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?

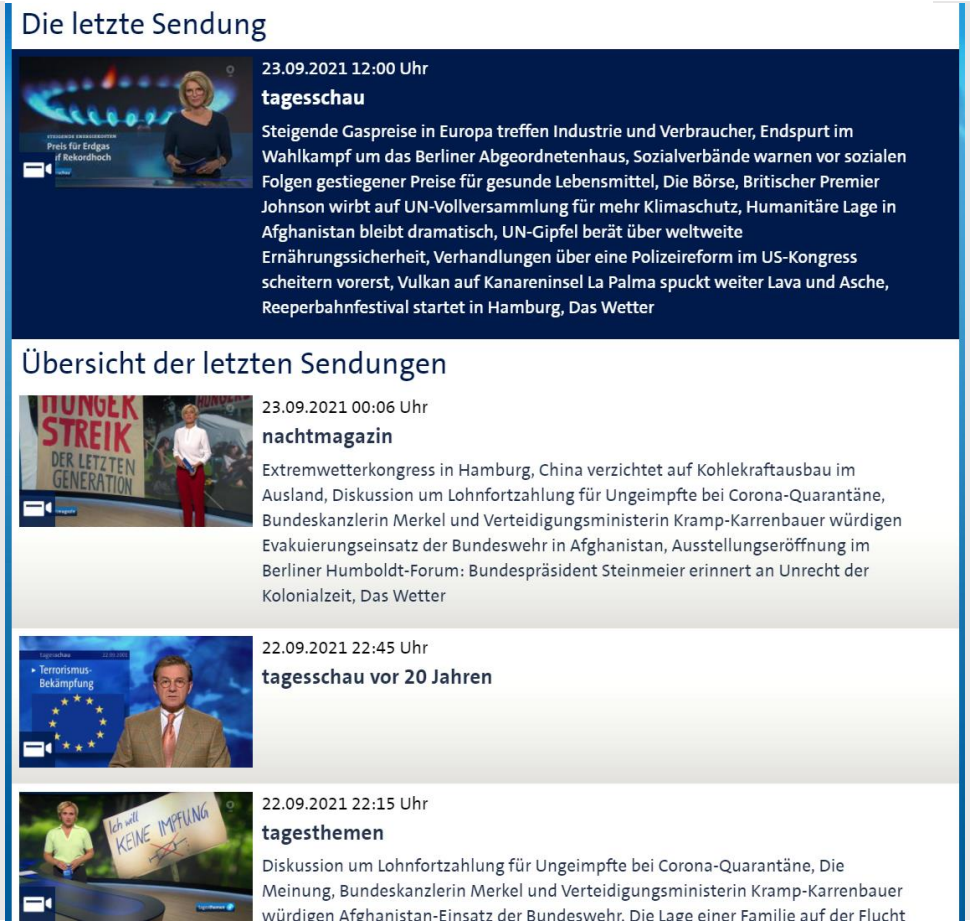
Erste Schritte

Datenbeschaffung

Wo kriege ich die Daten her?

Sind die Daten vollständig?

Sind die Daten vertrauenswürdig?



Was gibt es zu beachten?

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

Serverauslastung beachten (ggf. sleep zwischen Requests einbauen)

