

# iNews-Pipeline

Implementierung einer Dokumenten-Pipeline für  
Nachrichtenartikel

Wintersemester 2019/20

# Gruppenmitglieder

Ahmad Alkurdi  
Lisa Hillebrand  
Flip Jansen  
Damian Lippert

Marie Kemker  
Rosanna Baltzer  
Laura Liegener  
Philipp Arndt

betreut durch Prof. Dr.-Ing. Hendrik Gärtner

# Gliederung

1. Gruppenorganisation
2. Use Case
3. Technischer Aufbau
  - a. Architektur
  - b. Scraping
  - c. Pipeline mit UIMA
  - d. mongoDB
  - e. Elasticsearch
  - f. HTTP-API
  - g. Frontend
4. Schwierigkeiten und Ausblick

# 1. Gruppenorganisation

## **Wöchentliches Teamtreffen**

mit Update über aktuellen Arbeitsstand,  
Aufgabenverteilung und Protokollführung





## **Aufteilung in Arbeitsgruppen**

Scraping, Frontend, REST-Schnittstelle,  
Elasticsearch, UIMA-Pipeline



## 2. Use Case





coronavirus

trump

auschwitz

virus

wuhan

bryant

israel

merkel

apple


gabriel

Stichwortsuche

Suche nach AutorInnen ▾

Quellen ▾


RESET




Wissen

Gesellschaft


Regional

 Lesezeit: 2 Minuten


Coronavirus: Drei weitere




Politik

 Lesezeit: 4 Minuten

Trump plant "realistische

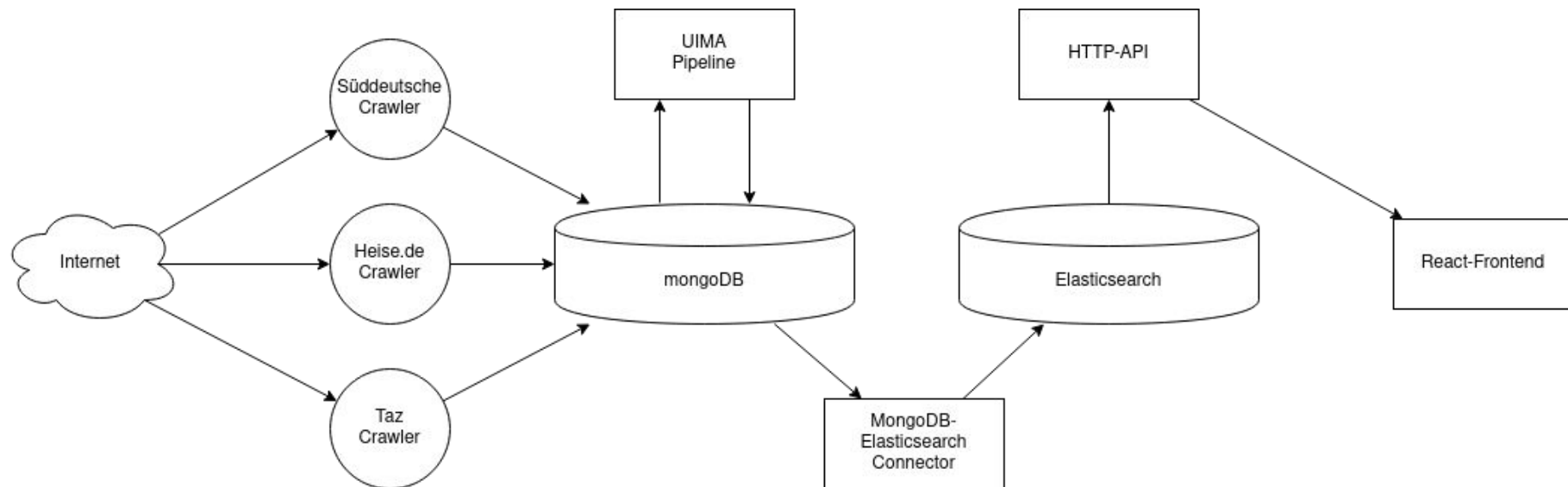


Kultur

 Lesezeit: 3 Minuten

Semperopernball:

