



Tagesschau Mining

Datenanalyse mit NLP

Inhalt

Ein paar Fragen:

Wie setze ich ein Data Science Projekt auf?

Welche interessanten Insights finden sich?

Wie benutze ich NLP für meine Analysen?



python

Erste Schritte



Datenbeschaffung

Wo kriege ich die Daten her?

Sind die Daten vollständig?

Sind die Daten vertrauenswürdig?

Die letzte Sendung

23.09.2021 12:00 Uhr
tagesschau

Steigende Gaspreise in Europa treffen Industrie und Verbraucher, Endspurt im Wahlkampf um das Berliner Abgeordnetenhaus, Sozialverbände warnen vor sozialen Folgen gestiegener Preise für gesunde Lebensmittel, Die Börse, Britischer Premier Johnson wirbt auf UN-Vollversammlung für mehr Klimaschutz, Humanitäre Lage in Afghanistan bleibt dramatisch, UN-Gipfel berät über weltweite Ernährungssicherheit, Verhandlungen über eine Polizeireform im US-Kongress scheitern vorerst, Vulkan auf Kanareninsel La Palma spuckt weiter Lava und Asche, Reeperbahnfestival startet in Hamburg, Das Wetter

Übersicht der letzten Sendungen

23.09.2021 00:06 Uhr
nachtmagazin

Extremwetterkongress in Hamburg, China verzichtet auf Kohlekraftausbau im Ausland, Diskussion um Lohnfortzahlung für Ungeimpfte bei Corona-Quarantäne, Bundeskanzlerin Merkel und Verteidigungsministerin Kramp-Karrenbauer würdigen Evakuierungseinsatz der Bundeswehr in Afghanistan, Ausstellungseröffnung im Berliner Humboldt-Forum: Bundespräsident Steinmeier erinnert an Unrecht der Kolonialzeit, Das Wetter

22.09.2021 22:45 Uhr
tagesschau vor 20 Jahren

22.09.2021 22:15 Uhr
tagesthemen

Diskussion um Lohnfortzahlung für Ungeimpfte bei Corona-Quarantäne, Die Meinung, Bundeskanzlerin Merkel und Verteidigungsministerin Kramp-Karrenbauer würdigen Afghanistan-Einsatz der Bundeswehr. Die Lage einer Familie auf der Flucht



Was gibt es zu beachten?

robots.txt überprüfen
(tagesschau.de/robots.txt)

Serverauslastung beachten
(ggf. sleep zwischen Requests einbauen)

Datenbeschaffung

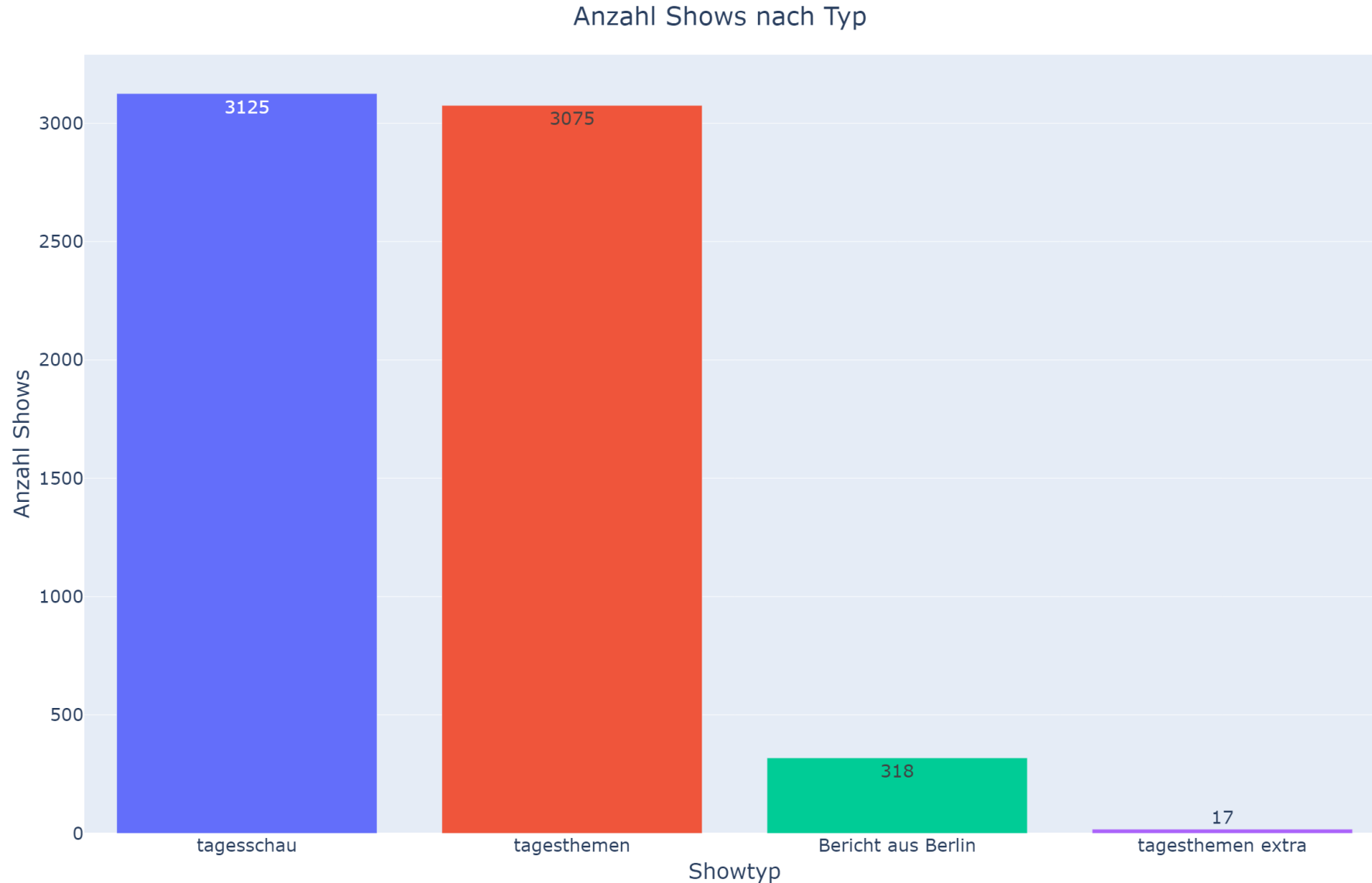
- Aus dem Tagesschau Archiv
- 3000 Requests
- Meta Daten aus dem HTML