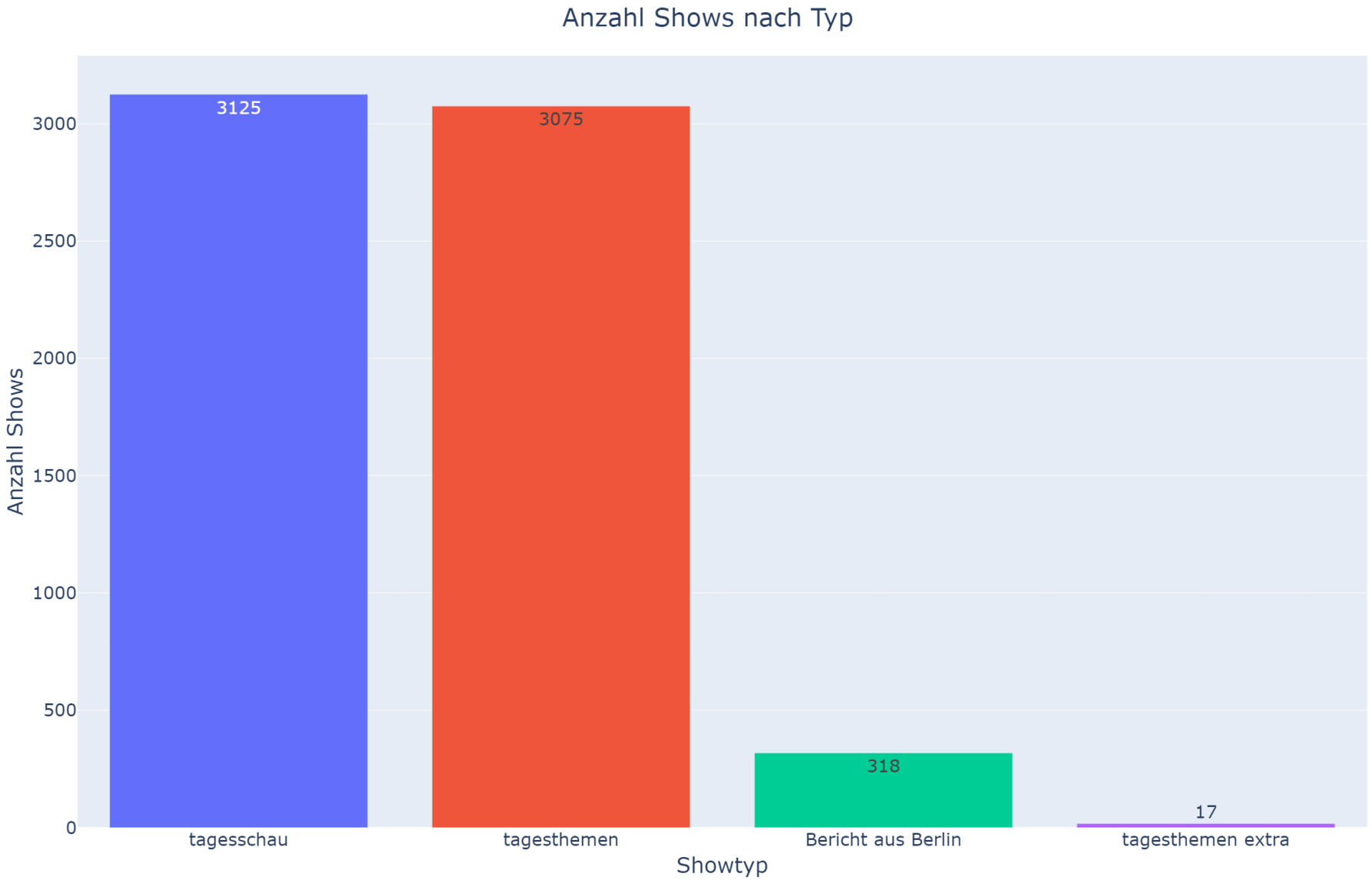
Datenbeschaffung

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

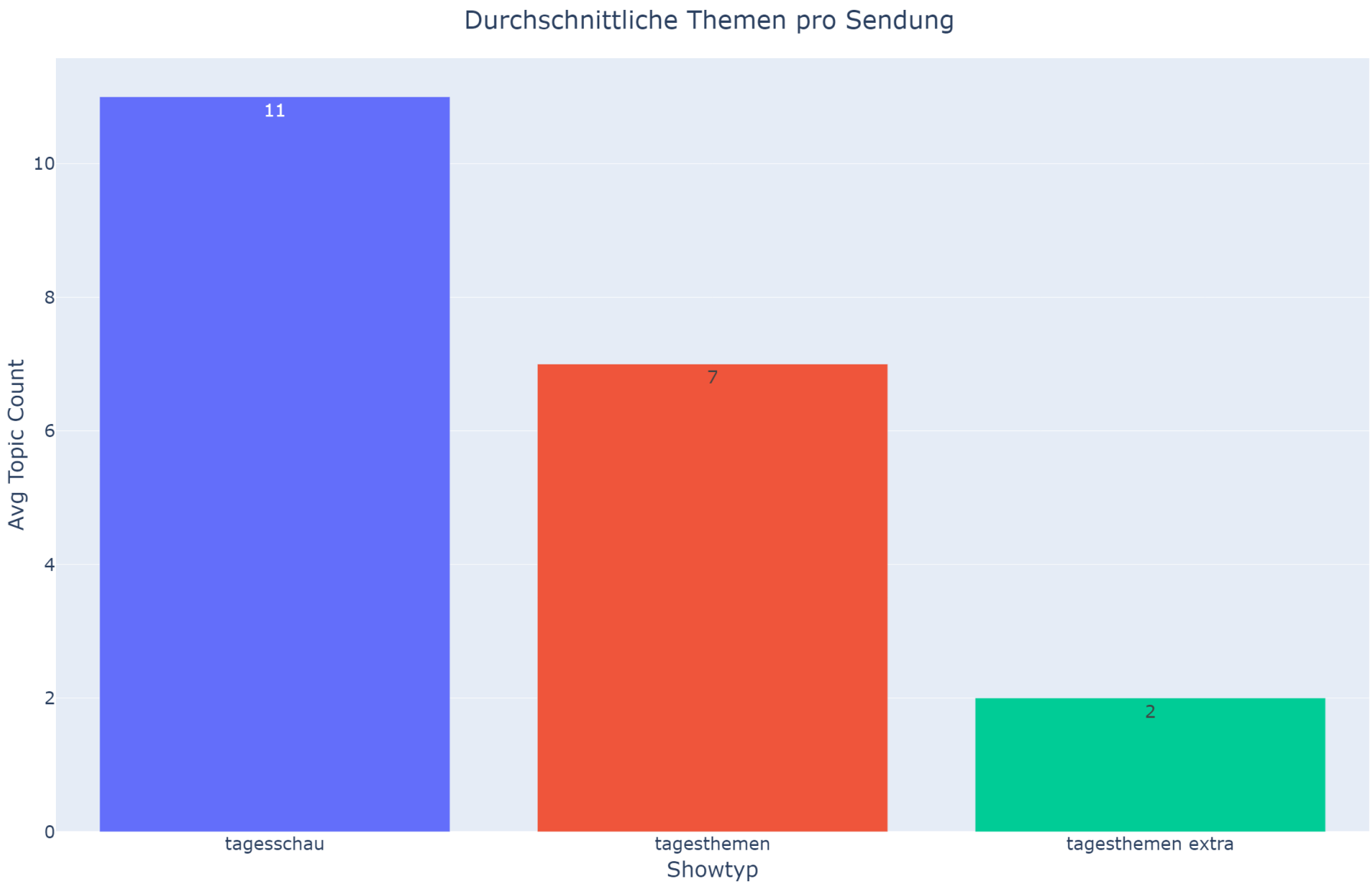