# 3a. Architektur



## 3b. Scraping

### Aufgaben:

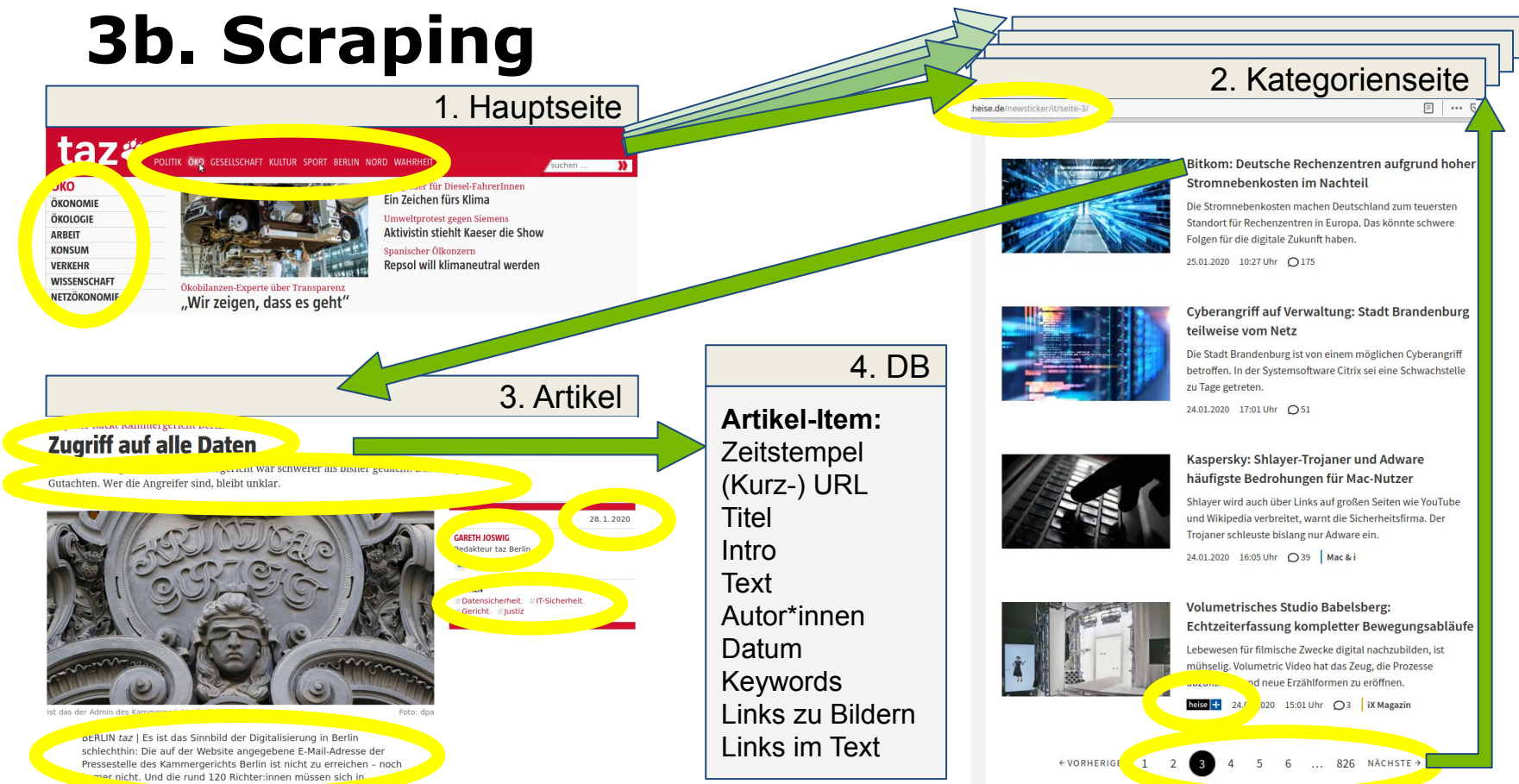
- regelmäßig neue Artikel von News-Seiten “scrapen”, bestimmte Merkmale erfassen und in eine Datenbank schreiben.
- aktuell:
  - taz
  - Süddeutsche Zeitung
  - Heise

### Technologien:

- Scrapy
- mongoDB
- Kibana (mit Elasticsearch)