Wie sieht der Datensatz aus?

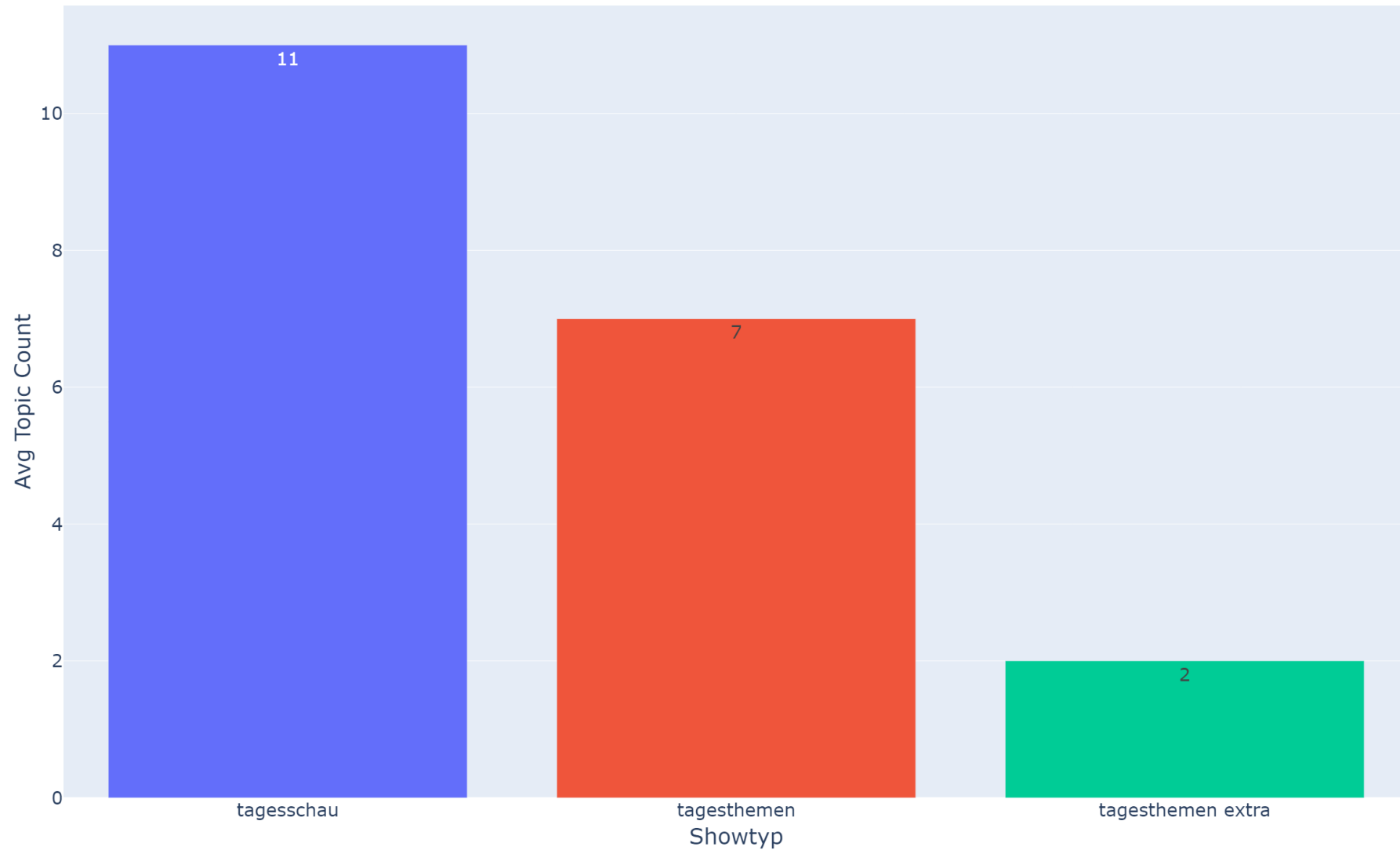
	date_and_time	title	topic
12424	2021-09-22 09:00:00	tagesschau	Gesundheitsminister beraten über künftige Corona-Regeln für Ungeimpfte
12424	2021-09-22 09:00:00	tagesschau	Vor Bundestagswahl: Bundeswahlleiter Thiel ruft zur Stimmabgabe auf
12424	2021-09-22 09:00:00	tagesschau	Taliban in Afghanistan fordern Rederecht bei UN-Generalversammlung in New York
12424	2021-09-22 09:00:00	tagesschau	Chinas Staatschef Xi Jinping betont bei UN-Generaldebatte Wichtigkeit internationaler Zusammena...
12424	2021-09-22 09:00:00	tagesschau	Nach Angriff in Idar-Oberstein: GdP warnt vor einer Radikalisierung der Corona-Leugner-Szene
12424	2021-09-22 09:00:00	tagesschau	Lavaströme nach Vulkanausbruch auf La Palma richten immer größere Schäden an
12424	2021-09-22 09:00:00	tagesschau	Das Wetter