Wie sieht der Datensatz aus?

	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 Vulkanaubruch auf La Palma richten immer größere Schäden an
12424	2021-09-22 09:00:00	tagesschau	Das Wetter

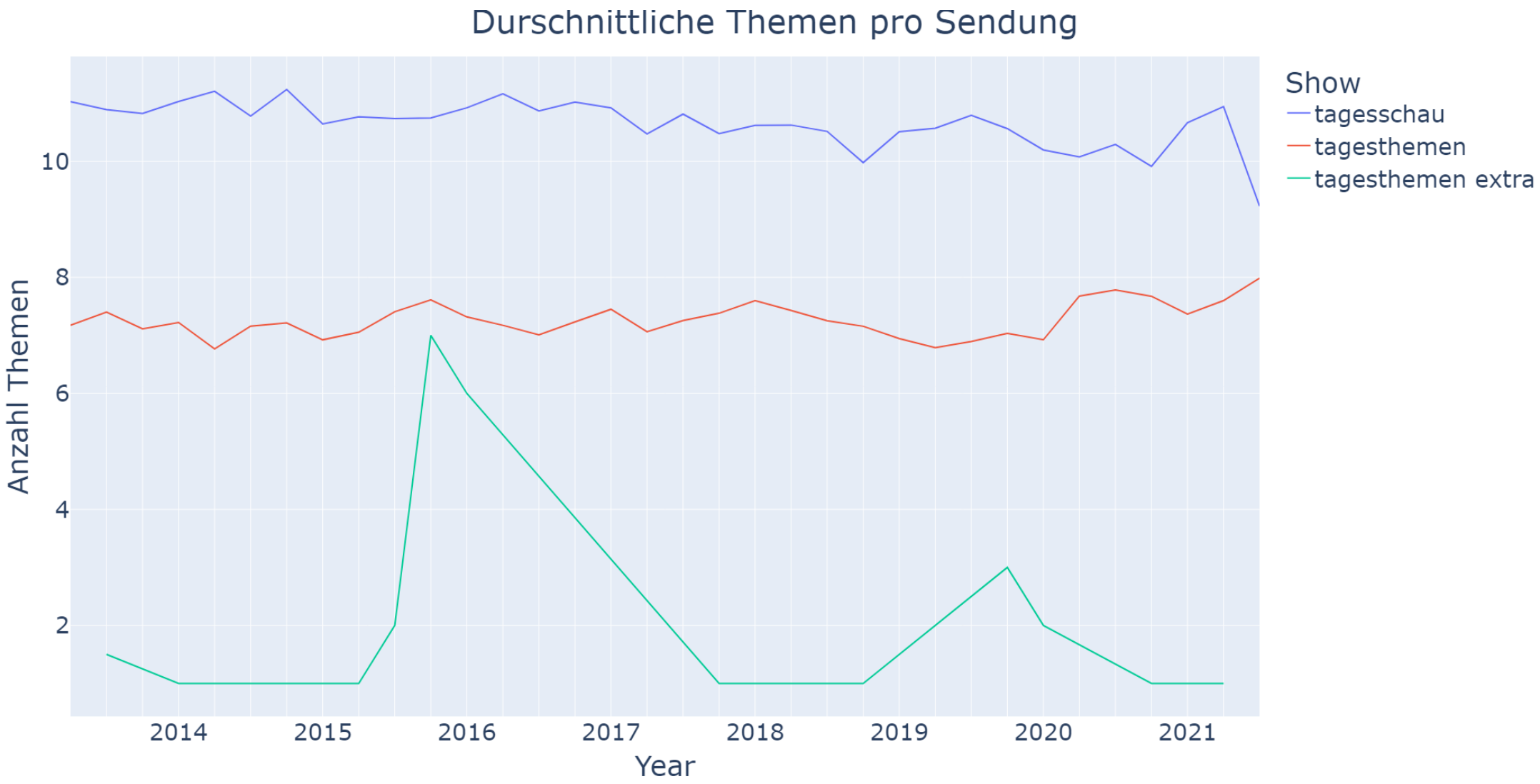
Wie sieht der Datensatz aus?



