iNews-Pipeline

Implementierung einer Dokumenten-Pipeline für Nachrichtenartikel

Wintersemester 2020/21

Hochschule für Technik und Wirtschaft Berlin
University of Applied Sciences

Gruppenmitglieder

Lukas Drews
Janis Schanbacher
Katharina Koslowski
Malte Till Asbjörn Gronwald
Lennart Hendrik Döring
Dennis Dominik Lehmann
Andreas Scharnetzki
Steven Schütte

Betreut durch Prof. Dr. -Ing. Hendrik Gärtner

Gliederung

1. Einleitung
2. Gruppenorganisation
3. Technischer Aufbau
   a. Architektur
   b. Scraping
   c. Pipeline mit Apache Spark
   d. Webanwendung
4. Learnings
5. Ausblick
6. Ethik

2. Gruppenorganisation

Wöchentliche Video-Konferenz

Update über aktuellen Arbeitsstand,
Planung der Wochenziele & Meilensteine,
Besprechung von Problemen &
Entscheidungen

Aufteilung in Arbeitsgruppen

Scraping (Janis, Lukas)
NLP-Pipeline (Dennis, Andreas, Lennart, Steven)
Backend (Malte, Katharina)
Frontend (Katharina)

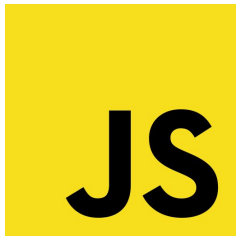
3a. Architektur - Techstack

Get Data

Analyze&Store Data

Provide Data

Visualize Data



3a. Architektur - Key Tech: SPARK

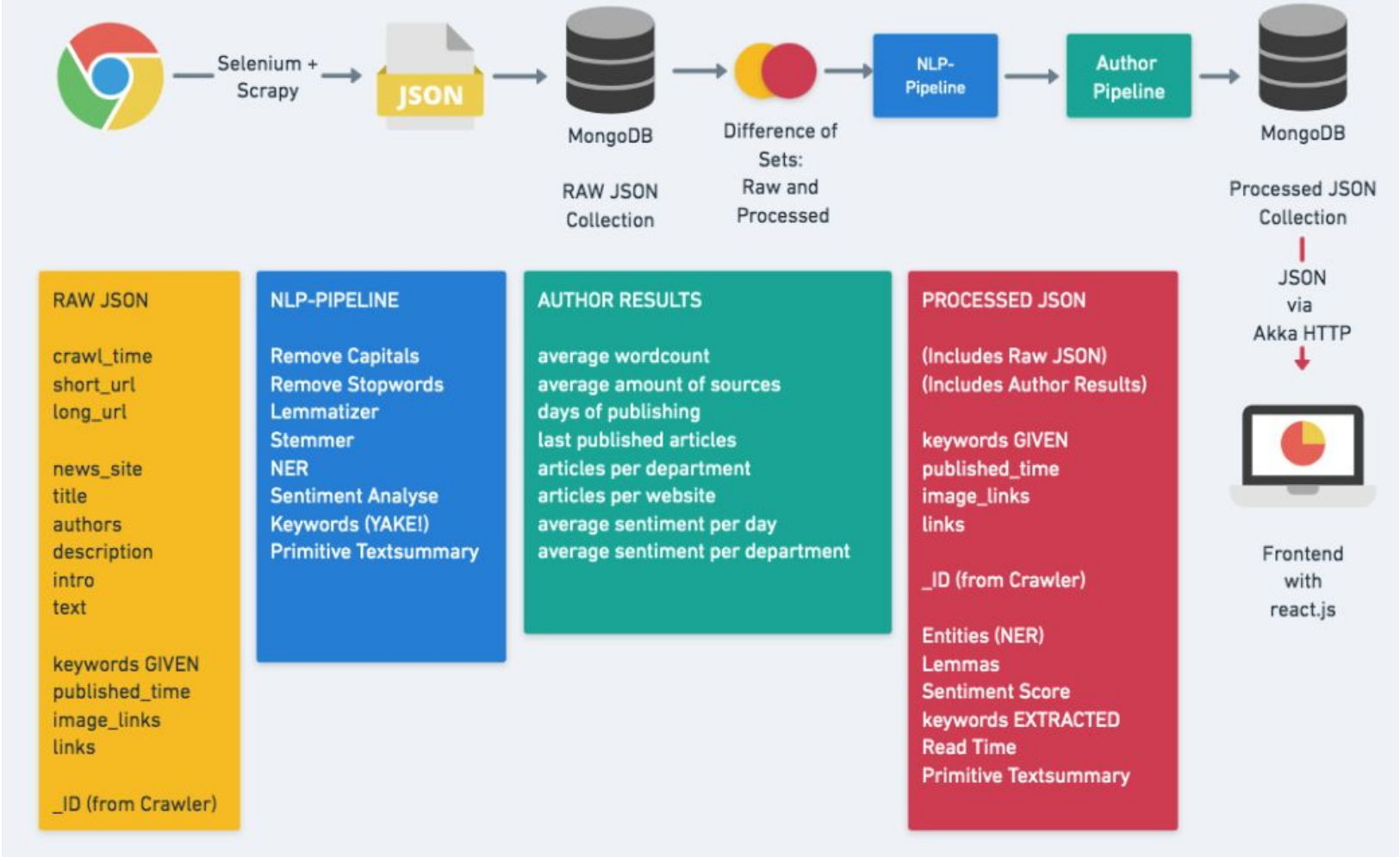
Framework für verteilte hoch parallelisierte In-Memory Berechnung von Big Data auf mehreren virtuellen oder physischen Maschinen.

Spark macht das Arbeiten mit Daten im Tera- oder Petabyte Bereich deutlich einfacher und oftmals deutlich effizienter.

Wir haben hauptsächlich die externe Library “Spark NLP” von John Snow Labs verwendet und eigene Algorithmen in Spark geschrieben.