Wie sieht der Datensatz aus?



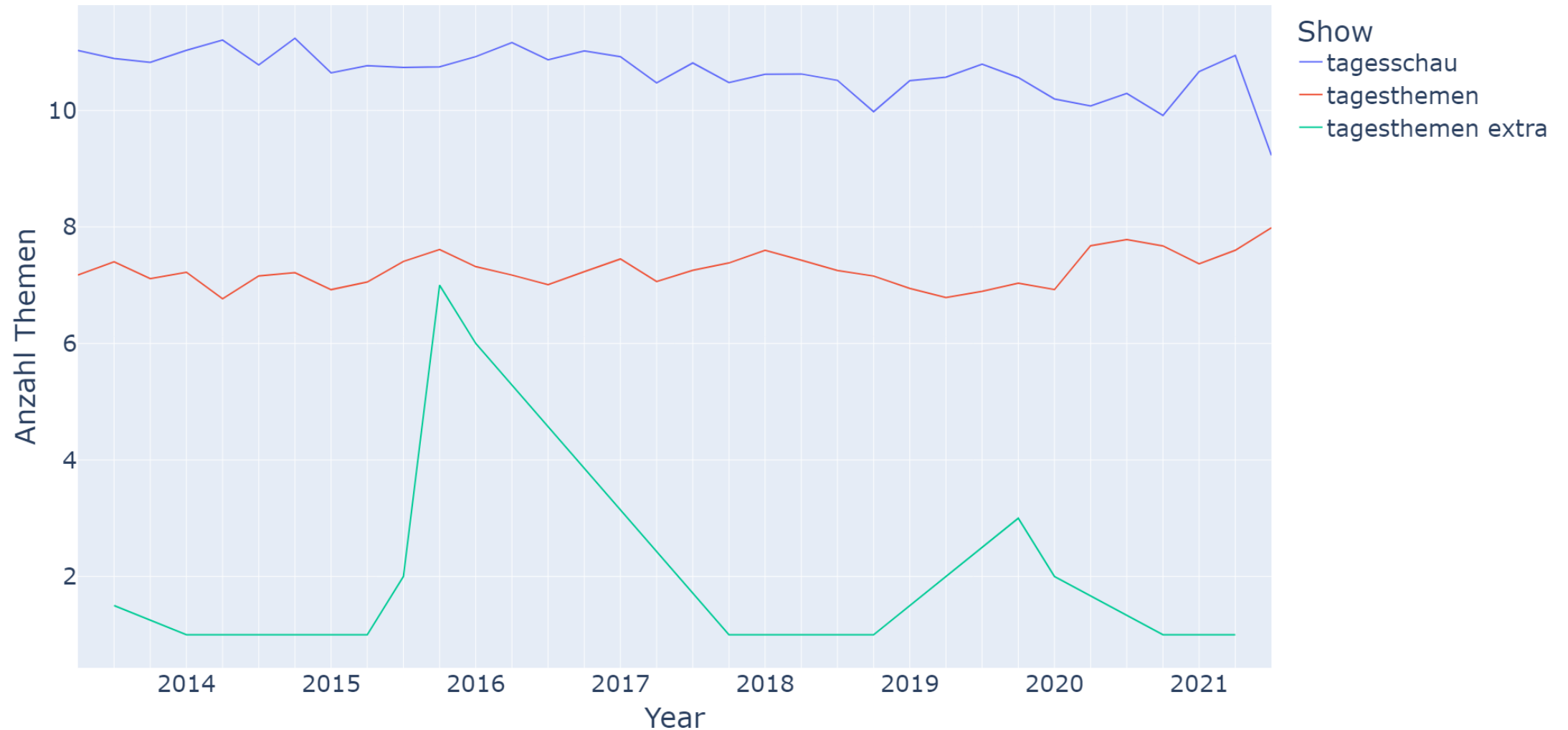
Insights

Durchschnittliche Themen pro Sendung



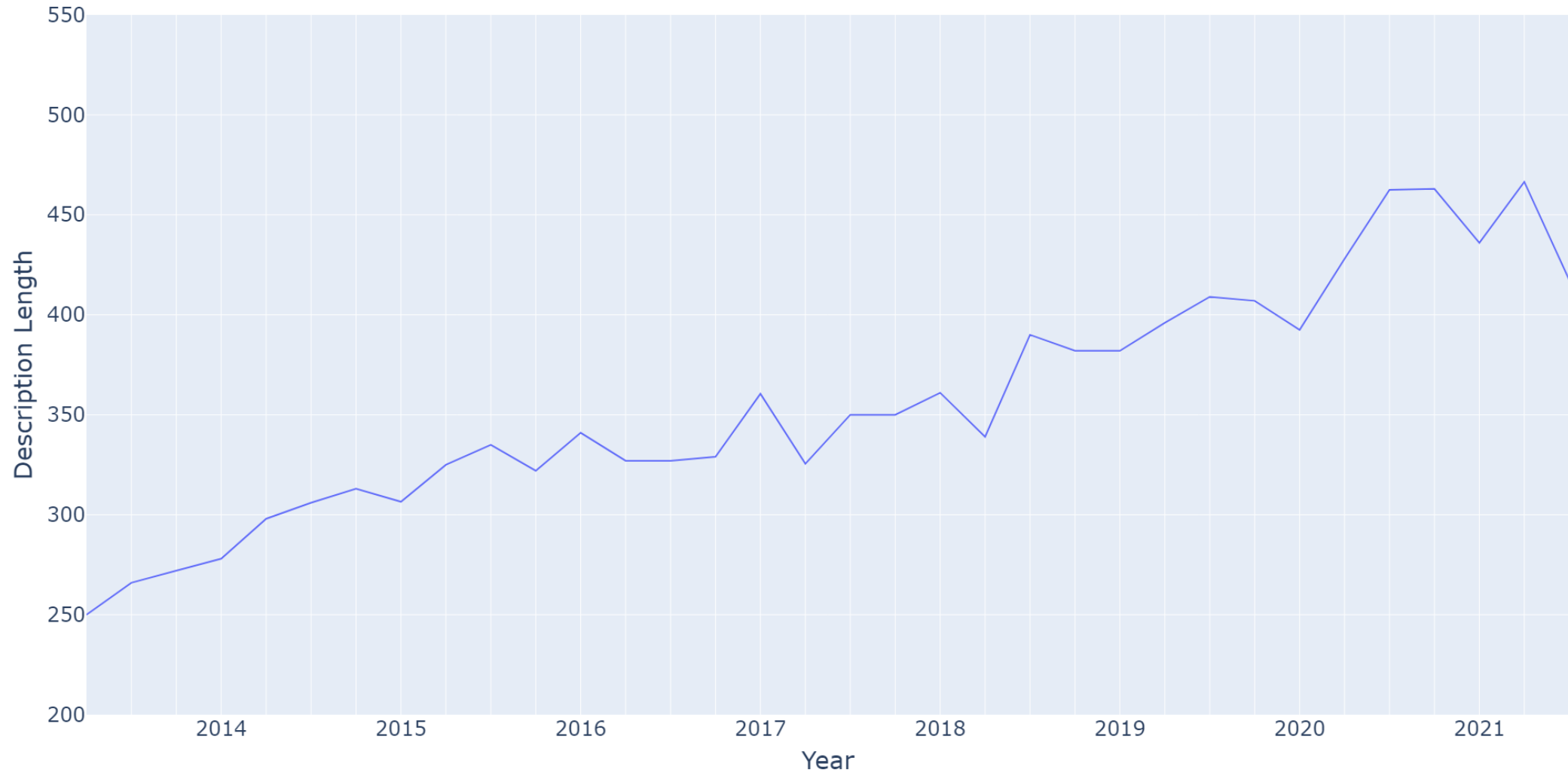
Insights

Durchschnittliche Themen pro Sendung



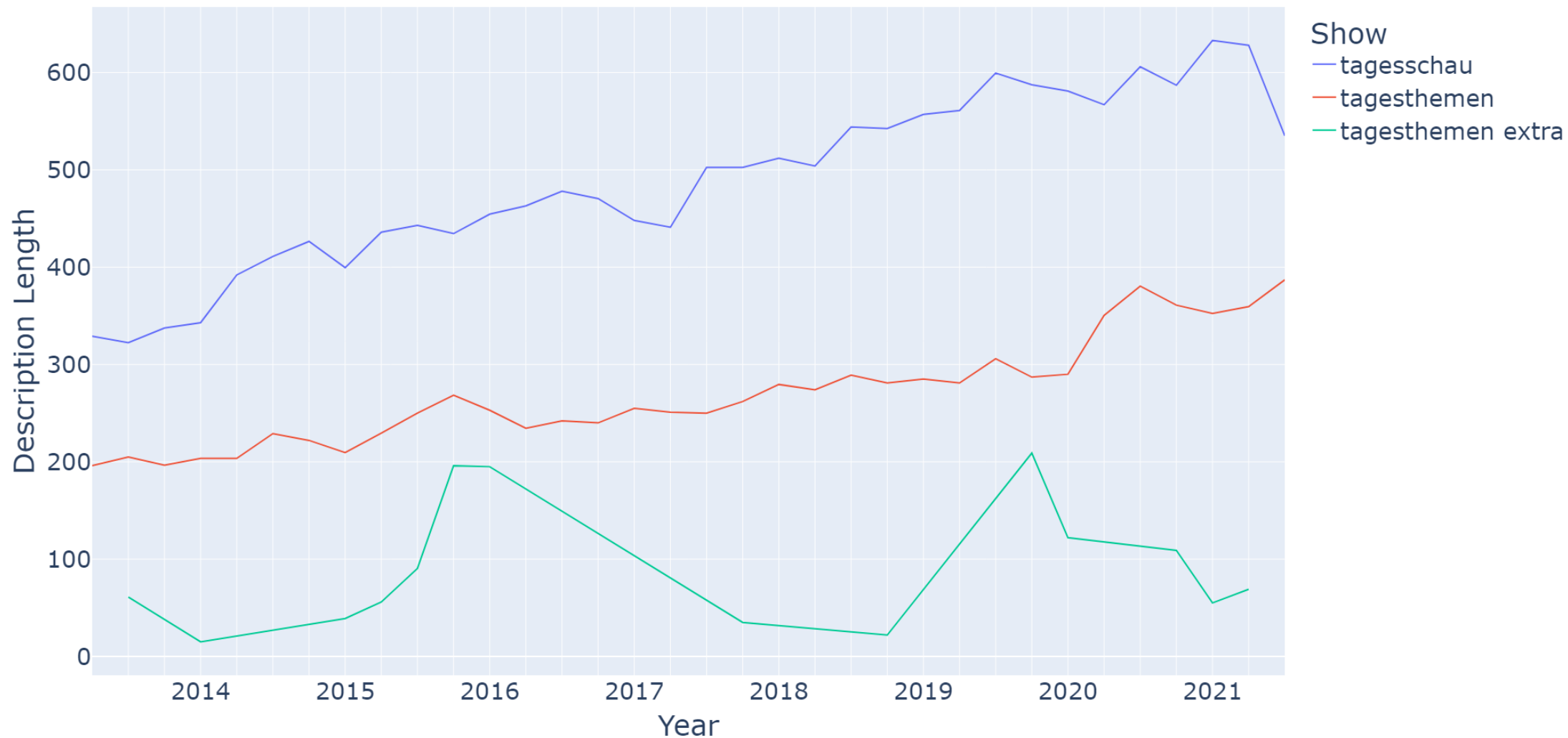
Insights

Median Beschreibungslänge nach Quartal



Insights

Median Beschreibungslänge nach Quartal



Natural Language Processing in Python

```
from transformers import pipeline
import torch

classifier = pipeline(
    "zero-shot-classification",
    model="Sahajtomar/German_Zeroshot",
    device=torch.cuda.current_device()
)

categories = [
    "Politik",
    "Wirtschaft",
    "Sport",
    "Naturkatastrophe",
    "Kunst und Kultur",
    "Terrorismus",
]
```