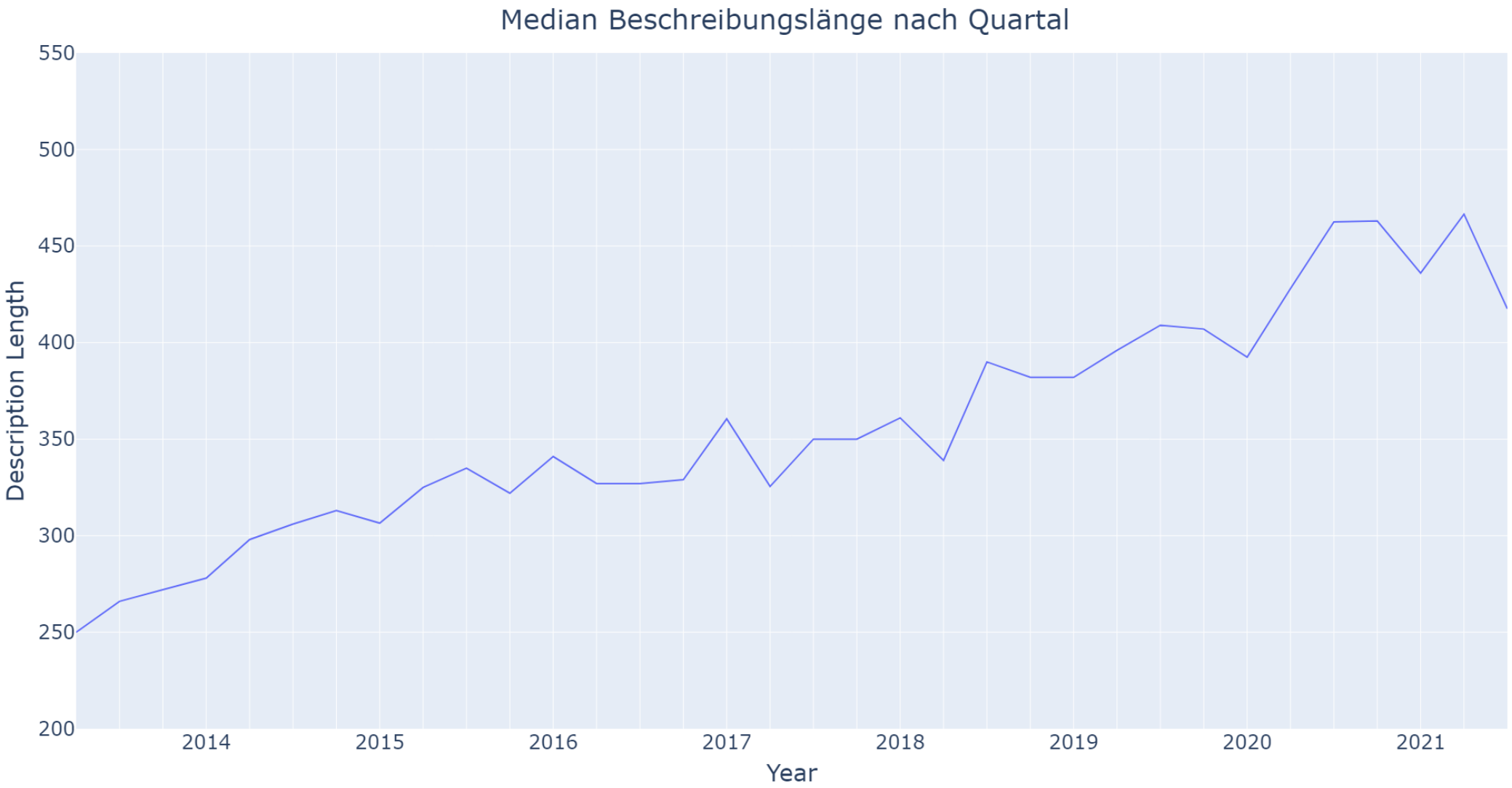
Insights


Durchschnittliche Themen pro Sendung
Avg Topic Count
0
2
4
6
8
10
11
7
2
tagesschau
tagesthemen
tagesthemen extra
Showtyp