3a. Architektur - Key Tech: Elastic Search

- Basiert auf Erstellung eines invertierten Index
- Index teilt sich in Typen und Dokumente
- Typen ähnlich wie Tabellen einer Datenbank
- Index kann viele Typen beinhalten
- Unterhalb der Typen sind die JSON-Dokumente
- Sie sind die kleinste Einheit des Index (enthalten die Daten)
- JSON-Dokumente bestehen aus Paaren von Schlüsseln und Werten
- Suchanfragen über REST API


Selenium + Scrapy
JSON
MongoDB
RAW JSON Collection
Difference of Sets: Raw and Processed
NLP-Pipeline
Author Pipeline
MongoDB
Processed JSON Collection
JSON via Akka HTTP
Frontend with react.js
RAW JSON
crawl_time
short_url
long_url
news_site
title
authors
description
intro
text
keywords GIVEN
published_time
image_links
links
_ID (from Crawler)
NLP-PIPELINE
Remove Capitals
Remove Stopwords
Lemmatizer
Stemmer
NER
Sentiment Analyse
Keywords (YAKE!)
Primitive Textsummary
AUTHOR RESULTS
average wordcount
average amount of sources
days of publishing
last published articles
articles per department
articles per website
average sentiment per day
average sentiment per department
PROCESSED JSON
(Includes Raw JSON)
(Includes Author Results)
keywords GIVEN
published_time
image_links
links
_ID (from Crawler)
Entities (NER)
Lemmas
Sentiment Score
keywords EXTRACTED
Read Time
Primitive Textsummary

3b. Scraping im iNews Projekt

- Regelmäßig neue Artikel von News-Seiten “scrapen”, ausgewählte Metadaten und Inhalte erfassen und in eine Datenbank schreiben
- Übernommene Seiten:
  - Taz
  - Süddeutsche Zeitung
  - Heise
- Neue Seiten:
  - Golem
  - Der Postillon

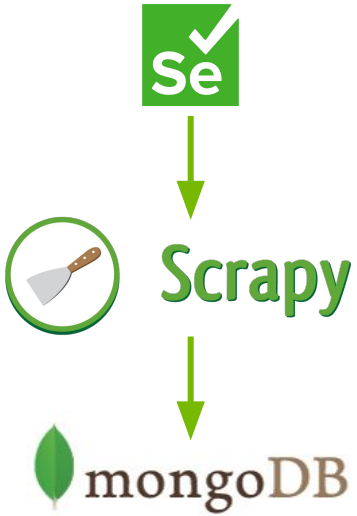
3b. Scraping & Crawling

- Web Scraping: Datenextraktion von Websites
- Web Crawling: Durchsuchen des Webs nach Links/Inhalten

- Crawling um die Links der Artikel zu sammeln, deren Inhalte und Metadaten mittels Scraping extrahiert werden

3b. Scraping - Technologien

- Python
- Selenium (Der Postillon)
  - Crawling: Dynamische Inhalte laden
- Scrapy
  - Crawling, Scraping
- mongoDB



3b. Scrapy vs. Selenium

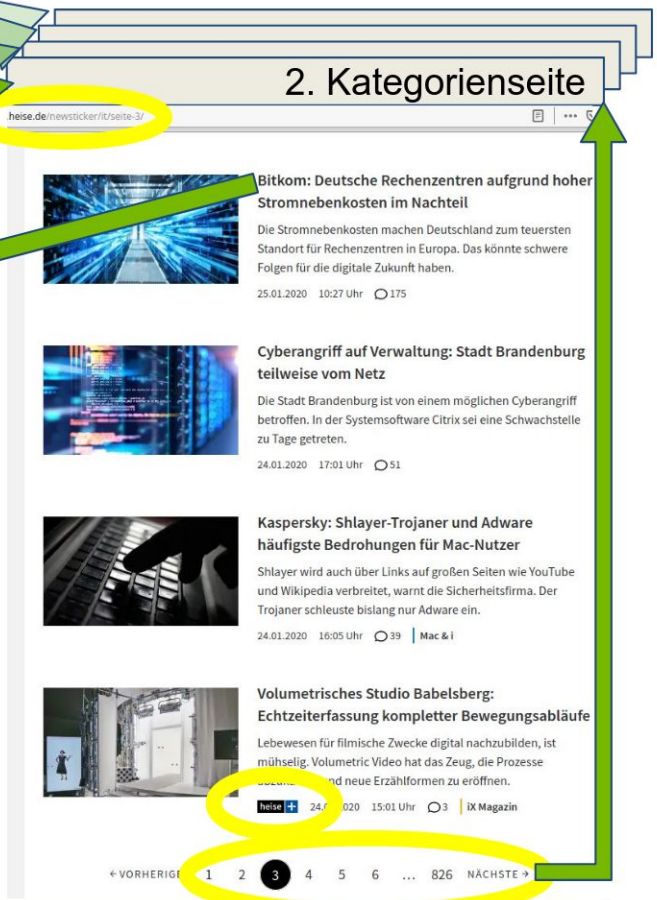
Scrapy:

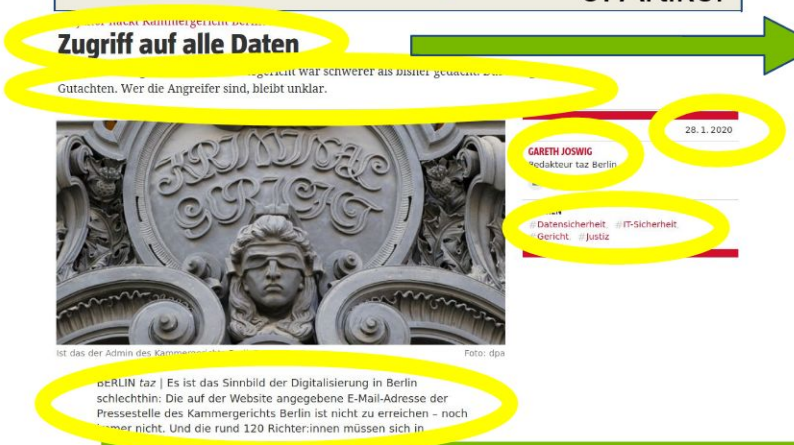
- Framework zum Crawlen und Scrapen
- Effizient Inhalte auslesen und Extrahieren
- Laden dynamischer Inhalte teilweise durch Requests möglich