Natural Language Processing in Python

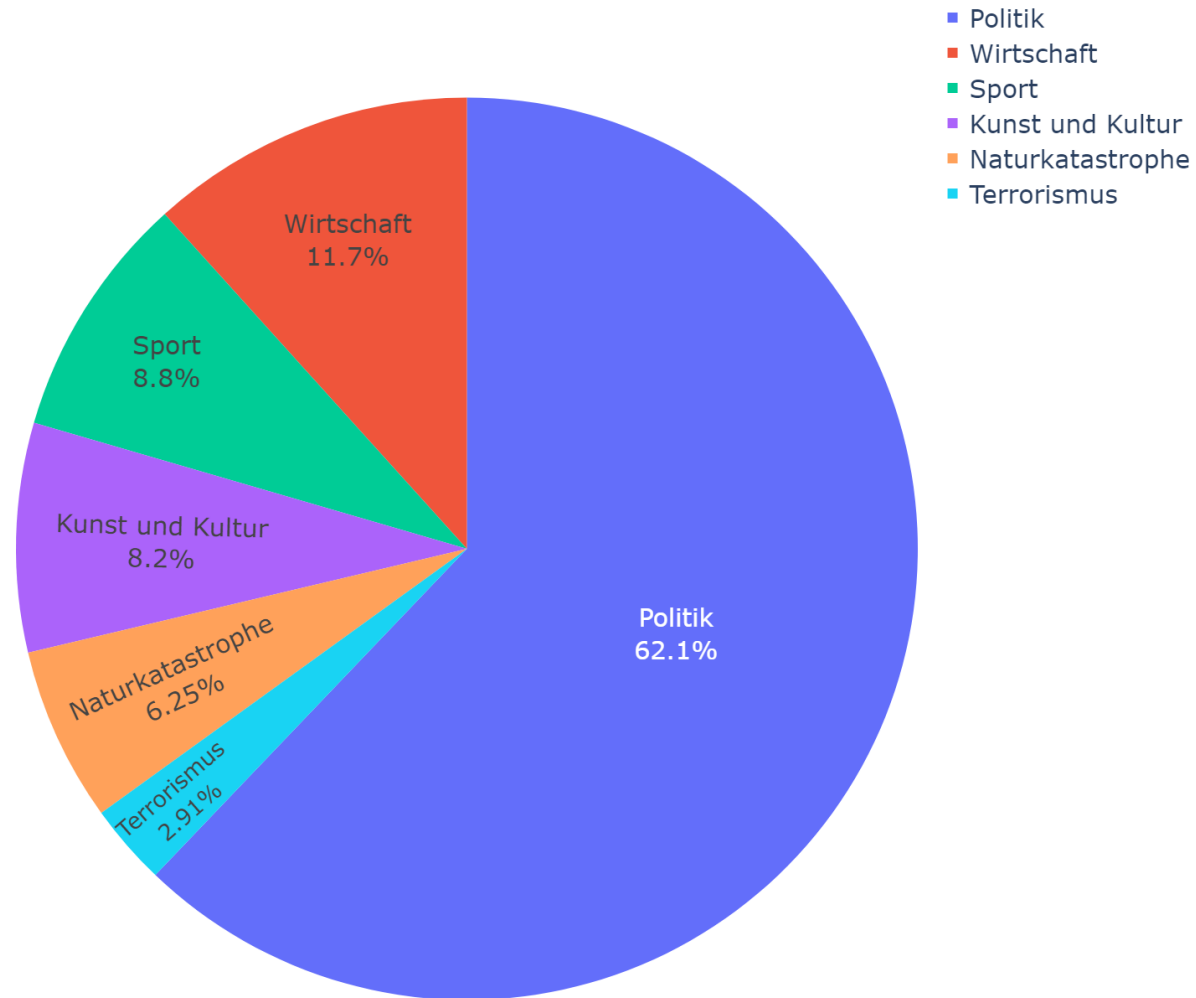
```
output = classifier(  
    "Biathlon-WM: Frauen-Staffel holt Gold",  
    candidate_labels = categories  
)  
list(zip(output["labels"], output["scores"]))
```

✓ 0.6s

```
[('Sport', 0.6599025726318359),  
 ('Wirtschaft', 0.11006765812635422),  
 ('Politik', 0.07161978632211685),  
 ('Naturkatastrophe', 0.059991877526044846),  
 ('Kunst und Kultur', 0.059368737041950226),  
 ('Terrorismus', 0.03904937580227852)]
```