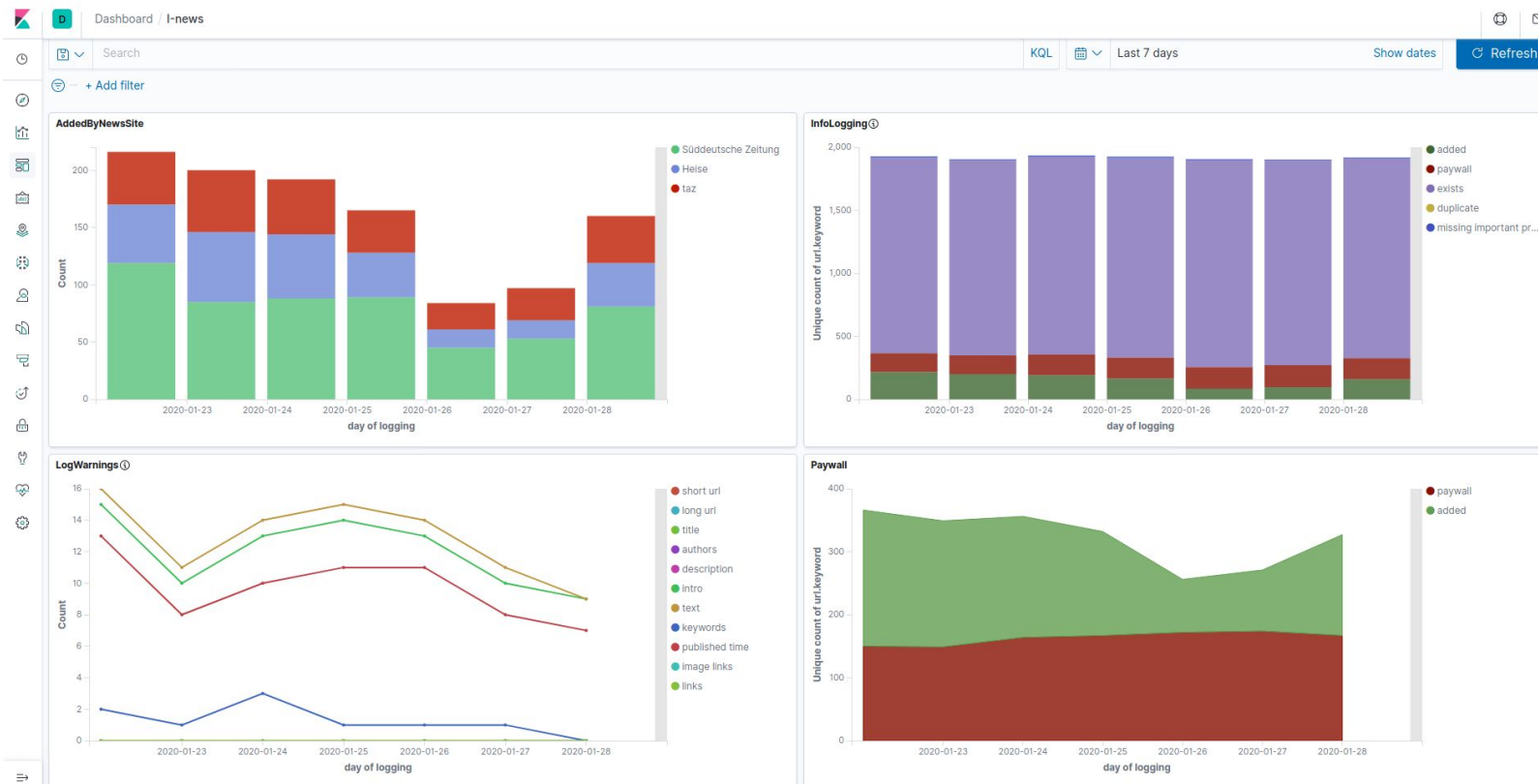


# 3b. Scraping





# 3b. Scraping



# 3c. UIMA-Pipeline

## Technologien:

- UIMA
- DKPro
- Spark

## Aufgaben:

- die gescrapten Artikel analysieren
- aktuell:
  - Lesezeit
  - relevanteste Wörter bzw. Objekte (nach Tf-Idf)
  - Lemmas
  - Zuordnung eines Departments

# 3c. UIMA-Pipeline

Reader

Lange galt Florian Schmidt, seit 2016 grüner Baustadtrat in Berlin-Kreuzberg, als „Robin Hood“ der Mieter. Nun gerät er unter Druck. Wofür steht dieser Mann?