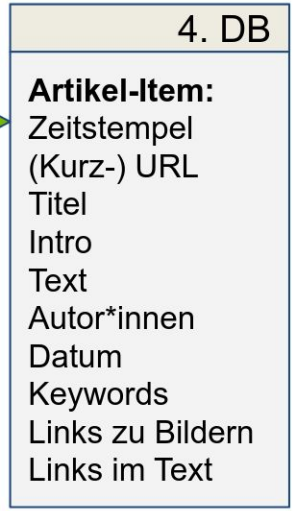
Selenium:

- Framework, welches für automatisiertes Testen ausgelegt ist
- Ermöglicht laden der dynamischen Inhalte/URLs durch Interaktion
- Headless Chromium (Browser ohne GUI)
- Zeit- und Speicherintensiv
  → Übergabe an Scrapy, sobald Seite vollständig geladen ist

3b. Scraping - Ablauf







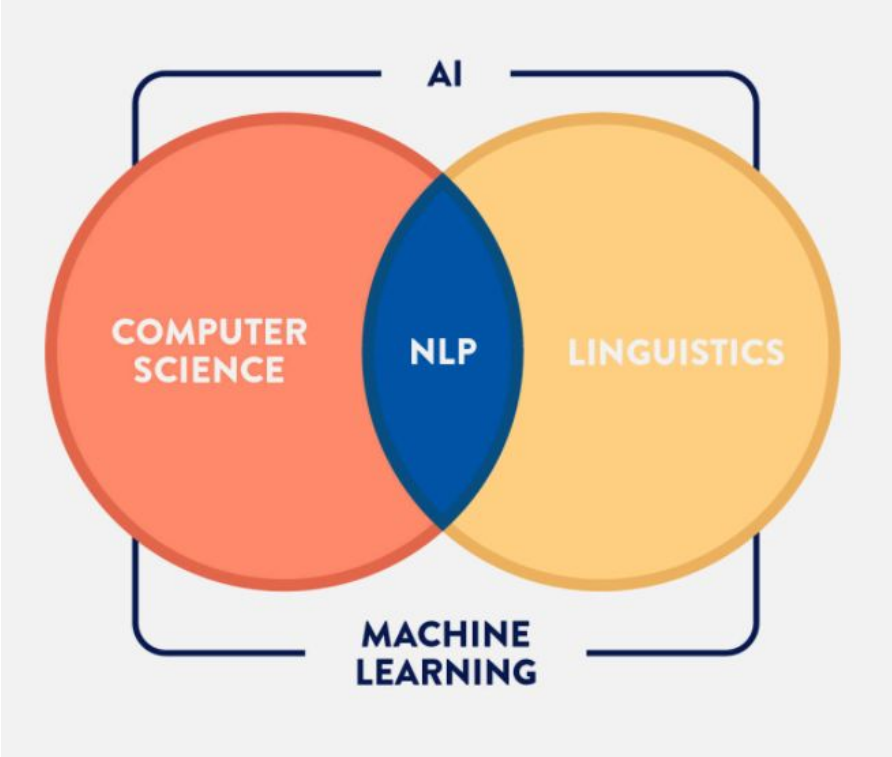
3b. Scraping - Ablauf: Der Postillon

- Archiv ist nach Daten statt Kategorien unterteilt
- Artikel-URLs sind erst nach Anklicken der entsprechenden Jahre/Monate im Quellcode enthalten
  → Selenium zum Laden URLs
- Scrapy: Sammeln der URLs, Scrapen der Artikel

Regelmäßiges Auslesen neuer Artikel

- Regelmäßiges Ausführen der Crawler mittels Cronjob
- Cronjob: Menge von Befehlen, die zu festgelegten Zeiten im Hintergrund ausgeführt werden

3c. What is Natural Language Processing



Quelle:
https://clevertap.com/blog/natural-language-processing
Stand: 12. Februar 2021

3c. NLP Pipeline mit Apache Spark Step By Step

Input Raw Text:

“Greta Thunberg sieht, wie meistens, besorgt aus, als sie im Stockholmer Nobelpreismuseum in einen Computer spricht. Ebenfalls im Raum, aber in sicherer Entfernung, ist der Moderator Gustav Källstrand, Nobelpreisexperte des Museums.”

3c. NLP Pipeline mit Apache Spark Step By Step: Capitals

Remove Capitals:

“greta thunberg sieht, wie meistens, besorgt aus, als sie im stockholmer nobelpreismuseum in einen computer spricht ebenfalls im raum, aber in sicherer entfernung, ist der moderator gustav källstrand, nobelpreisexperte des museums.”

3c. NLP Pipeline mit Apache Spark Step By Step: Stopwords

Remove Stopwords:

“greta thunberg sieht meistens besorgt stockholmer nobelpreismuseum computer spricht”

“ebenfalls raum sicherer entfernung moderator gustav källstrand nobelpreisexperte museums”

3c. NLP Pipeline mit Apache Spark Step By Step: Lemmas

Transformiere zu Lemmas mit Spark NLP:

“greta thunberg sieht meistens besorgt stockholmer nobelpreismuseum computer spricht. ebenfalls raum sicherer entfernung moderator gustav källstrand nobelpreisexperte museums.”

Ergebnis:

*“greta thunberg sehen meistens **besorgen** stockholmer nobelpreismuseum computer sprechen. ebenfalls raum sicher entfernung moderator gustav källstrand nobelpreisexperte museum.”*

3c. NLP Pipeline mit Apache Spark Step By Step: Entities

Identifiziere Entitäten

"Greta Thunberg" = Person
"Stockholmer Nobelpreismuseum" = Ort
"Gustav Källstrand" = Person

Gesetzentwurf: Online-Ausweis soll aufs Handy, wird aber teuer

Hervorheben

Lesezeit: 5 Minuten

ZUSAMMENFASSUNG BETA

Der mit dem elektronischen Personalausweis verknüpfte Online-Ausweis soll künftig einfacher eingesetzt werden können. Das Bundesinnenministerium (BMI) hat einen Referentenentwurf für ein Gesetz "zur Einführung eines elektronischen Identitätsnachweises mit einem mobilen Endgerät" vorgelegt. Bürger sollen ihren Online-Ausweis künftig direkt im Smartphone speichern können. Das soll die Akzeptanz elektronischer Ausweise verbessern und die Nutzung um mindestens die Hälfte steigern.