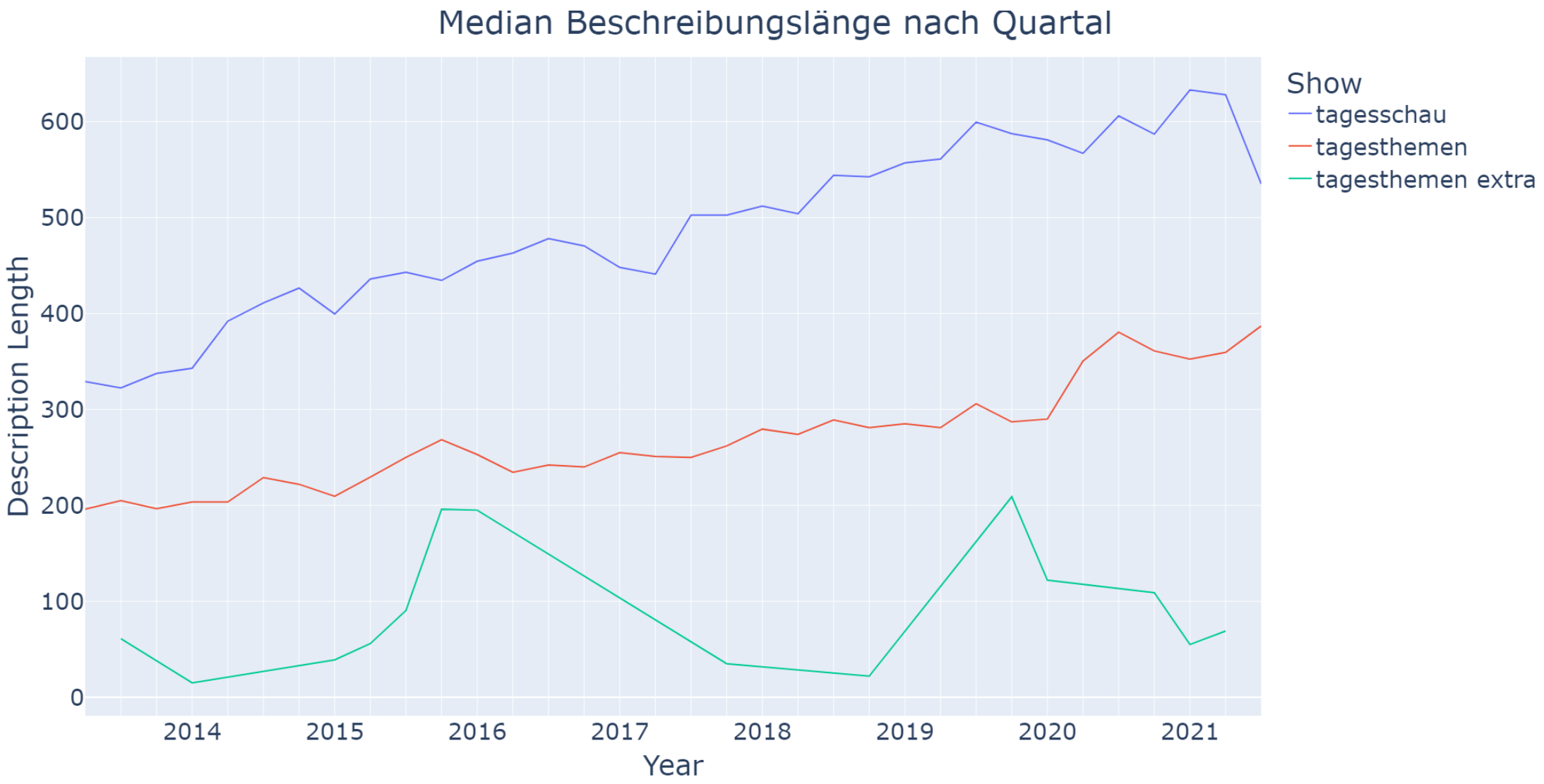
Insights


Durschnittliche Themen pro Sendung
Anzahl Themen
10
8
6
4
2
2014
2015
2016
2017
2018
2019
2020
2021
Year
Show
tagesschau
tagesthemen
tagesthemen extra

Insights


Median Beschreibungslänge nach Quartal
Description Length
550
500
450
400
350
300
250
200
2014
2015
2016
2017
2018
2019
2020
2021
Year

Insights


Median Beschreibungslänge nach Quartal
Description Length
600
500
400
300
200
100
0
2014
2015
2016
2017
2018
2019
2020
2021
Year
Show
tagesschau
tagesthemen
tagesthemen extra