Segmenter

Token: [“Lange“, “galt“, “Florian“, “Schmidt“, ““, „seit“, „2016“, “grüner“, “Baustadtrat“, “in“, “Berlin-Kreuzberg“, ““, “als“, ““Robin“, “Hood“, “der“, “Mieter“, “.“]

TokenTrimmer

Token: [“Lange“, “galt“, “Florian“, “Schmidt“, „seit“, „2016“, “grüner“, “Baustadtrat“, “in“, “Berlin-Kreuzberg“, “als“, “Robin“, “Hood“, “der“, “Mieter“]

NumberAndPunctuationRemover

Token: [“Lange“, “galt“, “Florian“, “Schmidt“, „seit“, “grüner“, “Baustadtrat“, “in“, “Berlin-Kreuzberg“, “als“, “Robin“, “Hood“, “der“, “Mieter“]

## 3c. UIMA-Pipeline

### NamedEntityRecognizer

NamedEntities: ["Florian" (Person),  
"Schmidt" (Person),  
"Berlin-Kreuzberg" (Ort),  
"Robin" (Person),  
"Hood" (Person)]

### NamedEntityMapper

NamedEntities: ["Florian Schmidt",  
"Robin Hood"]

### ReadingTimeEstimator

300  
words



ReadingTime: 2

### StopwordRemover

Token: ["Lange", "galt", "Florian",  
"Schmidt", "grüner", "Baustadtrat",  
"Berlin-Kreuzberg", "Robin", "Hood",  
"Mieter"]

# 3c. UIMA-Pipeline

Lemmatizer

Lemmas: ["lang", "gelten", "florian",  
"schmidt", "grün", "baustadtrat",  
"berlin-kreuzberg", "robin", "hood",  
"mieter"]

IdfModelWriter

Lemmas  
NamedEntities



```
{..., „mieter“:  
8,76, „florian  
schmidt“:  
9,45,  
... }
```

TfIdfCalculator

```
{..., „mieter“:  
8,76, „florian  
schmidt“:  
9,45,  
... }
```



MostRelevantLemmas:  
["mieter", "florian  
schmidt"]

Writer


# 3c. UIMA-Pipeline

iNews

coronavirus trump auschwitz israel virus wuhan merkel bryant apple gabriel

Suche nach AutorInnen

Quellen

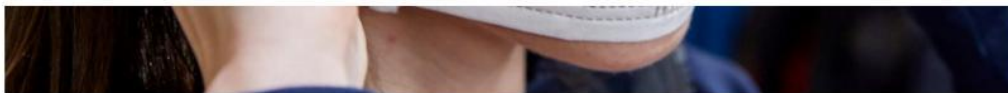


Meinung

Lesezeit: 1 Minuten

Rentenreform in Frankreich - Mehr Finte als Fortschritt

Leo Klimm



Auch Bayern hat das Coronavirus mittlerweile erreicht. Obwohl das Gesundheitsministerium eine weitere Ausbreitung für nicht wahrscheinlich hält, bleibt ein Restrisiko. Wie man einer Ansteckung vorbeugen kann.

 Lesezeit: 2 Minuten

Eine Infektion mit dem neuen Coronavirus zu diagnostizieren, ist ohne eine molekularbiologische Untersuchung der Erreger kaum möglich. Manche Infizierte haben gar keine oder nur milde Symptome, die an eine Erkältung erinnern. Fieber, Husten- und Kurzatmigkeit listet die amerikanische Seuchenschutzbehörde CDC auf.

Wenn es überhaupt Symptome gibt, dann treten sie zwei bis 14 Tage nach der Infektion auf. Doch es gibt eben auch schwere Krankheitsverläufe; vor allem für Menschen mit Vorerkrankungen kann das Virus lebensgefährlich werden. Wer sich krank fühlt, in China war oder mit Menschen Kontakt hatte, die in China waren, sollte das dem Arzt unbedingt mitteilen.