Dazu sollen die Gesetze für den Personalausweis, für die Karte für den elektronischen Identitätsnachweis (eID) und für den Aufenthalt ausländischer Staatsbürger "nutzerfreundlich" weiterentwickelt werden, schreibt das Innenressort. Der Online-Ausweis sei in seiner gegenwärtigen Form zwar allgemein zur sicheren Identifizierungs anerkannt. Seine Verbreitung "ist jedoch hinter den Erwartungen zurückgeblieben". Das Bundeskabinett wird den Vorschlag voraussichtlich am Mittwoch billigen und dann in den Bundestag einbringen.

Identitätsnachweis per Smartphone

Aktuell wird der elektronische Identitätsnachweis durch zwei Faktoren gewährleistet: Das Wissen der sechsstelligen Geheimnummer, und der Besitz von Personalausweis, eID-Karte oder elektronischem Aufenthaltstitel. Nun soll auch das Smartphone samt staatlicher App als Besitzmittel gelten.

Bürger sollen die Übertragung der notwendigen Schlüssel aus dem Speicher des Personalausweises auf das Handy online beantragen können, wobei sie sich per eID identifizieren müssen. Der Ausweishersteller muss Maßnahmen gegen missbräuchliche Verwendung der im Handy gespeicherten Daten treffen, beispielsweise durch einen weiteren Sperrschlüssel. Der Ausweisinhaber kann dem Plan nach die Daten der eID-Funktion auf seinem Endgerät selbst löschen.

Nicht billig

3c. NLP Pipeline mit Apache Spark Step By Step: Keywords

Extracted from whole Text with YAKE! Algorithm:

"thunberg",
"källstrand",
"chance",
"krise",
"pandemie".

Der Gesamtverband der deutschen Versicherungswirtschaft fordert, "dass neben einer abgeleiteten digitalen Identität aus dem Personalausweis auch andere Identifizierungsverfahren auf gleichem Vertrauensniveau anerkannt werden". Nötig sei ein "eID-Ökosystem".

(olb)

Online-Ausweis | Sicherheit | Smartphones | Tablets | Verwaltung | ePerso | elektronischer Identitätsnachweis | elektronischer Personalausweis | Netzpolitik | elektronisch | personalausweis | onlineausweis | smartphone | euro

09.02.2021

Stefan Krempl
https://www.heise.de/news/Gesetzentwurf-Online-Ausweis-soll-aufs-Handy-wird-aber-teuer-5049183.html

3c. NLP Pipeline mit Apache Spark Step By Step: TextSum

Primitive TextSum using complete text
(Identify Top 3 Sentences with most Keywords)

„Wir wissen, dass 2020 das entscheidende Jahr ist, wenn wir noch eine Chance haben wollen, die Erderwärmung auf 1,5 Grad Celsius zu begrenzen“, sagt Thunberg, die selbst als Kandidatin für den Nobelpreis gilt, nachdem sie im vergangenen Jahr leer ausging. Die Menschen hätten alles beiseite gelegt, um gemeinsam der Pandemie entgegenzutreten – „wie es in Krisen notwendig ist“, so Thunberg. Das Problem sei aber, dass die Klimakrise nicht wie eine Krise behandelt werde, sagt Thunberg."

3c. NLP Pipeline mit Apache Spark Step By Step: Analytics

Sentiment Analyse: In diesem Prozess wird jedem Wort des vorprozessierten Texts ein Sentiment-Wert zugeschrieben. Am Ende entsteht ein Wert der darauf hinweisen soll, welche emotionale Stimmung in einem Text überwiegt. Worte mit negativer Konnotation sind mit einem negativen Wert besetzt, wohingegen es sich umgekehrt bei Worten verhält, denen eine positive Bedeutung zugeschrieben wird.

"Das Problem sei aber, dass die Klimakrise nicht wie eine Krise behandelt werde, sagt Thunberg."

"**Problem** -> -0,3865, **Klimakrise** -> 0, **Krise** -> -0,3621."

Lesezeit: 5 Minuten

ZUSAMMENFASSUNG BETA

Der mit dem elektronischen Personalausweis verknüpfte Online-Ausweis soll künftig einfacher eingesetzt werden können. Bei ihm selbst und dem BSI (Bundesamt für Sicherheit in der Informationstechnik) entstünden einmalige Entwicklungskosten von mehr als 19 Millionen Euro und jährliche Kosten von über 26 Millionen Euro. Bürger sollen ihren Online-Ausweis künftig direkt im Smartphone speichern können.

