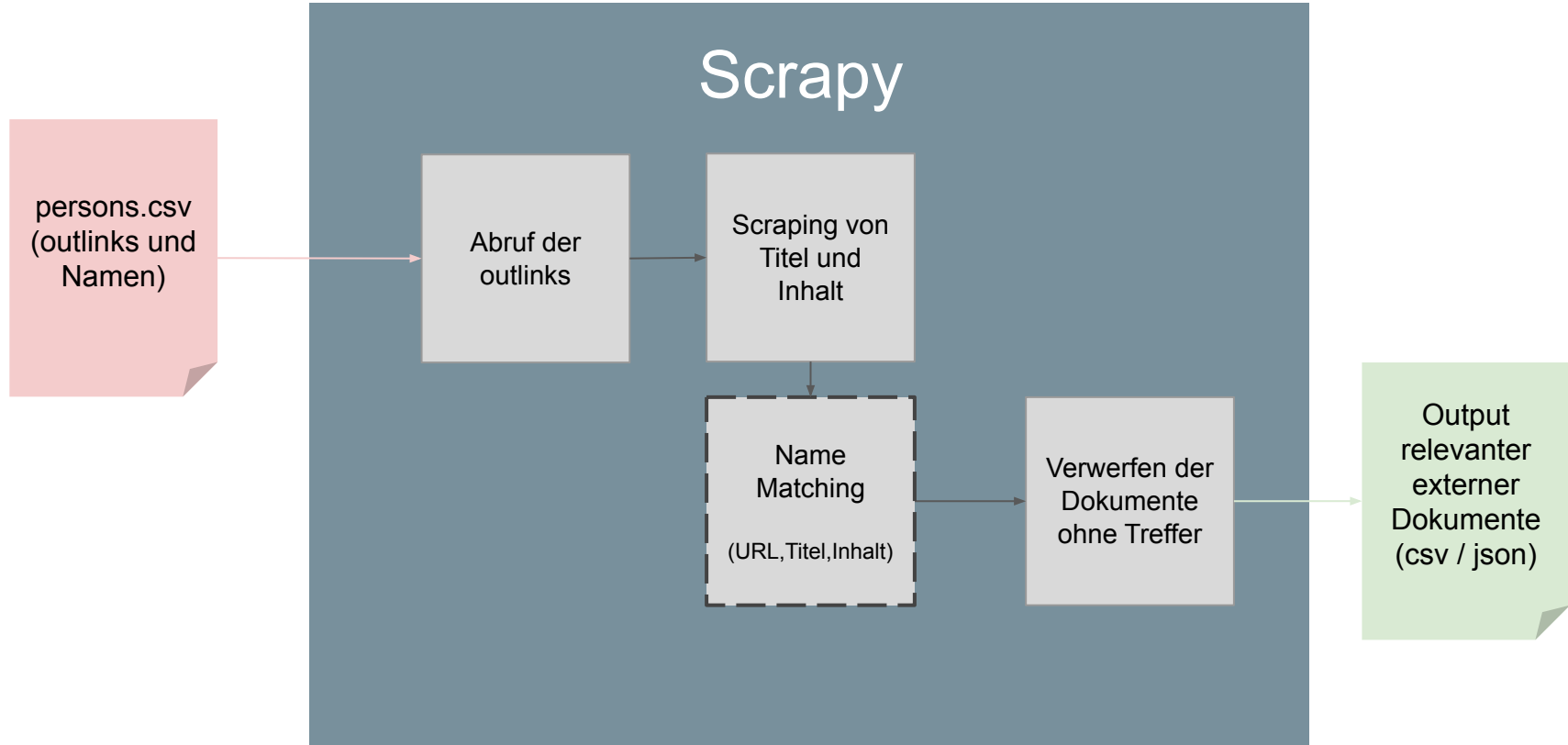
Nachname

Titel (*optional*)

Vorname ganz oder nur Initiale

Nachname

## 3.4 Externe Quellen (verworfen)

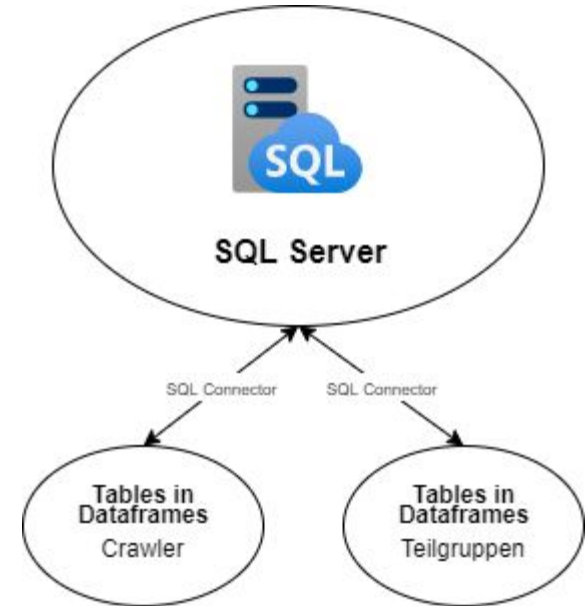


# 04 —

## Vorverarbeitung des Datensatzes

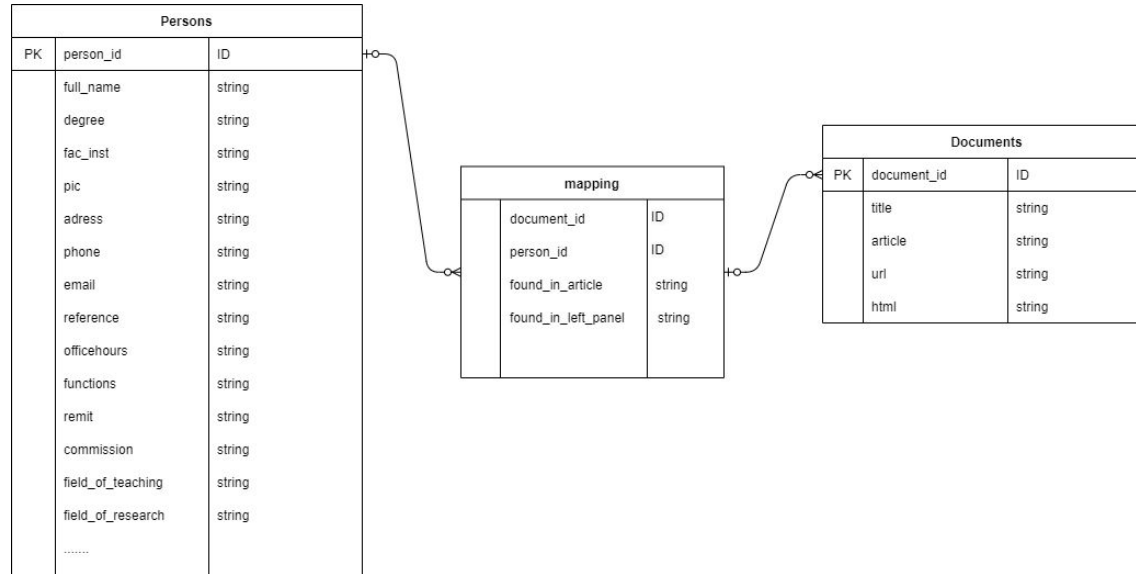
## 4.1 Datenbank

- webgehostete MariaDB
- Datenzugriff direkt via SQL Connector in MariaDB
- zunächst private MariaDB
- später Daten zur TH gehostete MariaDB migriert