Die üblichen Hygienemaßnahmen, die auch vor Erkältungen, Grippe- und Durchfallviren schützen, helfen am besten gegen eine Ansteckung: regelmäßiges und sorgfältiges Händewaschen mit Seife und am besten mit warmem Wasser, "Husten- und Nies-Etikette" wahren, wie das Robert-Koch-Institut schreibt, also in die Armbeuge zielen oder besser in ein Taschentuch und dieses gleich entsorgen, Abstand halten zu Erkrankten.

Gesichtsmasken, wie sie nun häufig in China zu sehen sind, bieten nur bedingt Schutz. Ob sie funktionieren, hängt von vielen Faktoren ab: Wird die Maske richtig und konsequent getragen? Schließt sie also überall bündig mit der Haut ab? Wird sie regelmäßig gewechselt? Wie gut filtert die Maske? Ein Mundschutz hilft jedoch auch unabhängig von der Porengröße: Er erinnert den Träger und die Trägerin daran, nicht mit den eigenen Fingern ins Gesicht zu fassen. Denn die Schleimhäute von Augen, Mund und Nase sind die wichtigsten Einfallstore für viele Virusarten. Und ziemlich sicher schützt man mit einer Maske andere Menschen, wenn man selbst krank ist.

coronavirus

maske

ansteckung

<https://www.sueddeutsche.de/bayern/bayern-coronavirus-schutz-gesundheit-tipps-1.4775954>

## 3d. mongoDB



- Dokumentenbasierte NoSQL Datenbank
- Basiert quasi auf JSON-Dokumenten
- unser zentraler Datenspeicher
  - hält die Daten der Scraper und der UIMA-Pipeline

```
1 {  
2   "" : ObjectId( "5ded44bbd15247dbbad979b2" ),  
3   "short_url" : "https://sueddeutsche.de/1.4714672",  
4   "keywords" : [ "Boxen", "Sport" ],  
5   "crawl_time" : ISODate( "2019-12-08T19:45:15.557Z" ),  
6   "published_time" : ISODate( "2019-12-08T14:51:32Z" ),  
7   "news_site" : "sz",  
8   "...": "..."  
9 }
```



## 3e. Elasticsearch

```
1 GET processed_mongo_data/_search
2 {
3   "query":{
4     "match_phrase": {
5       "title": "München vor 30 Jahren: Als Tausende DDR-Bürger kamen"
6     }
7   }
8 }
```

- Auf JSON basierende Suchmaschine
  - nutzt JSON für Anfragen und Antworten
- Teil des Elastic Stacks
  - Kibana ist ein Tool zur Visualisierung und Analyse
- Stellt die Suchfunktion bereit



# 3f. HTTP-API

## Aufgaben:

- **Zeitungsartikel nach außen sichtbar machen**
- **Analytics aggregieren**
- Elasticsearch-Abfragen (suchen, filtern und aggregieren)
- GET-Anfragen bearbeiten
- Antwort mit JSON-Daten

## 3f. HTTP-API: Technologiewahl



Scalatra



# 3f. HTTP-API: API Definition

## News Pipeline 1.0.0 OAS3

Api documentation for news pipeline

### analytics View analytics for news articles

**GET** **/analytics/terms** Get term occurrences per day

**GET** **/analytics/lemmas** Get most relevant lemmas for the last 7 days

### articles Retrieve news articles

**GET** **/articles** Get news articles

**GET** **/articles/{articleId}** Find article by id

**GET** **/articles/departments** Get departments

**GET** **/articles/newspapers** Get newspapers

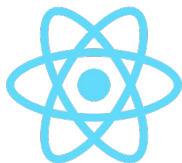
**GET** **/articles/authors** Get authors

## 3f. HTTP-API: Beispiel einer Antwort

```
▼ 0:
  authors:      [...]
  crawlTime:    1580259609607
  departments:  [...]
  description:  "Wie der erste deutsche P...rgehend seine Zentrale."
  id:           "5e30cb098a84911cfde51ee5"
  imageLinks:   [...]
  intro:        "Die drei weiteren Infizi...ndheitsministerium mit."
  keywords:     [...]
  lemmas:       [...]
  links:        [...]
  longUrl:       "https://www.sueddeutsche...chland-webasto-1.4776016"
  mostRelevantLemmas: [...]
  newsSite:     "sz"
  publishedTime: 1580250569000
  readingTime:  2
  text:         "Das Coronavirus hat sich...ankheit gestorben sein."
  title:        "Coronavirus: Drei weitere Infizierte in Bayern"
```