3c. NLP Pipeline mit Apache Spark Analyse von Autoren: Idee

Die Qualität einer News Seite wird durch den Output ihrer Autoren bestimmt

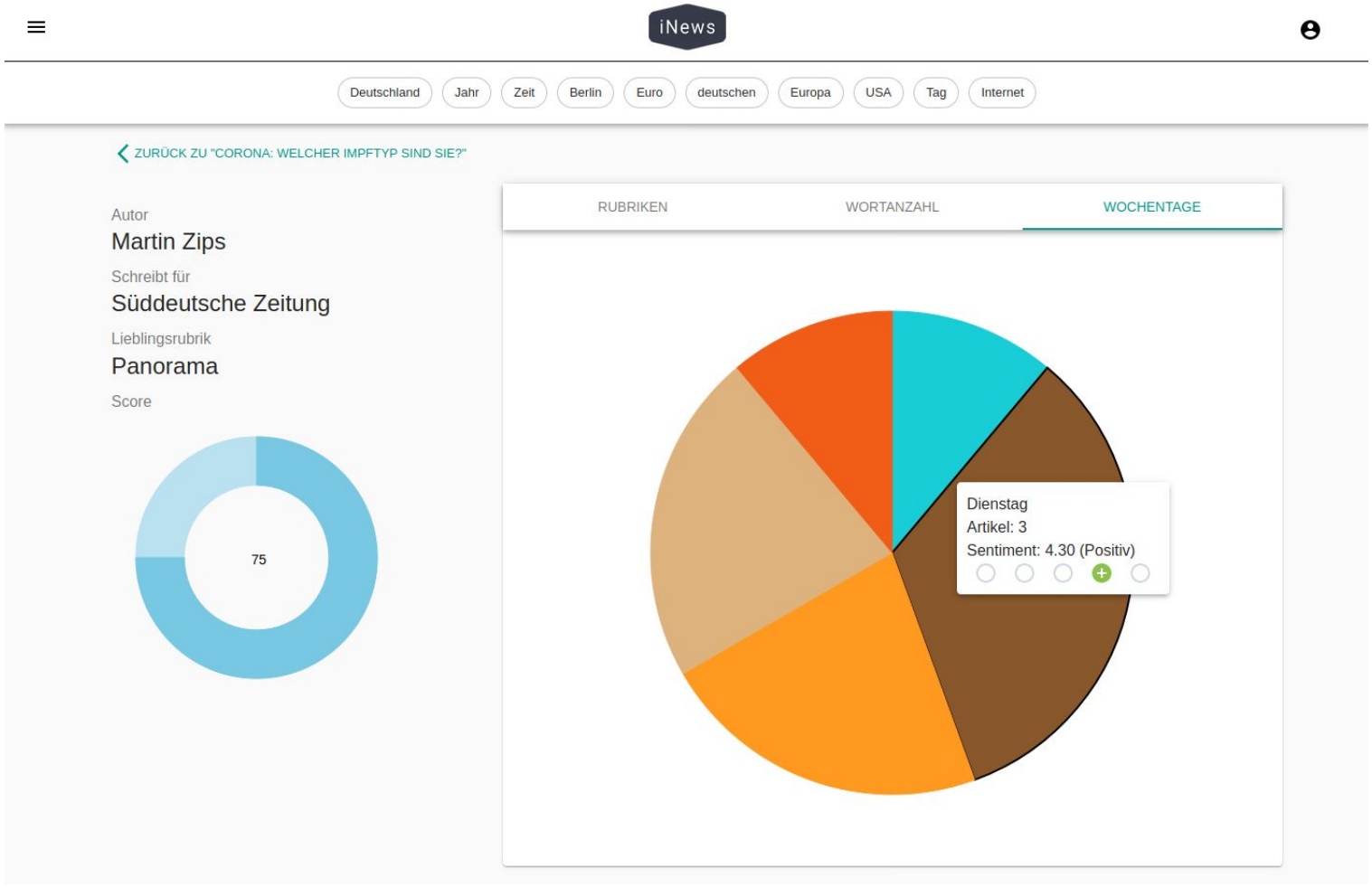
Autoren sind Menschen und daher nie ganz unabhängig.

Hypothese:
Wenn man die Autoren analysiert kann man die Qualität einer News Seite approximieren.

3c. NLP Pipeline mit Apache Spark Analyse von Autoren:

Ergebnisse:

- Durchschnittliche Sentiment Werte der Artikel der Autoren
- An welchen Wochentagen veröffentlichen Autoren
- Trust Score anhand der Anzahl und Quellen der Artikel
- Durchschnittliche Anzahl an Wörtern pro Artikel
- In welchen Ressorts


iNews
Deutschland
Jahr
Zeit
Berlin
Euro
deutschen
Europa
USA
Tag
Internet
ZURÜCK ZU "CORONA: WELCHER IMPFTYP SIND SIE?"
Autor
Martin Zips
Schreibt für
Süddeutsche Zeitung
Lieblingsrubrik
Panorama
Score
75
RUBRIKEN
WORTANZAHL
WOCHENTAGE
Dienstag
Artikel: 3
Sentiment: 4.30 (Positiv)

Deutschland | Jahr | Zeit | Berlin | Euro | deutschen | Europa | USA | Tag | Internet

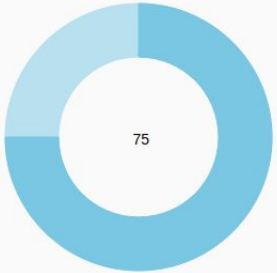
ZURÜCK ZU "CORONA: WELCHER IMPFTYP SIND SIE?"

Autor
Martin Zips

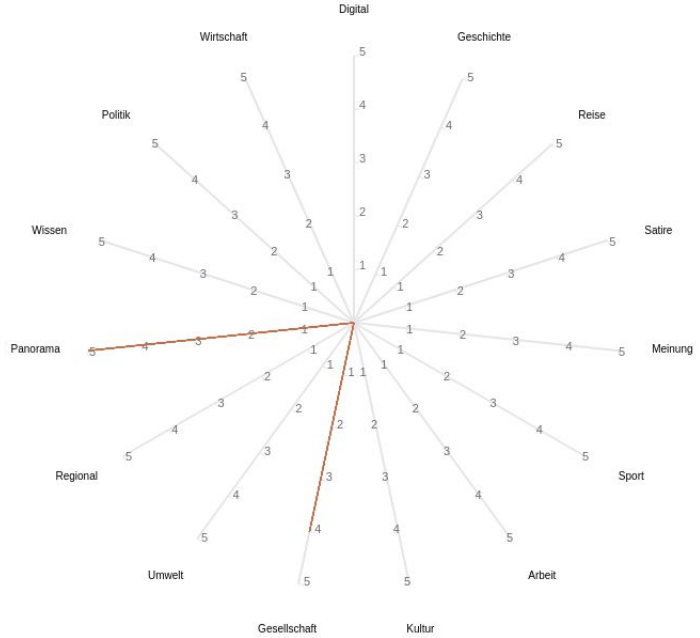
Schreibt für
Süddeutsche Zeitung

Lieblingsrubrik
Panorama

Score



RUBRIKEN | WORTANZAHL | WOCHENTAGE



iNews

Deutschland | Jahr | Zeit | Berlin | Euro | deutschen | Europa | USA | Tag | Internet

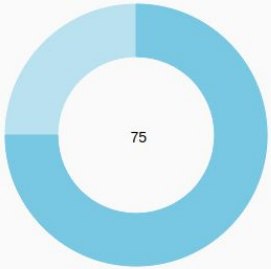
ZURÜCK ZU "CORONA: WELCHER IMPFTYP SIND SIE?"