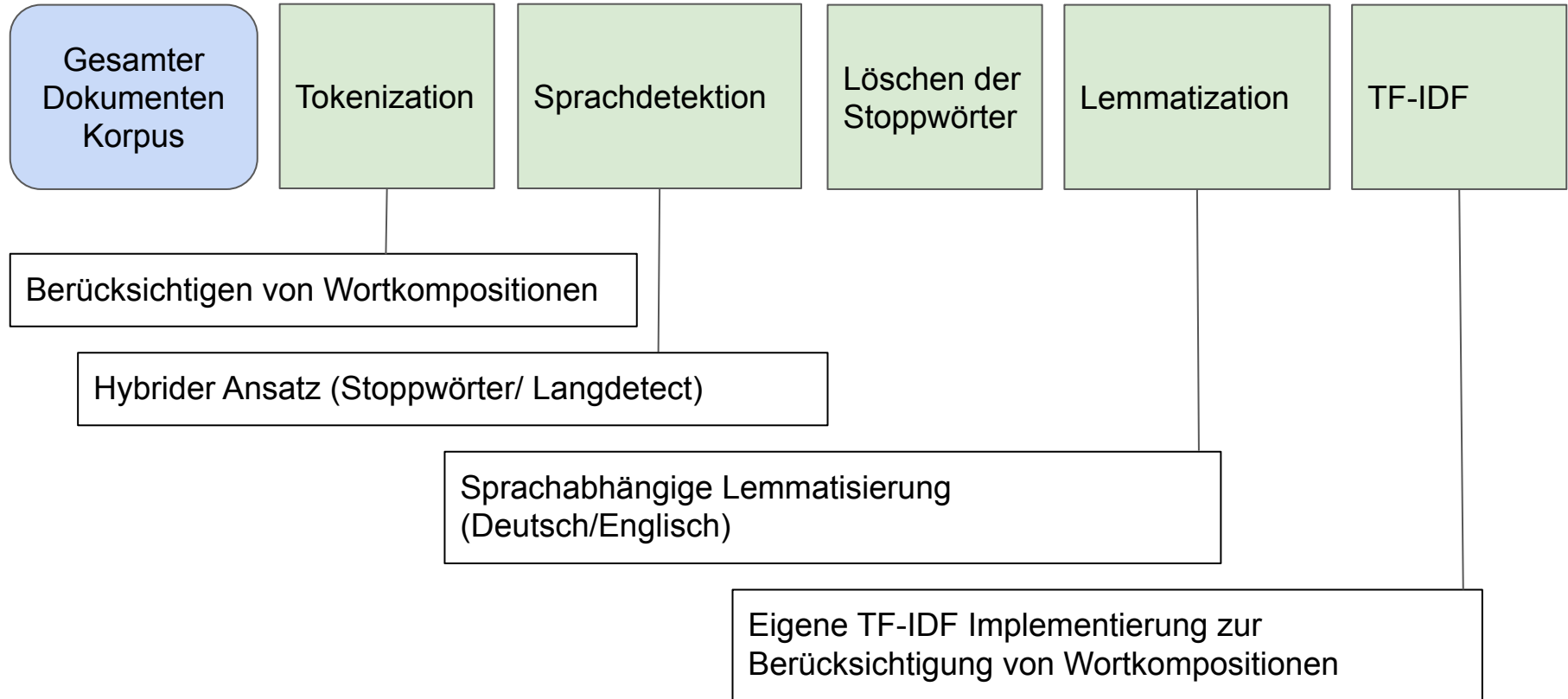
## 4.1 Datenbank

### Entwicklung des Datenbankschemas



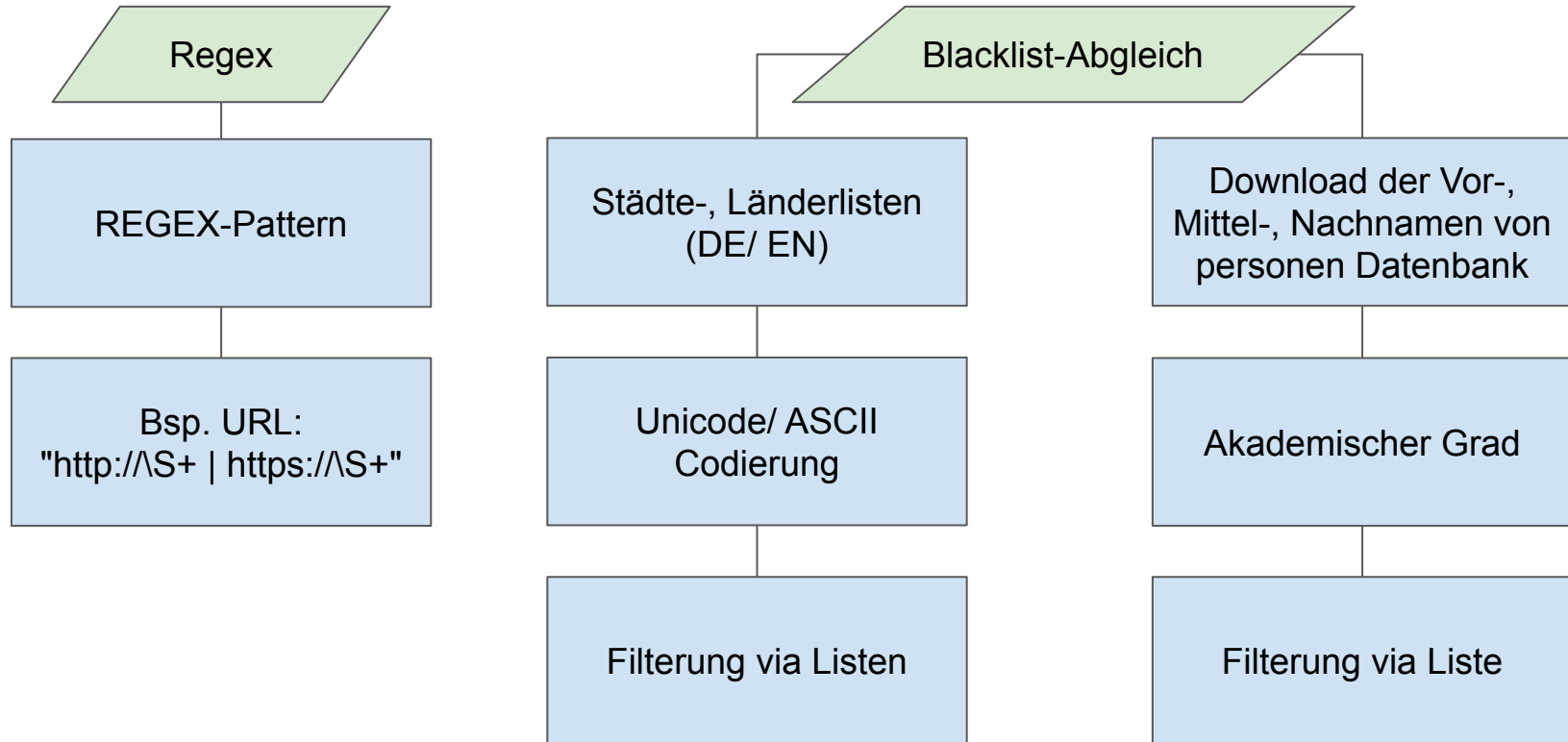


## 4.2 Pre Processing Pipeline



## 4.2 Pre Processing Pipeline

Identifizierung möglicher seltener Terme (Namen, Locations)

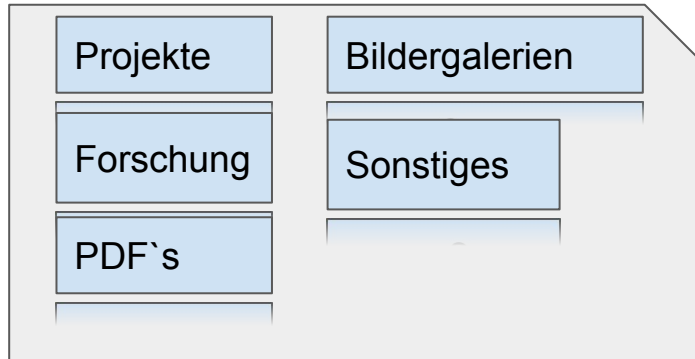


## 4.2.1 Designentscheidung

Klassenstruktur:

- separat verwendbare Funktionen
- Pipeline Prozess individuell anpassbar
- Nutzung von Vererbung
- Nutzung von Globalen Variablen