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",
]
```

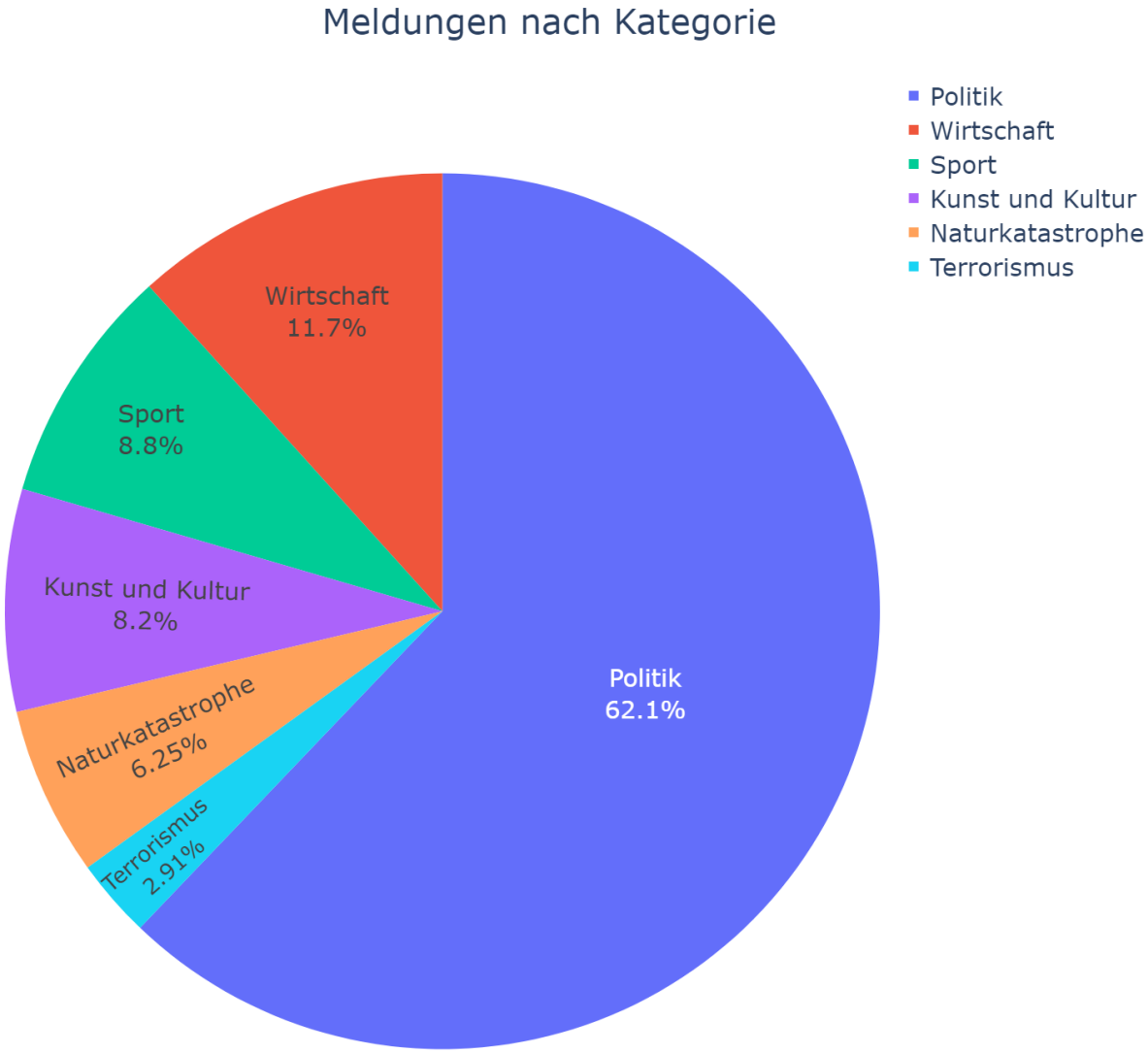
Natural Language Processing in Python

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

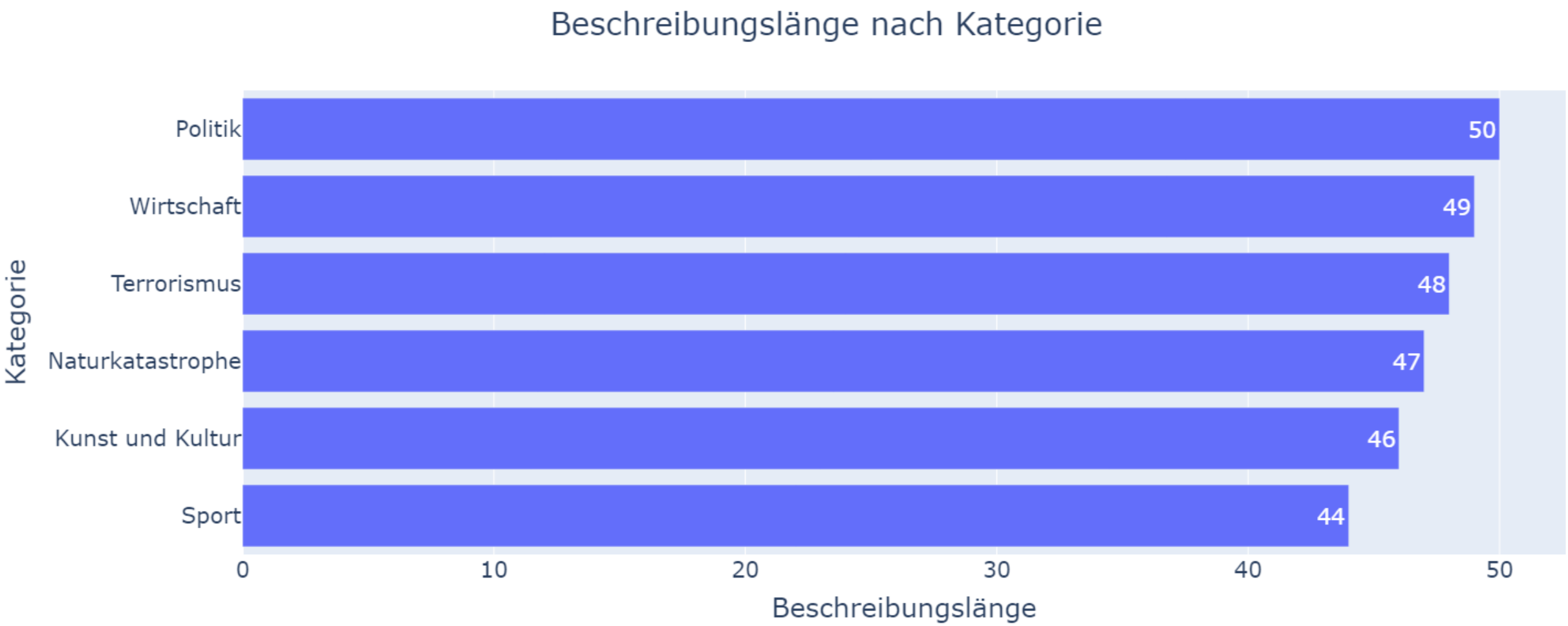
✓ 0.6s