Autor
Martin Zips

Schreibt für
Süddeutsche Zeitung

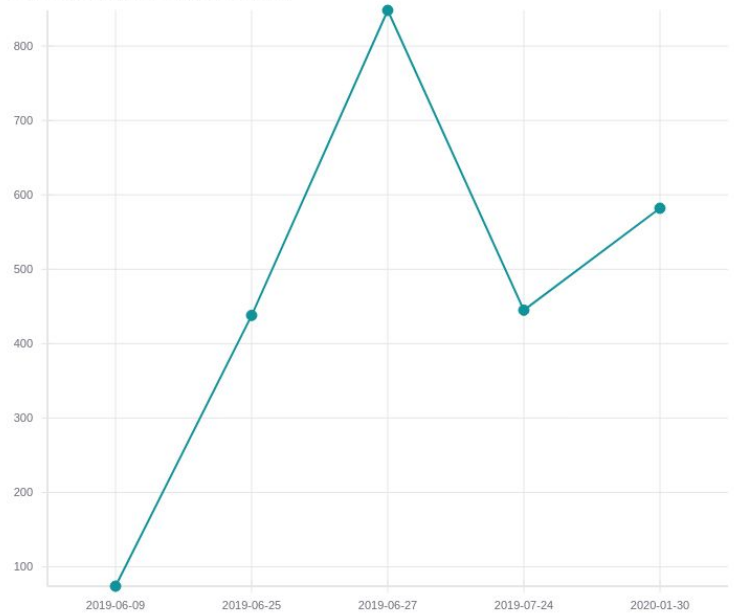
Lieblingsrubrik
Panorama

Score

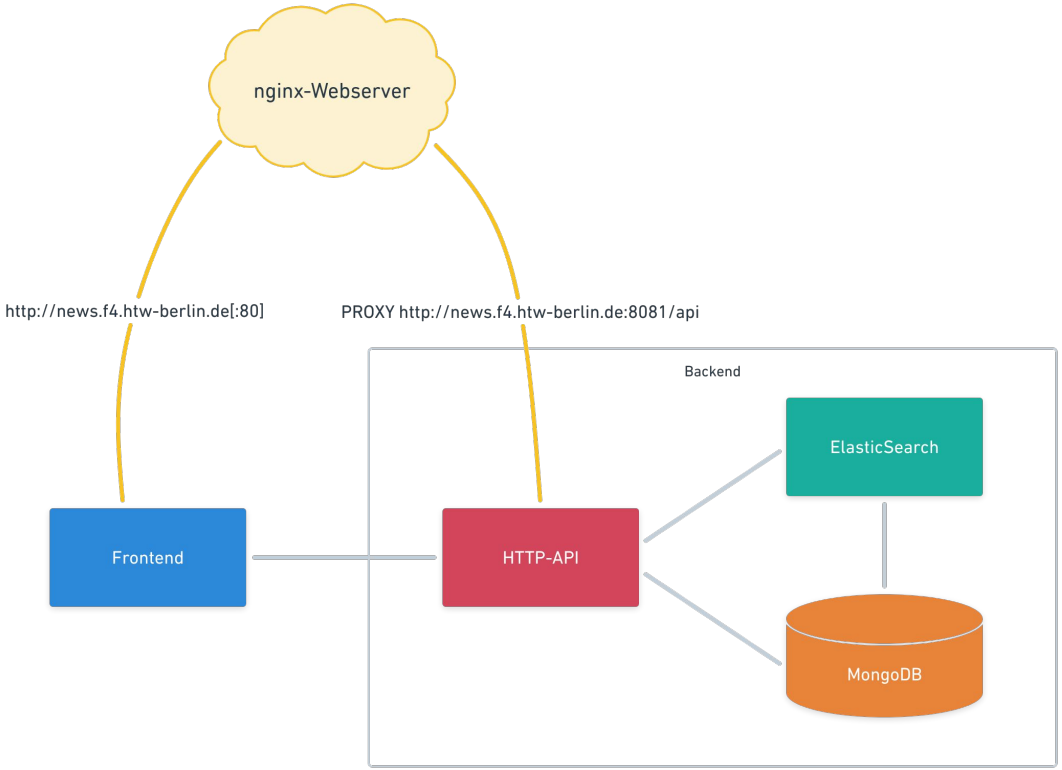


RUBRIKEN | WORTANZAHL | WOCHENTAGE

Durchschnitt: **513.44** Worte pro Artikel



3d. Webanwendung



3d. Webanwendung: HTTP-API

REST-Schnittstelle → Verarbeitung von Anfragen und Bereitstellung von Ressourcen (Artikel, Autoren, Userdaten)

Genutzte Technologien:

- Programmiersprache: Scala
- Akka-HTTP: Routing und Request-Verarbeitung
- elastic4s: Client für ElasticSearch
- MongoDB Scala Driver: Treiber Interaktion mit MongoDB