Kompletter Korpus



Selektierter Korpus



## 4.2.2 Problematik Implementierung

- |                 |  |
|-----------------|--|
| <b>Problem:</b> | Vorverarbeitung nur auf den aktuellsten Korpus optimierbar.                |
| <b>Lösung:</b>  | Stetige Evaluierung der Anforderungen für das Preprocessing                |
| <b>Problem:</b> | Die eigene Implementierung von TF-IDF-Berechnung nicht Laufzeit optimiert. |
| <b>Lösung:</b>  | Implementierung via Pandarallel.   |
| <b>Problem:</b> | Limitierung von Zeichenmenge in der SQL Datenbank                          |
| <b>Lösung:</b>  | ...  |
| <b>Problem:</b> | Filterung der Dokumenten anhand semantischer Kategorienlisten              |
| <b>Lösung:</b>  | Synsemantische Netze (Wordnet, Germanet), Embeddings, Wortvektoren, Listen |
| <b>Problem:</b> | Blacklist-Ansatz bei der Filterung   |
| <b>Lösung:</b>  | Whitelist-Ansatz verwenden   |

## 4.2.2 Problematik Implementierung

Bsp. Problematik Blacklist vs Whitelist Ansatz

### Blacklist-Ansatz:

Entferne alle Zeichen innerhalb der Unicode Ranges

**Kyrillisch:** \u0400–\u04FF

**Arabisch:** \u0600–\u06FF

**Asiatische Schriftzeichen:**

\u4e00–\u9FFF

\uac00–\ud7a3

\u3040–\u30ff

### Whitelist-Ansatz:

Entferne alle Zeichen außerhalb der gegebenen Unicode Range

**Basic-Latin:** u\u0000–u\u007F

## 4.2.2 Problematik Implementierung

Bsp. Problematik Bindestriche

Bindestriche sollten erhalten bleiben um fachbezogene Wortkompositionen als Tokens zu behalten:

“IT-Anwendungen in Bibliotheken”

“EDV-Anwendungen im Bauwesen”

“Entwicklung von PHP-Anwendungen”

Welche Probleme sich durch die PDF's ergaben:

qua-litätskriterien

'insbesonde-'

'ei-ner'

## 4.2.2 Problematik Implementierung

Bsp. Problematik Fehlende Whitespaces bei den PDF`s

Teilweise fehlende Whitespaces in den PDF`s:

dasseinmaterialverlustsch

allerdingskanderherkömmlicheauftragvonfestigungsmittelnmitpinselneinrisikofürdasobjekt

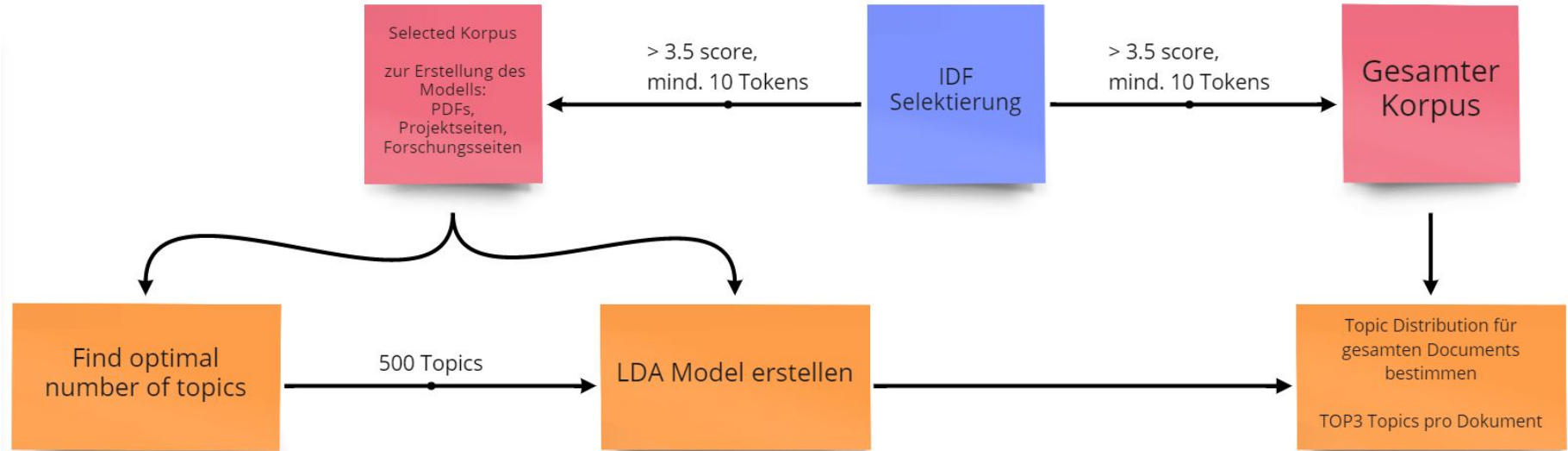


# 05 —

Extraktion der Topics aus den  
Dokumenten



## 5.1 LDA-Pipeline



# 5.2 Korpusauswahl

▼ 50 Jahre TH Köln

## Für Studierende der TH Köln

Studierende können das Postkartenset "Ansichtssache, Teil I" kostenfrei bestellen. Schreiben Sie mit Ihrer email-Adresse an 50jahre@th-koeln.de

## Kontakt

### Jubiläumskoordinator



**Matthias Kötter**  
Hochschulreferat Kommunikation  
und Marketing

☎ +49 221-8275-5289

✉ 50jahre@th-koeln.de

## Ansichtssache! Postkarten zum 50-jährigen Jubiläum

Die TH Köln ist bunt, groß, vielfältig. Man kann sie aus vielen Perspektiven betrachten. Wir zeigen Ihnen auf unseren sechs Jubiläums-Postkarten einen besonderen Blick auf die Hochschule. Außerdem rufen wir zu einem Kreativwettbewerb auf: Wie sieht Ihre TH-Postkarte aus? Reichen Sie Ihre Idee ein – die besten Vorschläge werden umgesetzt.

### Bildergalerie



7 / 7

Die TH Köln lehrt und forscht im Bergischen Land: Am Campus Gummersbach stehen Informatik und Ingenieurwissenschaften auf dem Plan, in Lindlar geht es um nachhaltiges Wirtschaften mit Abfällen. (Bild: Bergischer Abfallwirtschaftsverband (BAV), Thilo Schmülgen/TH Köln)

- ◀ Hochschule
- ▼ Aktuelles
- ▶ Nachrichten

Termine  
Standortentwicklung  
Presse und Kommunikation  
Amtliche Mitteilungen

» Die Hochschulbibliothek

» Standorte & Öffnungszeiten

» Fragen Sie uns (Kontaktformular)

## Buchbare Einzelarbeitsplätze



Das Lernen vor Ort ist bald wieder möglich: Ab Montag, 21.06.21 können Angehörige der TH Köln ausgewählte Lernplätze nutzen. Bitte buchen Sie dafür ab Mittwoch, 16.06.21 online in ILIAS einen Einzelarbeitsplatz.

### Buchungsregeln

- Nur Angehörige der TH Köln können buchen
- Ausschließlich online über ILIAS kann gebucht werden
- Bis zu 4 Stunden vor Öffnung am jeweiligen Tag möglich
- Jede NutzerIn kann pro Tag einen eigenen Lernplatz buchen. Mehrfachbuchungen werden storniert.
- In der persönlichen Buchungsübersicht in ILIAS können Sie Ihre Reservierungen stornieren
- Die Buchung gilt für einen gesamten » Öffnungstag
- Jeweils die nächsten 5 Öffnungstage sind zur Buchung freigeschaltet
- Alle anderen Gruppen- und PC-Arbeitsplätze bleiben leider weiterhin gesperrt
- Ohne vorherige Buchung können die Einzelarbeitsplätze in der Bibliothek nicht genutzt werden
- Ein Negativtest muss nicht vorgelegt werden
- Achtung: Die Buchung per Smartphone ist nicht möglich