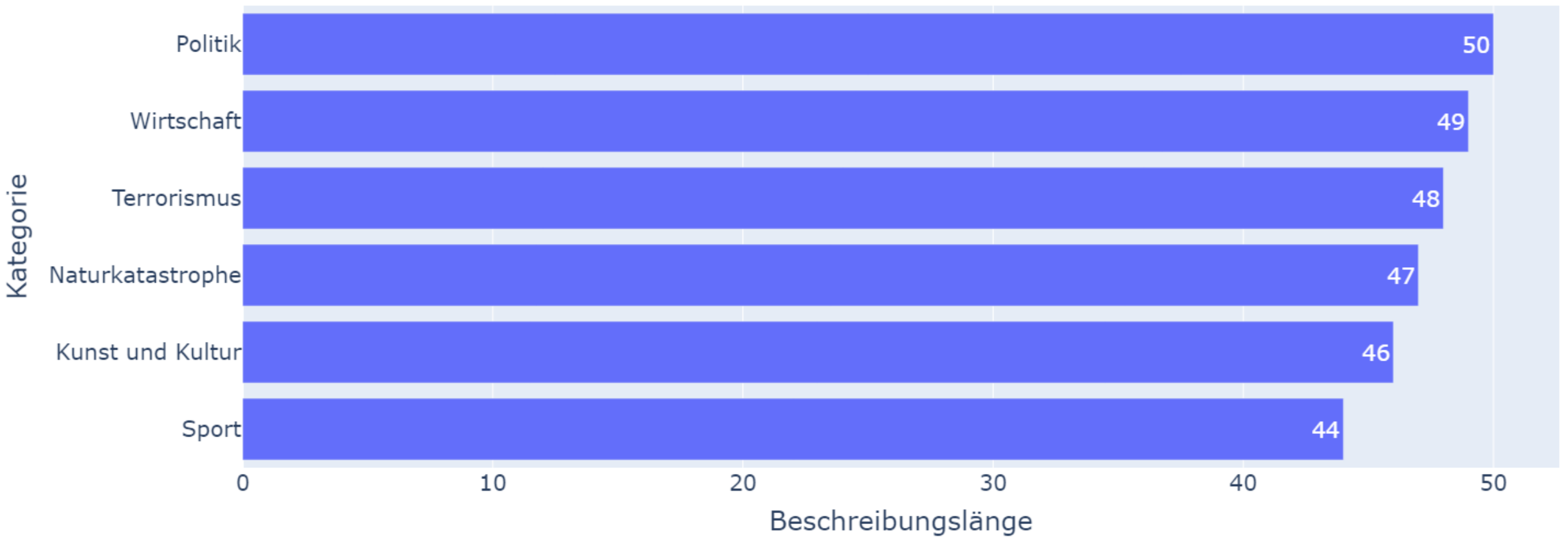
Insights

Meldungen nach Kategorie



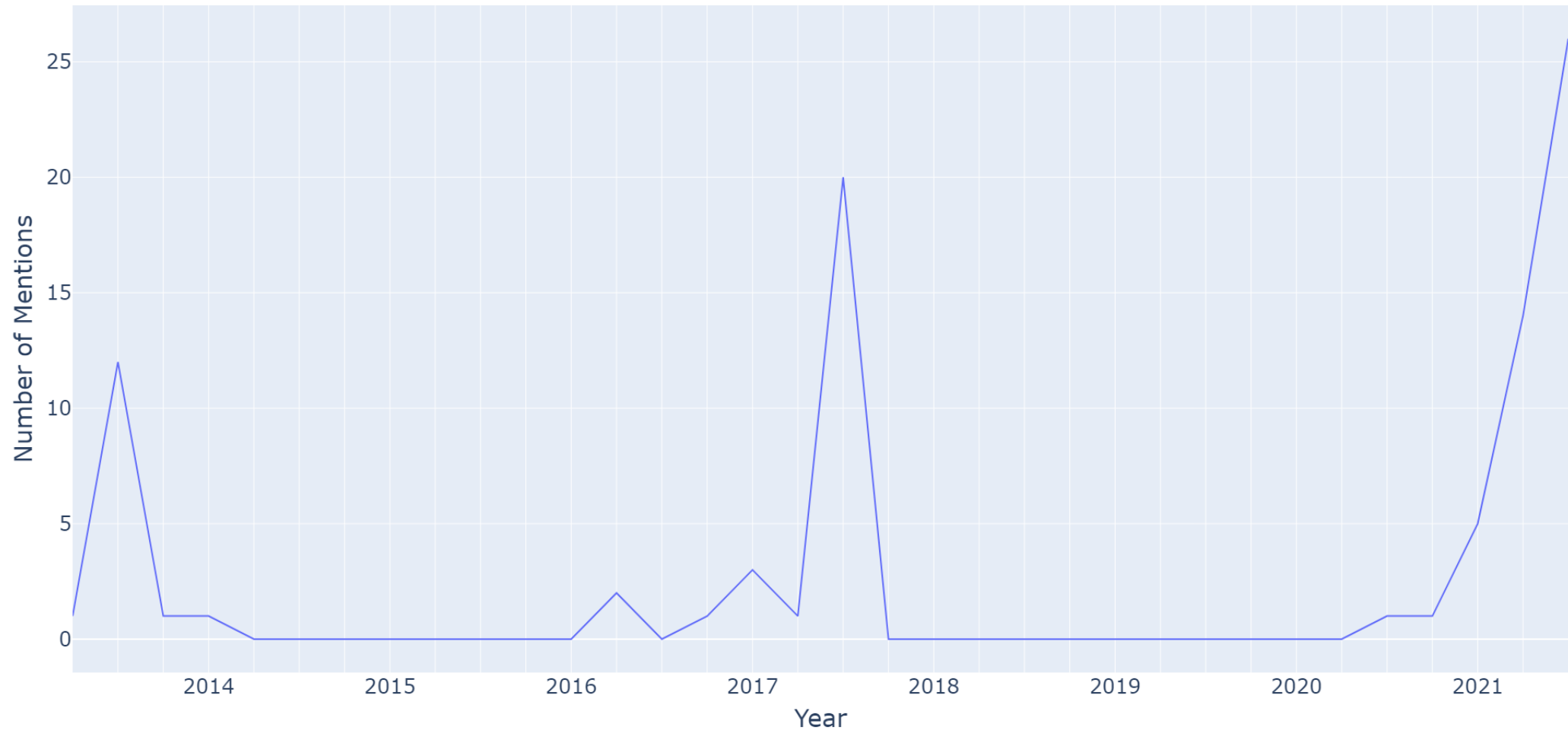
Insights

Beschreibungslänge nach Kategorie



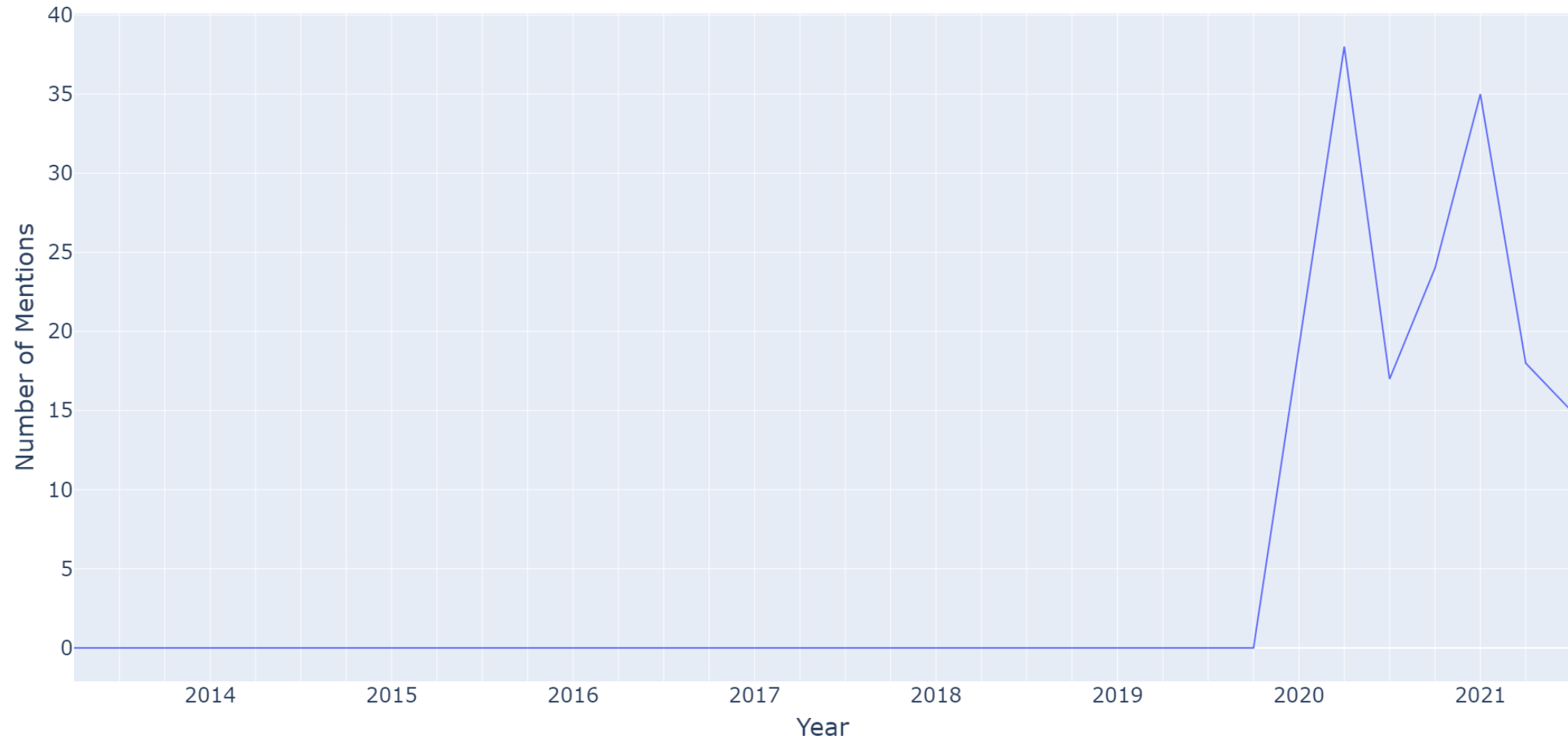
Insights

Mentions of word "Bundestagswahl" per quarter



Insights

Mentions of word "Pandemie" per quarter





Fragen?

✉ johnny.kessler@studium.uni-hamburg.de

🌐 <https://www.linkedin.com/in/johnny-kessler/>

🐙 github.com/jhKessler/Tagesschau-Analysis