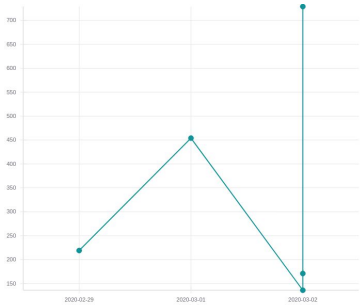
3d. Webanwendung: Frontend

Single-Page-Web-Application → Interaktion mit API über graphisches UI

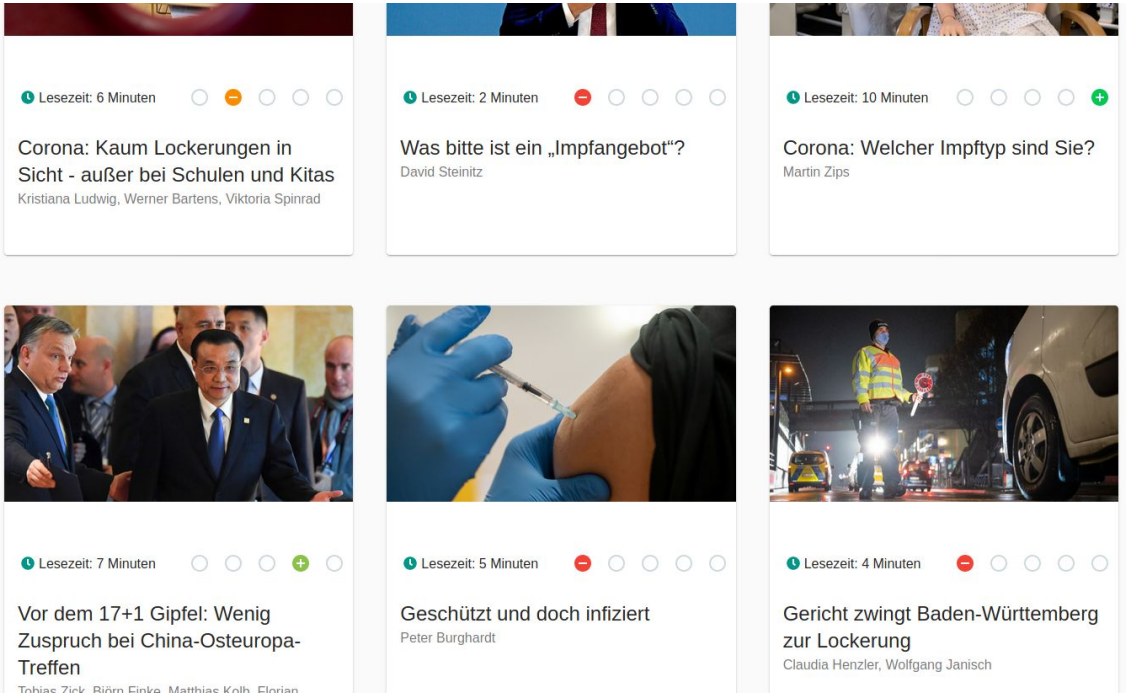
Genutzte Technologien:

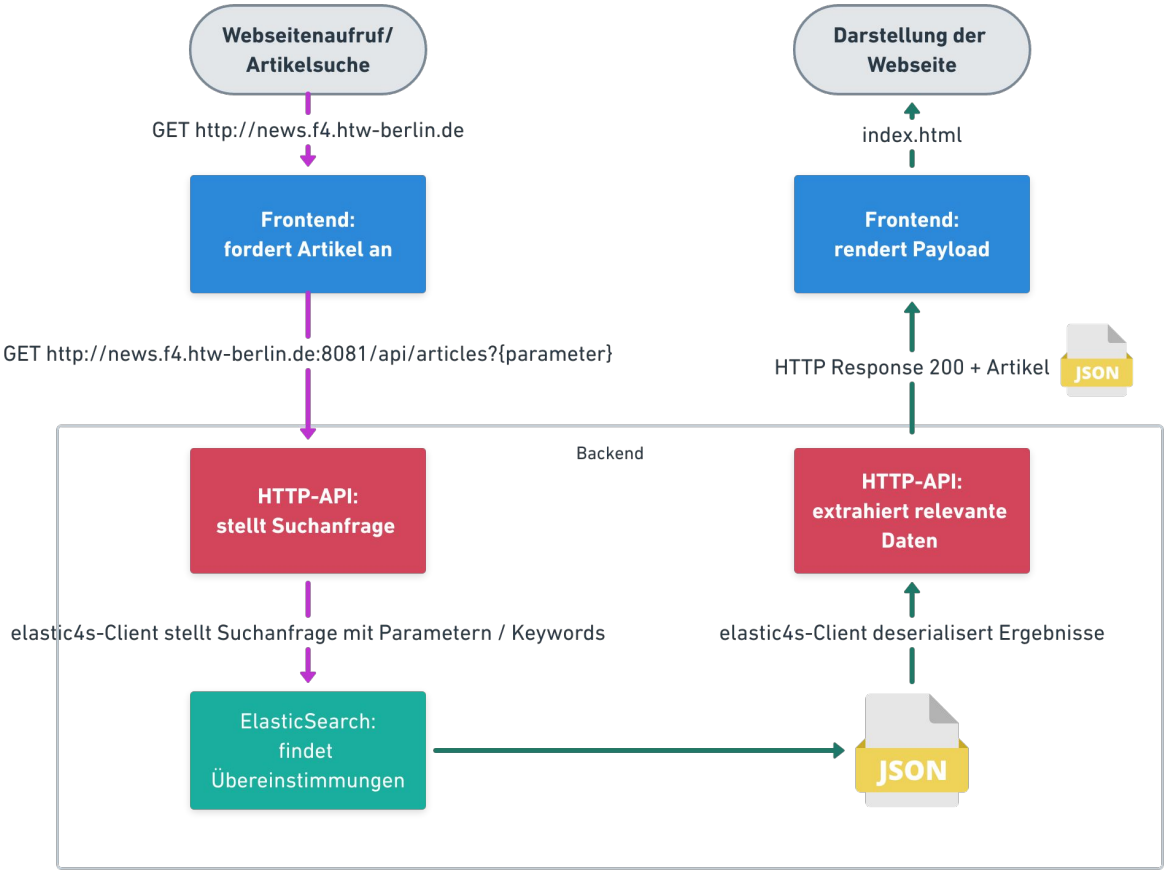
- Programmiersprache: JavaScript
- React: Bibliothek zum Erstellen von UIs und UI-Komponenten
- Material-UI: Framework für UI-Komponenten