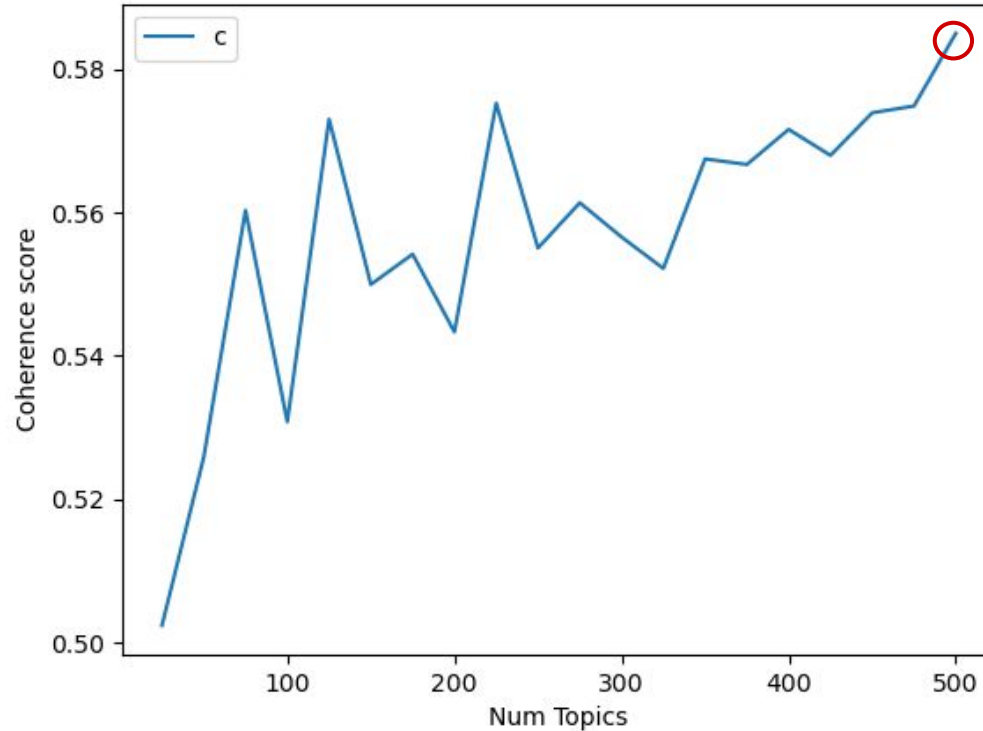
**Fair Play:** Bitte stornieren Sie rechtzeitig, damit andere NutzerInnen eine Chance auf einen Lernplatz haben. Falls Sie mehrfach die Buchung ungenutzt verfallen lassen, können Sie gesperrt werden.

→ Nicht fachspezifisch

## 5.2 Korpusauswahl

- Für die Erstellung fachspezifischer Topics benötigt es selektierten Korpus:
  - PDFs
  - Forschungsseiten
  - Projekte
- Ausschlusskriterien:
  - mindestens 10 Tokens pro Dokument
  - mindestens 3.5er IDF Score pro Term

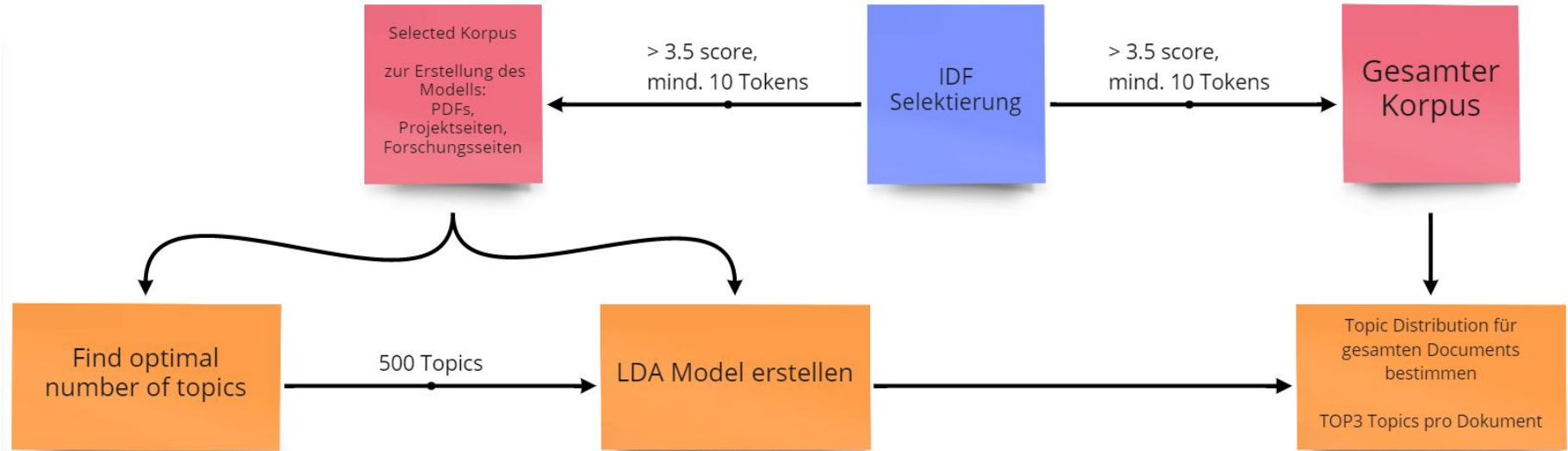
## 5.3 Optimale Anzahl der Topics bestimmen



Höchster Coherence Score  
bei Num Topics = 500

Optimale Topicanzahl = 500

## 5.4 LDA-Pipeline



## 5.5 Topicverteilung (500 Topics)

Sinnvolle Topics:

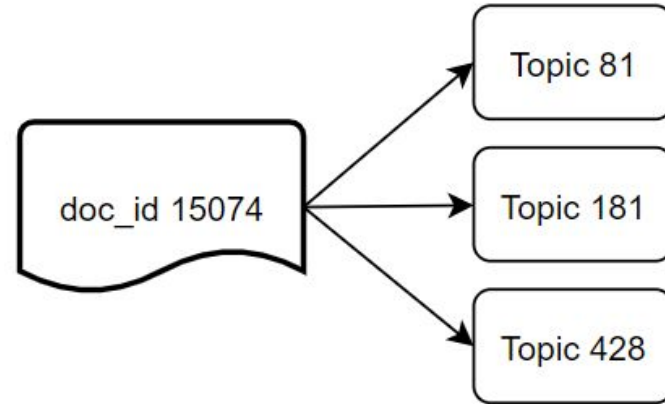
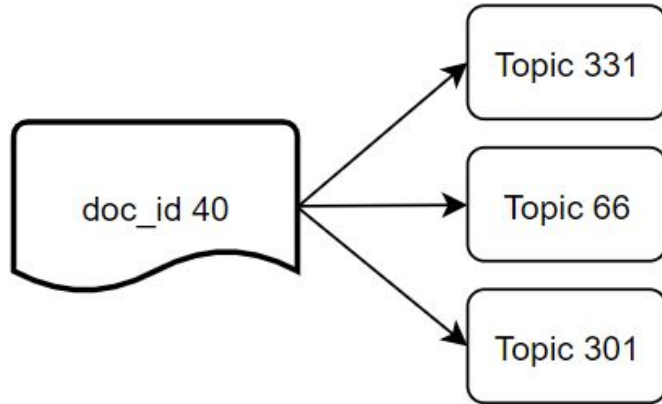
Topic ID	Terme
322	mediumen, weißabgleich, lichtfarbe, photokina, mischlicht, bildaufnahme, ungenauen, camera, kamera, lter
326	cloud, schema, sensorcloud, db, fdb, dbms, datenbanksysteme, Ä, fördert, hersteller

Nicht sinnvolle Topics:

Topic ID	Terme
94	mediennutzung, lv, #, israelisch, societies, rückversicherung, turnus, hochschuldidaktik, whatsapp, nadev
428	forschungsstrategie, stärke, forum, vision, forschungsprofil, forschungsförderung, funk, weiter-, technologietransfer, organisationsentwicklung

## 5.6 Topic-Dokument-Relation

- max. Top 3 Topics pro Dokument



## 5.6.1

# Topic-Dokument-Relation

### Dokument Nr. 40

Forschungsschwerpunkt  
Sozial • Raum • Management

- ◀ Forschungsstrukturen
- ▼ Sozial Raum Management (FSP)
  - Projekte
  - Publikationen