# 3g. Frontend

ReactJS



- 2011 innerhalb von Facebook entwickelt
- Komponenten
- State und Props
- Virtual DOM

# 3g. Frontend

Libraries:

- react-router
- MaterialUI
- react-vis

# 4. Schwierigkeiten und Ausblick

## Herausforderndes Zusammenspiel der Komponenten

- einfache Anbindung:  
Scrapy zu MongoDB, ElasticSearch zu Kibana
- aufwendigere Workarounds z.B. bei UIMA:  
schlecht dokumentiert, insb. in Kombination mit Scala

## Scraping: Veränderung des Aufbaus von News-Websites

- in Zukunft erwartbar → Scraper muss angepasst werden



# 4. Schwierigkeiten und Ausblick

## **Features, die wir gerne noch implementiert hätten:**

- UIMA zur Verarbeitung von Suchanfragen in REST-API einbinden
- weitere Analytics:
  - Zu welchem Department (bspw. Politik) erscheinen die meisten Artikel?
  - Zu welcher Tageszeit erscheinen die meisten Artikel?
  - o.ä.
- Scraping: mehr News-Seiten einbinden

“

**Danke für eure  
Aufmerksamkeit!**

# Bildquellen

## 1. Gruppenorganisation

[<https://slack.com/intl/de-de/>, letzter Zugriff: 30.01.2020]

[<http://pngimg.com/download/73347>, letzter Zugriff: 30.01.2020]

## 3a. Scraping

[<https://medium.com/p/83d36732181/>, letzter Zugriff: 28.01.2020]

[<https://www.taz.de/>, letzter Zugriff: 28.01.2020]

[<https://taz.de/Trojaner-hackt-Kammergericht-Berlin/!5657177/>, letzter Zugriff: 28.01.2020]

[<https://www.heise.de/newsticker/it/seite-3/>, letzter Zugriff: 28.01.2020]

## 3d. mongoDB

[<https://www.mongodb.com/brand-resources>, letzter Zugriff: 30.01.2020]

## 3e. Elasticsearch

[<https://www.elastic.co/brand>, letzter Zugriff: 30.01.2020]

## 3f. Http-API

[<https://www.playframework.com/>, letzter Zugriff: 30.01.2020]

[<https://akka.io/>, letzter Zugriff: 30.01.2020]

[<http://scalatra.org/>, letzter Zugriff: 30.01.2020]

## 3g. Frontend

[<https://upload.wikimedia.org/wikipedia/commons/a/a7/React-icon.svg>, letzter Zugriff: 30.01.2020]



**Hochschule für Technik  
und Wirtschaft Berlin**

University of Applied Sciences

[www.htw-berlin.de](http://www.htw-berlin.de)



coronavirus

trump

auschwitz

virus

wuhan

bryant

israel

merkel

apple

gabriel

Stichwortsuche

Suche nach AutorInnen ▼

Quellen ▼

RESET



Wissen

Gesellschaft

Regional

🕒 Lesezeit: 2 Minuten

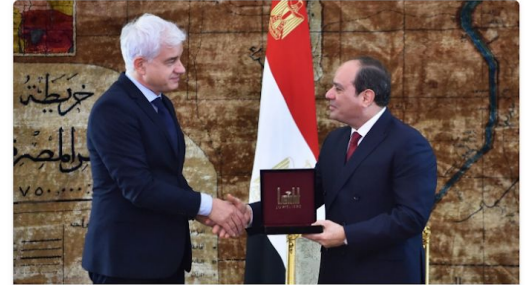
Coronavirus: Drei weitere



Politik

🕒 Lesezeit: 4 Minuten

Trump plant "realistische



Kultur

🕒 Lesezeit: 3 Minuten

Semperoperball:



Palästinenserpräsident Abbas zeigt sich empört über den von Trump als "realistische Zwei-Staaten-Lösung" angepriesenen Nahost-Plan. Den hatte der US-Präsident zuvor präsentiert - Seite an Seite mit Israels Ministerpräsident Netanjahu.

coronavirus

trump

auschwitz

virus

wuhan

bryant

israel

merkel

apple

gabriel

Trump

Putin

Erdogan