- react-vis: Bibliothek zum Erstellen von Charts

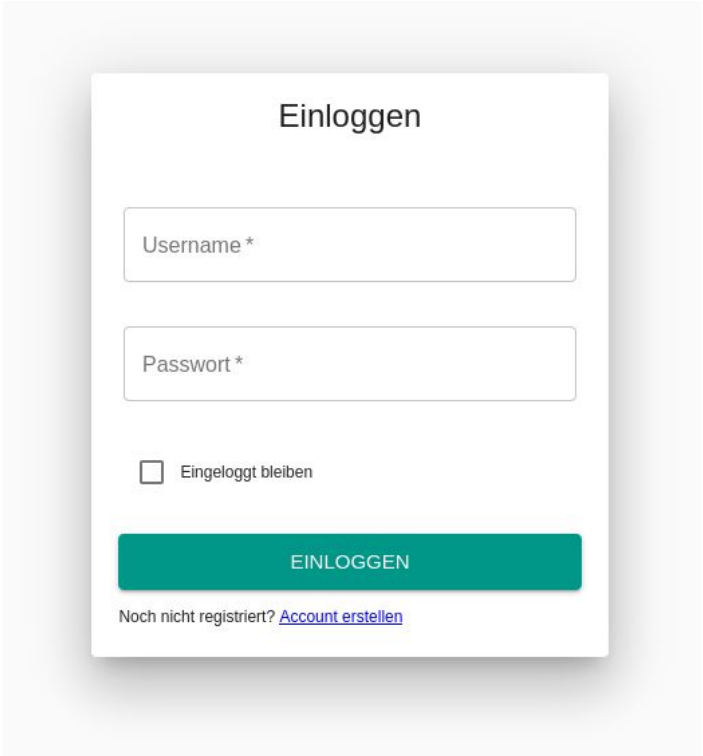


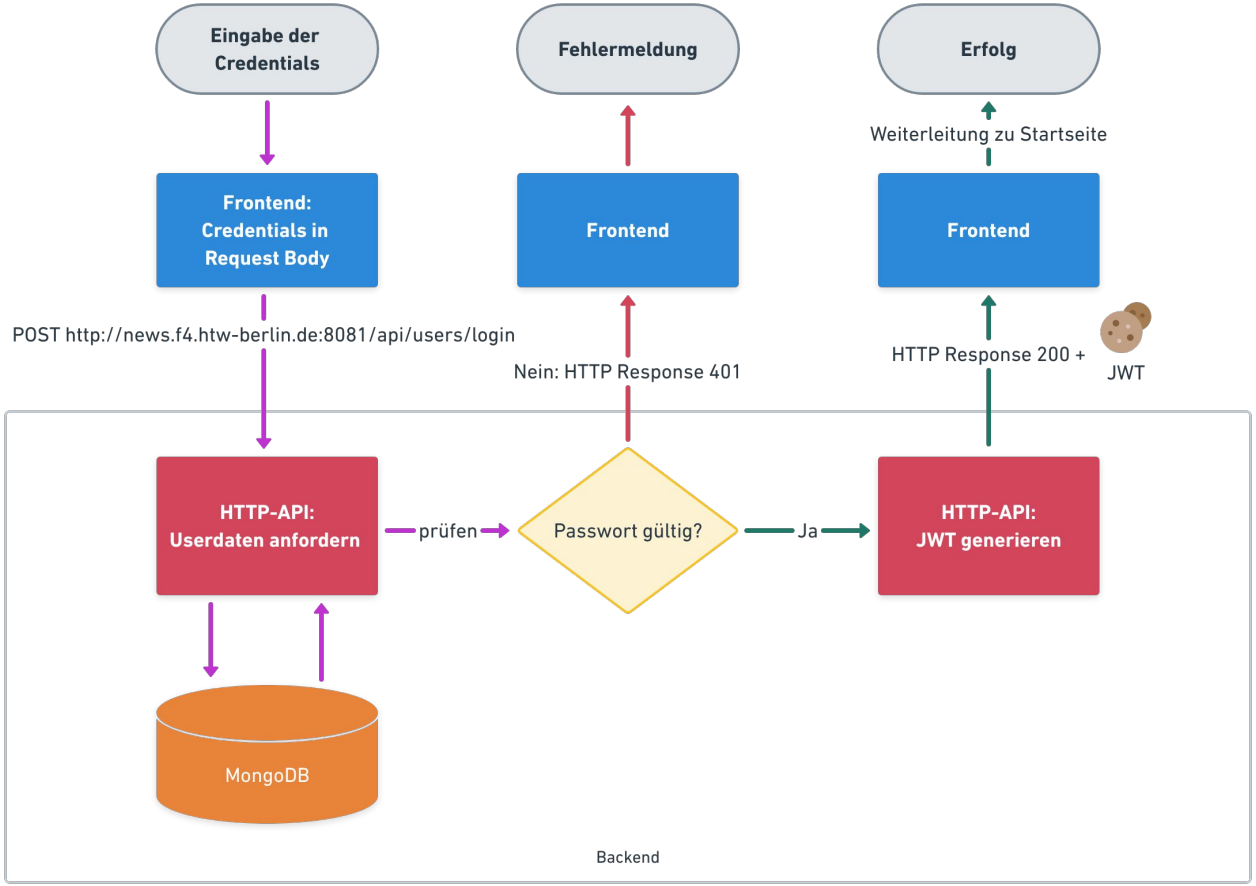
3d. Webanwendung: Artikel-Vorschau




Webseitenaufruf/
Artikelsuche
GET http://news.f4.htw-berlin.de
Frontend:
fordert Artikel an
GET http://news.f4.htw-berlin.de:8081/api/articles?{parameter}
Backend
HTTP-API:
stellt Suchanfrage
elastic4s-Client stellt Suchanfrage mit Parametern / Keywords
ElasticSearch:
findet
Übereinstimmungen
JSON
elastic4s-Client deserialisert Ergebnisse
HTTP-API:
extrahiert relevante
Daten
HTTP Response 200 + Artikel
JSON
Frontend:
rendert Payload
index.html
Darstellung der
Webseite

3d. Webanwendung: Login




Eingabe der Credentials
Fehlermeldung
Erfolg
Weiterleitung zu Startseite
Frontend: Credentials in Request Body
Frontend
Frontend
POST http://news.f4.htw-berlin.de:8081/api/users/login
Nein: HTTP Response 401
HTTP Response 200 +
JWT
HTTP-API: Userdaten anfordern
prüfen
Passwort gültig?
Ja
HTTP-API: JWT generieren
MongoDB
Backend

4. Learnings

- Online Teamwork funktioniert!
- Continuous Integration
- Zeitmangel
- Gute Doku
- Realitätsnah bezüglich “Arbeitswelt”
- Spannendes Projekt

5. Ausblick

- Weitere NLP Analysen
  - Moderne Transformer Models für TextSum
  - Keywords mit TF-IDF im Vergleich zu Yake?
  - Generelle Spark Optimierung
- Continuous Integration Realisieren
- Weitere Seiten für Scraper
- Jenkins/Airflow für Scheduling und Visualisierung des loggings
- MicroServices

6. Ethik

Wir stehlen hier Daten. Es muss erklärt werden dass dies NUR für wissenschaftliche Zwecke ist und die Werkzeuge die wir hier NICHT in der realen Welt anwenden.

Vielen Dank für eure Aufmerksamkeit

Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

www.htw-berlin.de