#### Forschungsschwerpunkt

Sozial • Raum • Management  
Fakultät für Angewandte  
Sozialwissenschaften

### Öffnung des Wohnquartiers für das Alter

**Entwicklung einer kommunikativen Informationsinfrastruktur zur Überbrückung struktureller Lücken im Sozialraum**

Herbert Schubert

Sigrid Leitner

Katja Veil

Marina Vukoman

Verlag Sozial • Raum • Management, Band 13

Köln 2014

#### Zum Inhalt

Im Blickpunkt stehen ältere Menschen, die in ihrer privaten Lebensführung zurückgezogen leben, wenig in lokale Beziehungsnetzwerke involviert sind und die von Informationen und Angeboten der Altenhilfeträger bisher nicht erreicht werden. Für diese Menschen wurde die Idee einer „kommunikativen Informationsinfrastruktur“ im Sozialraum des Wohnviertels und Stadtteils entwickelt. Mit dem Infrastrukturmodell soll vermieden werden, dass solche Personen unerkannt in Notsituationen geraten, aber auch sichergestellt werden, dass sie kontinuierlich über Gelegenheiten zur erfolgreichen Bewältigung ihrer Lebenssituation

#### Bestellung

- Printausgabe bei Amazon
- Download als Open Access Ebook (PDF)





Öffnung des Wohnquartiers für das Alter (Bild: FH Köln/SRM)



## 5.6.1 Topic-Dokument-Relation

Dokument Nr. 40

Topic 1	Topic 2	Topic 3
Topic ID 301: vermittler, öffna, sozialraum, seniorenberatung, wohnquartier, öffnung, beratungsstelle, beratungsangebot, silqua, kontaktaufnahme 	Topic ID 66: beratungsbedarf, unerkant, informationsangebot, überreichen, überbrückung, wohnquartier, notsituation, projektlaufzeit, friseure, ehrenfeld 	Topic ID 331: Mehrfach, sterberaten, unsicherheit, mirko, letztlich, stromnetz, wei-, nahe, koordinieren, ordnen 

## 5.6.2

## Topic-Dokument-Relation

## Dokument Nr. 15074

- ◄ Institut für Produktentwicklung und Konstruktionstechnik
- ▼ Forschung
  - Projekte

» Labor für Fertigungssysteme  
» F&E Projekte

## Kontakt

Prof. Dr. Ulf Müller

☎ +49 221-8275-2914  
✉ ulf.mueller@th-koeln.de

» Zur Personenseite

## KI: Mobil

## Künstliche Intelligenz in virtueller Realität erleben

- Gewinner Hochschulwettbewerb im Wissenschaftsjahr 2019 - Künstliche Intelligenz
- Interaktive Darstellung der Funktion künstlicher, neuronaler Netze
- Mobiles Präsentationskonzept zur proaktiven Kommunikation mit der Öffentlichkeit



KI: Mobil - Künstliche Intelligenz in virtueller Realität erleben  
(Bild: TH-Köln, IPK, Ifk)




## Kurzfassung der Projektbeschreibung:

Das Projekt KI:Mobil wurde ins Leben gerufen, um die Thematik der künstlichen Intelligenz einer breiten Öffentlichkeit zugänglich zu machen und einfach zu erklären. Ziel ist es die zentralen Fragestellungen „Was ist KI?, Wie funktioniert sie? und Wie kann diese eingesetzt werden?“ zu beantworten. Dazu wurde eine mobile Lern- und Spiele **nach oben** entwickelt, die es ermöglicht, das Thema KI an jedem beliebigen Ort zu vermitteln und zu erleben. Der proaktive Ansatz ermöglicht es Verständnis und Akzeptanz für diese Technologie zu schaffen. Um dies zu erreichen, wurde ein interaktives Virtual Reality Spiel entwickelt, in dem kleine vorformulierte Aufgaben, mit Hilfe gängiger Modelle der künstlichen Intelligenzen, gelöst werden müssen. Die Plattform KI:Mobil bietet die Möglichkeit die zukunftsweisende KI-Technologie vielen interessierten Unternehmen und Einrichtungen zugänglich zu machen. Neben der (Weiter-) Bildungsmöglichkeit für bspw. Schulen und Ausbildungsstätten, sind zudem Klein- und Mittelständigen Unternehmen eine der Hauptzielgruppen. Künftig soll das KI:Mobil darüber hinaus als Wissensvermittlungsplattform für neue KI-Fragestellungen aus Bildung, Gesellschaft und Wirtschaft verwendet werden.

[https://www.th-koeln.de/anlagen-energie-und-maschinensysteme/ki-mobil\\_74226.php](https://www.th-koeln.de/anlagen-energie-und-maschinensysteme/ki-mobil_74226.php)

## 5.6.2 Topic-Dokument-Relation

Dokument Nr. 15074

Topic 1	Topic 2	Topic 3
Topic ID 81: künstliche intelligenz, intelligenz, künstlich, big, algorithmen, maschinell, handschrift, fahre, telematik, rezept 	Topic ID 181: duroplasten, prozessparameter, projektbeschreibung, präsenzveranstaltung, studierendenzentriert, ipk, lfk, kurzfassung, materialeigenschaft, einsparung 	Topic ID 428: forschungsstrategie, stärkung, forum-, vision, forschungsprofil, forschungsförderung, funk, weiter-, technologietransfer, organisationsentwicklung 

## 5.7 Herausforderungen, Probleme & Zwischenergebnisse

**Problem:** Laufzeit der Modellierung

**Lösung:** Nutzung einer VM der TH-Köln, Verwendung von LDA Multicoremodel

**Problem:** optimale Anzahl an Topics finden

**Lösung:** Topic Coherence

**Problem:** Orte, Namen, typische Hochschulsbegriffe, Sonderzeichen

**Lösung:** Verbessertes Preprocessing

**Problem:** fachspezifische Topicerstellung

**Mögliche Lösung:** Spezialisierung des Korpus für die Topicerstellung (Projekte, Forschung, PDF)

**Problem:** mehrere Fachgebiete in einem Topic

**Mögliche Lösung:** Eindeutigerer Korpus; Verbessertes Preprocessing; Händische Topicauswahl

# 06 —

## Intellektuelle Erschließung der Topics

## 6. Automatische Erschließung

- Erschließung der Lehrgebiete über TH-Köln interne Struktur
  - Fakultäten → Institute → Lehrgebiete

Kandidaten sind an  
Fakultäten und  
Instituten beschäftigt.

```
"Prof. Dr. Jürgen Danielzik":{  
  "faculty":"Fakultät für Bauingenieurwesen und Umwelttechnik",  
  "institute":"Institut für Baubetrieb und Vermessung (IBV)",  
  "fot":["Baubetrieb, Bauwirtschaft und Baumanagement, Bauprojektmanagement",  
        "Nachtragsforderungen aus gestörten Bauabläufen, BIM"]  
},  
"Prof. Dr. Maik Dapper":{  
  "faculty":"Fakultät für Anlagen, Energie- und Maschinensysteme",  
  "institute":"Institut für Technische Gebäudeausrüstung (TGA)",  
  "fot":["Anlagenhydraulik, Kreiselpumpen, hydraulischer Abgleich, hydraulische Schaltungen",  
        "Gastechnik in der Hausinstallation, Erdgas, Flüssiggas"]  
},  
"Prof. Hannelore Damm":{  
  "faculty":"Fakultät für Bauingenieurwesen und Umwelttechnik",  
  "institute":"Institut für Konstruktiven Ingenieurbau (IKI)",  
  "fot":["Holzbau, Baustatik, Ingenieurholzbau und EDV-Anwendungen im Bauwesen"]  
},  
"Dr. Cornelia Dahmer":{  
  "faculty":"Fakultät für Informations- und Kommunikationswissenschaften",  
  "institute":"Institut für Translation und Mehrsprachige Kommunikation (ITMK)",  
  "fot":["Angewandte Deutsche Sprach- und Kulturwissenschaft"]  
},  
"Alban de Lausun":{  
  "faculty":"Fakultät für Informations- und Kommunikationswissenschaften",  
  "institute":"Institut für Translation und Mehrsprachige Kommunikation (ITMK)",  
  "fot":["französische Wirtschaft, französische Recht, französisch Politikwissenschaften"]  
}
```



vorverarbeiteter  
Input

## 6. Automatische Erschließung

- Über Kandidaten eine Zugehörigkeit von Instituten zu Fakultäten herstellen
- Zwischenergebnis:
  - 12 Fakultäten
  - 53 Institute
- Nächster Schritt:  
Ermitteln von Lehrgebieten an Instituten

```
"Fakultät für Anlagen, Energie- und Maschinensysteme":{  
  "Institut für Rettungsingenieurwesen und Gefahrenabwehr (IRG)":[ ... ],  
  "Institut für Werkstoffanwendung (IWA)":[ ... ],  
  "Institut für Produktentwicklung und Konstruktionstechnik (IPK)":[ ... ],  
  "Cologne Institute for Renewable Energy (CIRe)":[ ... ],  
  "Institut für Bau- und Landmaschinentechnik Köln (IBL)":[ ... ],  
  "Institut für Technische Gebäudeausrüstung (TGA)":[ ... ],  
  "Institut für Anlagen- und Verfahrenstechnik (IAV)":[ ... ],  
  "Institut für Elektrische Energietechnik (IET)":[ ... ]  
},  
"Fakultät für Bauingenieurwesen und Umwelttechnik":{  
  "Institut für Baustoffe, Geotechnik, Verkehr und Wasser (IBGVW)":[ ... ],  
  "Institut für Konstruktiven Ingenieurbau (IKI)":[ ... ],  
  "Institut für Baubetrieb und Vermessung (IBV)":[ ... ]  
},
```

## 6. Automatische Erschließung

- Ergebnis sind 3 Hierarchisierungsstufen

1. Fakultät
2. Institut
3. Liste der Lehrgebiete

```
"Fakultät für Angewandte Naturwissenschaften":{
  "Institut für Anlagen- und Verfahrenstechnik (IAV)":[
    "Thermodynamik", "Thermische Verfahrenstechnik", "Membranprozesse", "Wasseraufbereitung",
    "Anlagenbau und Energieverfahrenstechnik", "Project Management", "Project Controlling",
    "Grundlagen der Verfahrenstechnik", "Wärme- und Stofftransport", "Prozesssimulation",
    "Stationäre Simulation (CHEMCAD) und Dynamische Simulation (Matlab)", "Reaktionstechnik",
    "Heterogene Reaktionssysteme und Polymerreaktionstechnik", "Maßstabsvergrößerung",
    "Vorplanung und Basisplanung", "Cost Engineering", "Kostenschätzung in frühen
    Projektstadien", "Produkt- und Prozessentwicklung", "Verfahrenstechnische Produkte",
    "Fluidverfahrenstechnik", "Membrantechnik", "Abwassertechnik",
    "Partikeltechnologie", "Feststoffverfahrenstechnik"
  ]
},
```

- Anwendung in der Web UI



## 6.2 Intellektuelle Erschließung

- Verknüpfung der Topics mit den Dokumenten

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	topic_id	topic_terms																								
2	0	(0, '0.070***kampagne' + 0.036***-innen' + 0.025***ö' + 0.024***d' + 0.020***spielgemeinschaft' + 0.019***spielraum' + 0.016***spieler' + 0.014***szene' + 0.014***äden' + 0.012***medienkompetenz'')																								
3	1	(1, '0.003***stellungsmessung' + 0.003***evaluationsergebniss' + 0.003***tessmer' + 0.003***rechnerarchitektur' + 0.001***gebögen' + 0.000***mentoring' + 0.000***beginners' + 0.000***©' + 0.000***evaluationsergebnisse' + 0.000***credit'')																								
4	2	(2, '0.011***investor' + 0.011***collateralized' + 0.010***capacity' + 0.009***rückversicherung' + 0.008***einhellig' + 0.007***rückversi' + 0.006***meisch' + 0.006***rückversicherer' + 0.005***rückver' + 0.005***sinken'')																								
5	3	(3, '0.015***informationsverarbeitung' + 0.013***hochwasser' + 0.013***clarke' + 0.012***ions' + 0.011***ars' + 0.010***sv' + 0.009***solutions' + 0.009***schulungs-' + 0.009***einsatzkräfte' + 0.009***aufenthaltort'')																								
6	4	(4, '0.026***step' + 0.026***metabolon' + 0.014***efre' + 0.013***forschungszentrum' + 0.011***sustainable' + 0.011***polyurethane' + 0.011***erneuerbare energien' + 0.010***reststoff' + 0.010***energie technik' + 0.009***renewable'')																								
7	5	(5, '0.000***i' + 0.000***aufsuchen' + 0.000***std' + 0.000***credit' + 0.000***x' + 0.000***nationalität' + 0.000***angeführt' + 0.000***öffna' + 0.000***x' + 0.000***silqua'')																								
8	6	(6, '0.021***maler' + 0.018***cranach' + 0.015***rdert' + 0.012***meisterwerk' + 0.012***gemälde' + 0.011***royal' + 0.010***kunstpalast' + 0.010***sohn' + 0.010***andrew' + 0.010***kooperatio'')																								
9	7	(7, '0.049***#' + 0.026***mediennutzung' + 0.025***-befragung' + 0.024***praxisstudium' + 0.020***nennung' + 0.017***praxiszentrum' + 0.010***innengruppe' + 0.008***möglichkeit' + 0.008***x' + 0.007***messenger'')																								
10	8	(8, '0.000***ects' + 0.000***\$' + 0.000***credit' + 0.000***lv' + 0.000***kulturwissenschaft' + 0.000***#' + 0.000***methods' + 0.000***sws' + 0.000***stadtteil' + 0.000***rückversicherung'')																								
11	9	(9, '0.027***risikomanagement' + 0.010***vage' + 0.010***berichterstattung' + 0.009***aggregation' + 0.008***faris' + 0.008***-und' + 0.008***bewertungsmethode' + 0.008***rm' + 0.007***beziehungsweise' + 0.006***risikomanagementsystem'')																								
12	10	(10, '0.019***gefahrenabwehr' + 0.014***rettungsingenieurwesen' + 0.013***irg' + 0.013***routing' + 0.012***esri' + 0.011***geography' + 0.008***masterstudent' + 0.008***marz' + 0.008***rohr' + 0.007***oraussetzung'')																								
13	11	(11, '0.000***lv' + 0.000***ects' + 0.000***obligatory' + 0.000***dolmetsch' + 0.000***sws' + 0.000***stauraum' + 0.000***konferenzdolmetsch' + 0.000***leistungspunkt' + 0.000***summer' + 0.000***lotsen'')																								
14	12	(12, '0.020***jah-' + 0.008***wikis' + 0.008***fraktion' + 0.006***wunder' + 0.006***referentenentwurf' + 0.006***masterstudiums' + 0.004***verabschiedet' + 0.004***wissensmanagement' + 0.004***stel-' + 0.004***landtag'')																								
15	13	(13, '0.025***zertifikatskurs' + 0.018***selbstlernphase' + 0.018***zbwi' + 0.016***können' + 0.015***xprtn' + 0.010***live' + 0.009***arbeitsaufwand' + 0.007***zzgl' + 0.007***zeitstunde' + 0.007***gudrun'')																								
16	14	(14, '0.012***studierendenwerk' + 0.000***zutreffend' + 0.000***teilprojekt' + 0.000***zufrieden' + 0.000***top' + 0.000***grenzüberschreitend' + 0.000***selbstverständnis' + 0.000***entwicklungspartnerschaft' + 0.000***unentschieden' + 0.000***lernend'')																								
17	15	(15, '0.000***-innen' + 0.000***credit' + 0.000***i' + 0.000***isbn' + 0.000***workload' + 0.000***öffnung' + 0.000***turnus' + 0.000***geflichtet' + 0.000***prüfungsform' + 0.000***sws'')																								
18	16	(16, '0.030***session' + 0.028***kompetenzorientierung' + 0.020***kompetenzentwicklung' + 0.019***kompetenzorientiert' + 0.016***zung' + 0.013***kompe-' + 0.011***entschei-' + 0.011***fö-' + 0.011***-tisch' + 0.011***kompe'')																								
19	17	(17, '0.000***baoinformatik' + 0.000***credit' + 0.000***-beratung' + 0.000***zutreffend' + 0.000***steps' + 0.000***d' + 0.000***leistungsbefugnis' + 0.000***stadtteil' + 0.000***programmierung' + 0.000***prüfungsform'')																								
20	18	(18, '0.024***cold' + 0.020***mashing' + 0.015***food' + 0.015***extract' + 0.015***c' + 0.011***kit' + 0.009***-dem' + 0.009***composition' + 0.009***fig' + 0.009***malt'')																								
21	19	(19, '0.015***current' + 0.013***power' + 0.009***concern' + 0.008***core' + 0.007***dc' + 0.007***inductor' + 0.007***ac' + 0.007***without' + 0.007***flux' + 0.006***pa'')																								
22	20	(20, '0.033***exp' + 0.020***pensionsfond' + 0.018***stochastisch' + 0.015***rente' + 0.015***gl' + 0.014***asset' + 0.012***variante' + 0.011***liability' + 0.011***rentenanpassung' + 0.010***sterblichkeit'')																								
23	21	(21, '0.000***credit' + 0.000***dual' + 0.000***lernergebnisse' + 0.000***turnus' + 0.000***prüfungsform' + 0.000***bestandene' + 0.000***lehrformen' + 0.000***kontaktzeit' + 0.000***modulbeauftragte' + 0.000***wahlpflichtfach'')																								
24	22	(22, '0.071***reise' + 0.017***quo' + 0.016***inklusion' + 0.015***abb' + 0.015***jugendreise' + 0.014***inklusion' + 0.013***halt' + 0.011***freizeit' + 0.009***äh' + 0.009***behin'')																								
25	23	(23, '0.009***doi' + 0.009***chemical' + 0.009***kurzfassung' + 0.008***humanities' + 0.007***archivierung' + 0.006***heal' + 0.006***diplomtag' + 0.005***rel' + 0.005***diplomandinnen' + 0.005***diplomanden'')																								
26	24	(24, '0.000***credit' + 0.000***psso' + 0.000***studiensemester' + 0.000***sws' + 0.000***modulnummer' + 0.000***workload' + 0.000***kontaktzeit' + 0.000***prüfungsform' + 0.000***pflichtfach' + 0.000***endnote'')																								
27	25	(25, '0.016***glas' + 0.013***pin' + 0.009***ingenieurnachwuchs' + 0.008***facette' + 0.008***brillenfassung' + 0.006***gst' + 0.006***brillenglas' + 0.005***nut' + 0.005***rand' + 0.005***pöpperl'')																								
28	26	(26, '0.020***ausfall' + 0.016***künstlich' + 0.015***krankenhaus' + 0.014***gesundheits-' + 0.013***aufrechterhaltung' + 0.012***künstliche intelligenz' + 0.012***cities' + 0.011***gesundheitswesen' + 0.010***großflächig' + 0.009***teilverhaben'')																								

## 6.2 Intellektuelle Erschließung

- Den Topic IDs wird jeweils ein sinnvoller Begriff zugeordnet, der das Topic zusammenfasst

1	Topic ID	Term
2	1	Mentoring
3	2	Kapital
4	3	Unternehmen
5	7	Animation
6	9	Thermodynamik
7	11	Handel, Globalisation
8	12	Vermessung
9	13	Dispersion
10	14	Forschungsdaten
11	16	Immersion
12	22	Medien
13	23	Inklusion, Bildung
14	26	Pharmazeutik
15	27	Architektur, Philosophie
16	29	Hybrid, Kinetik, Energiespeicher, Fahrzeugentwicklung
17	30	Statistik
18	40	Ökologie
19	41	Mechanik
20	43	Restaurierung
21	46	Textil
22	48	Informationswissenschaften, Kommunikationswissenschaften
23	55	Politik
24	57	Elektronik
25	62	Denkmal

## 6.2.1 Verknüpfung der Topics mit Dokumenten

- Verbindung der Topics und Dokumente sind durch LDA-Team gegeben

```
In [2]: 1 import pandas as pd
```

```
In [3]: 1 df = pd.read_excel('D:/StudiumDIS/5_Semester/knowledge_discovery/LDA_Thesaurus/Final/topics_distr_selected_whole_500.xlsx', '  
< >
```

```
In [4]: 1 print(df.loc[df['Topic_ID'] == 64])
```

	Index	Document_No	Topic_ID	Perc_Contribution
159	160	80	64	0.1040
2179	2180	1328	64	0.1574
2182	2183	1329	64	0.1792
2246	2247	1363	64	0.1055
2410	2411	1459	64	0.0771
2469	2470	1492	64	0.1423
2578	2579	1550	64	0.5128
2581	2582	1551	64	0.9159
2596	2597	1557	64	0.1621
2845	2846	1684	64	0.1375

## 6.2.1 Verknüpfung der Topics mit Dokumenten

- Diese Verknüpfung kann nun weiterverwendet werden

	A	B	C	D	E	F	G	H	I	J	K	L
1	Topic ID	Term	Document IDs									
2		1 Mentoring	224,400,532,1352,1614,1739,1891,2280,2415,2447,2655,2759,2830,2964,2967,2969,3269,3529,3662,4234,5151,5267,6062,6625,6698,7135,7138,7206,10048,10925,114									
3		2 Kapital	1079,1723,2501,3479,3501,4104,4150,5444,5451,5695,5724,5911,6516,7175,10434,10631,10666,12422,12808,12810,12813,12836,12856,12858,15034									
4		3 Unternehmen	58,1290,1616,1618,1621,1878,1890,2009,2301,2435,2601,2630,3843,4158,4684,5174,5292,5306,5398,5779,6668,13954									
5		7 Animation	714,867,973,2505,2742,2868,3256,3475,3490,3621,4023,4438,4506,4763,4868,4937,5199,5258,5332,5348,5760,5770,5771,5916,5923,6008,6033,6197,6386,6532,6536,6822,									
6		9 Thermodynamik	210,1666,1758,2506,2583,2851,3055,3558,4076,4892,5133,6637,6739,6864,7281,10772,11670,11729,11996,12954,13772,15148,15188									
7		11 Handel, Globalisation	443,1287,1422,1501,1906,2110,2273,2442,2560,2634,2640,3182,3245,3389,3418,3502,4521,4557,4635,6622,6919,7046,7497,10151,10870,11346,12201,12606,13037,13570									
8		12 Vermessung	235,3122,4328,4602,4659,5076,5388,13177									
9		13 Dispersion	244,252,272,1223,1246,2368,2726,3535,3662,4703,4767,4798,4904,5002,6670,7219,10260,11056,15036,15106,15150,15186,15325									
10		14 Forschungsdaten	121,336,1712,2408,3091,3926,4167,4839,4882,4936,4960,5448,6552,6617,6618,7504,13520,14532,14596,15355,15379									
11		16 Immersion	377,406,684,806,972,1012,1077,1288,1290,1340,1583,1878,1961,2043,2300,2358,2874,3144,3207,3402,3436,3441,3514,3534,4323,4325,4766,4823,4900,5050,5148,5172,57									
12		22 Medien	7,9,59,66,74,162,249,287,312,357,381,402,424,1149,1458,1543,1897,2104,2107,2108,2109,2293,2295,2494,2813,2815,3797,3798,4121,4324,4462,4507,5258,6152,6174,642									
13		23 Inklusion, Bildung	7,74,135,159,313,343,350,407,646,655,664,667,668,669,672,676,1074,7085,1485,1494,1617,1647,1799,1801,1899,1912,2171,2409,2414,2577,4487,5165,5305,5307,5523,57									
14		26 Pharmazeutik	190,221,246,254,286,341,551,1011,1112,1553,1607,1635,1860,1877,1951,2335,2354,2497,2613,2738,2897,3145,3178,3208,3230,3241,3453,4418,4508,4645,4884,4899,5154									
15		27 Architektur	1082,2134,2229,2392,2842,3117,3183,3246,4861,4940,5274,6471,10309,10970,13935									
16		29 Hybrid	45,203,204,831,1597,3422,3542,3582,3657,4678,7019,13128,15004,15060,15308,15309									
17		30 Statistik	254,257,483,877,909,1112,1535,2488,2495,2621,3047,3124,3666,4130,4669,4733,4909,4983,5055,5285,5313,5666,7173,10574,14111,14639,14640,15070,15078,15176									

- Nächster Schritt: Hierarchisierung der Topics

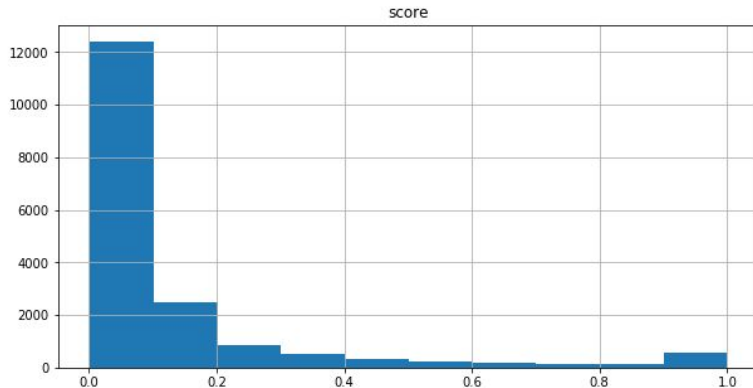
# 07 —

## Experten Empfehlung

# 7.1 Experten Ranking

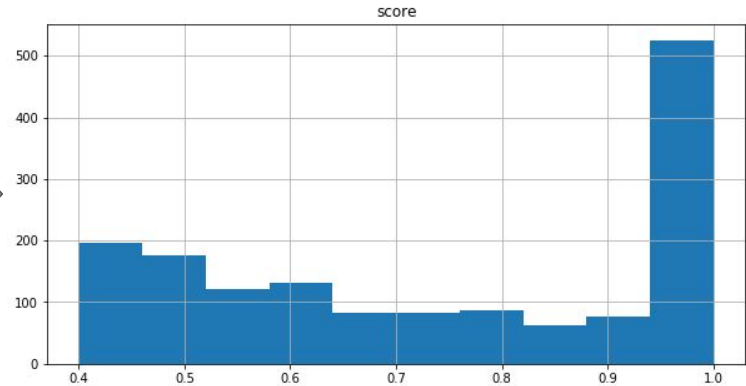
- prozentuale Zuordnung eines Topics zu den Dokumenten als Basis
- höhere Gewichtung von Personen im Left Panel
- Aggregation der Personen auf den höchsten Score
- Score-Normalisierung auf 0,1

person per topic: 35.587174348697395

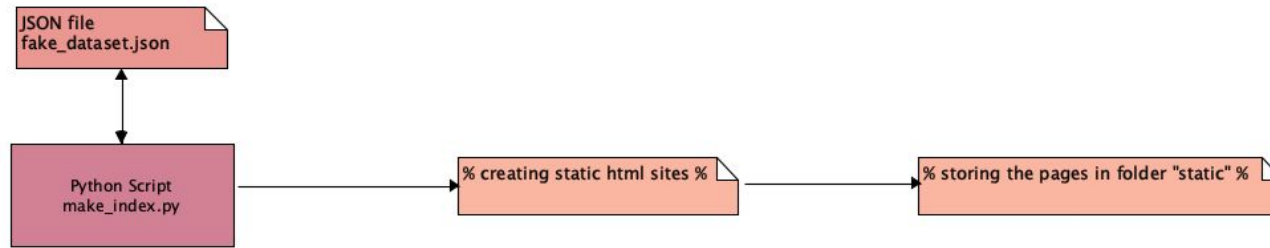


Threshold  
von 0.4

person per topic: 3.092184368737475



## 7.2 Das Frontend-System





# 7.2.1 Das Frontend-System

## TH Knowledge Discovery

### Top Level Topics:

- [Horror](#)
  - [Casey Jones](#)
  - [Broadway Bill](#)
  - [Inspector Bellamy \(Bellamy\)](#)
  - [Broken](#)
  - [Hunting Party, The](#)
  - [Animal Factory](#)
  - [Take Care of Your Scarf, Tatiana \(Pidä huivista kiinni, Tatjana\)](#)
- [Documentary](#)
  - [Clear History](#)
  - [Ballad of Nessie, The](#)
  - [Big Doll House, The](#)
  - [Monsieur Verdoux](#)
  - [I Am Legend](#)
  - [Diary of a Chambermaid, The](#)
  - [Xingu](#)



## TH Knowledge Discovery

[zurück](#)

### Sub Level Topics for "Horror":

[Casey Jones](#)  
[Broadway Bill](#)  
[Inspector Bellamy \(Bellamy\)](#)  
[Broken](#)  
[Hunting Party, The](#)  
[Animal Factory](#)  
[Take Care of Your Scarf, Tatiana \(Pidä huivista kiinni, Tatjana\)](#)



## TH Knowledge Discovery

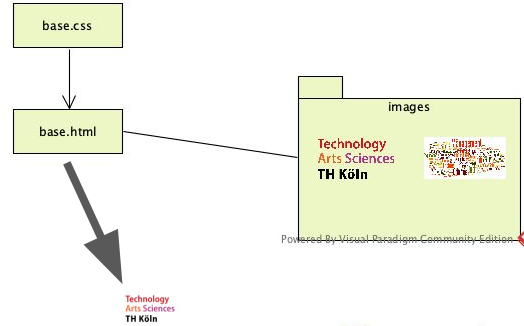
[home](#) [zurück](#)

### Experts for "Hunting Party, The":

Person	URL	Score
Dulcine Lefever	<a href="#">Link</a>	0.88
Jere Votier	<a href="#">Link</a>	0.73
Ring Dockray	<a href="#">Link</a>	0.3
Selena Bleue	<a href="#">Link</a>	0.28



# 7.2.2 Das Frontend-System



Auf der Suche nach Experten der TH Köln?  
Benutzen Sie unsere RKD Search/TH\_Knowledge\_Discovery

## Top Level Topics:

- Horror
  - Casey Jones
  - Broadway Bill
  - Inspector Bellamy (Bellamy)
  - Broken
  - Hunting Party, The
  - Animal Factory
  - Take Care of Your Scarf, Tatiana (Pida huivista kinni, Tatjana)
- Documentary
  - Clear History
  - Balled of Nettle, The
  - Big Doll House, The
  - Monsieur Verdoux
  - I Am Legend
  - Diary of a Chambermaid, The
  - Xingu

Technology  
Arts Sciences  
TH Köln



Auf der Suche nach Experten der TH Köln?  
Benutzen Sie unsere RKD Search/TH\_Knowledge\_Discovery

zurück

## Sub Level Topics for "Horror":

- Casey Jones
- Broadway Bill
- Inspector Bellamy (Bellamy)
- Broken
- Hunting Party, The
- Animal Factory
- Take Care of Your Scarf, Tatiana (Pida huivista kinni, Tatjana)

Technology  
Arts Sciences  
TH Köln



Auf der Suche nach Experten der TH Köln?  
Benutzen Sie unsere RKD Search/TH\_Knowledge\_Discovery

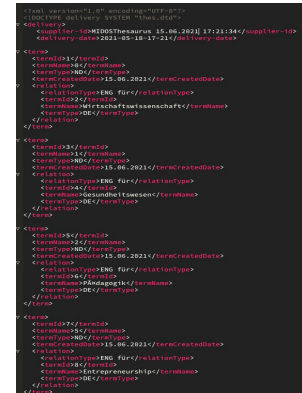
home zurück

## Experts for "Broadway Bill":

Person	URL Score
Berkly Yeude	Link 0.98
Shelby Chmarny	Link 0.61
Whitby Blusch	Link 0.61
Oates Dummings	Link 0.33

**Lösung:** Direktes Ansprechen der Datenbank bei Aufruf eines Topics. PHP und/oder Python Flask im Hintergrund laufen lassen.

**Lösung:** Veränderung des Scripts, damit XML-Dateien eingelesen werden können.



# 08 —

## Ergebnisse & Ausblick

## 8.1 Ergebnisse & Ausblick

- ★ Browsing Ansatz → Interviews zu eingehenden Anfragen
- ★ Dokumenten-Scraping → Ad-Hoc Ansatz
- ★ automatische Topic-Generierung → TH-Köln Thesaurus
- ★ Expertenranking → Implementierung in TH-Website
- ★ Web-Frontend → Zusammenhängendes Backend
- ★ **Alphaversion eines Recommender Systems**

**vielen Dank!**