```
[('Sport', 0.6599025726318359),
 ('Wirtschaft', 0.11006765812635422),
 ('Politik', 0.07161978632211685),
 ('Naturkatastrophe', 0.05999187752604484 6),
 ('Kunst und Kultur', 0.05936873704195022 6),
 ('Terrorismus', 0.03904937580227852)]
```

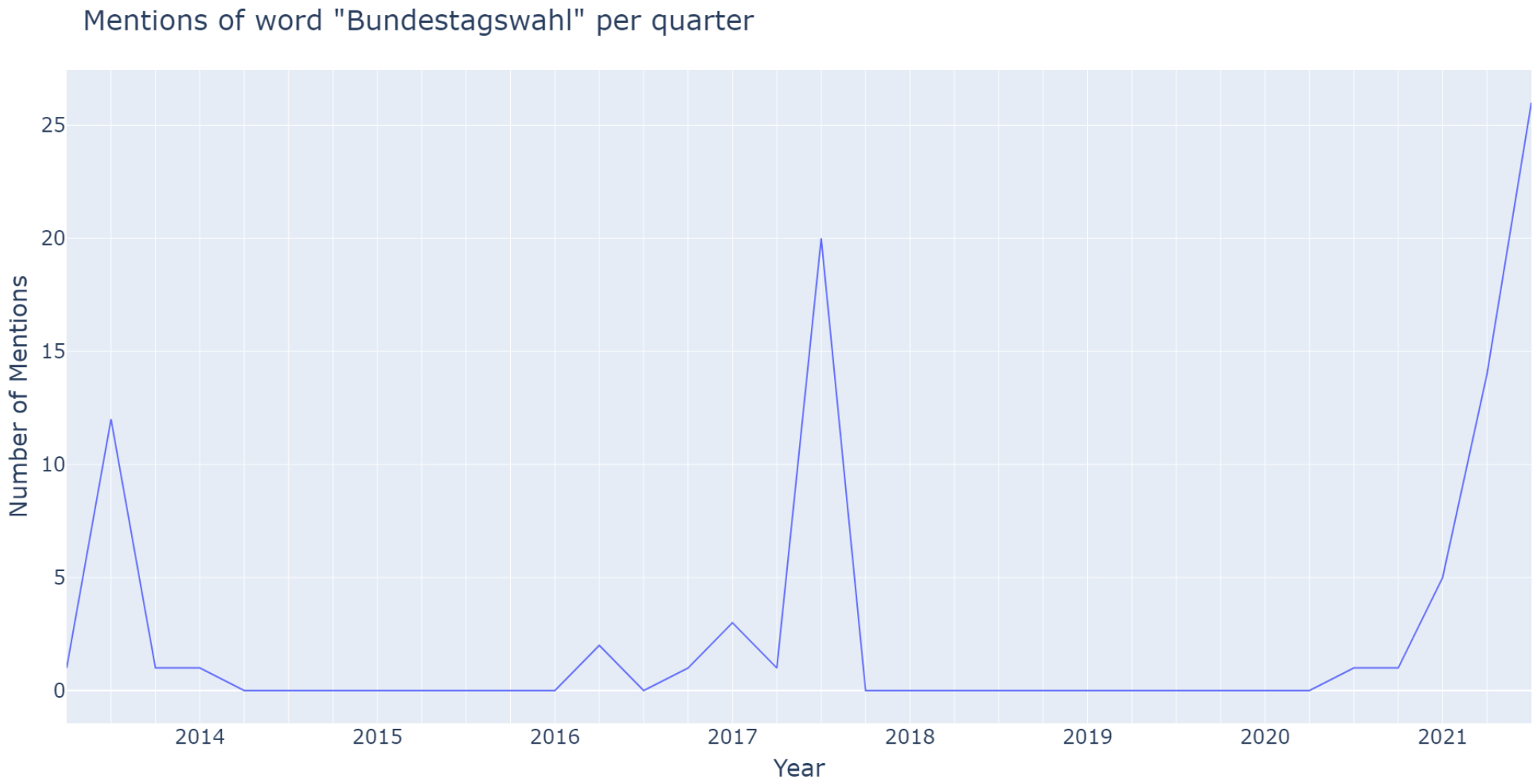
Insights


Meldungen nach Kategorie
Politik
Wirtschaft
Sport
Kunst und Kultur
Naturkatastrophe
Terrorismus
Wirtschaft
11.7%
Sport
8.8%
Kunst und Kultur
8.2%
Naturkatastrophe
6.25%
Terrorismus
2.91%
Politik
62.1%

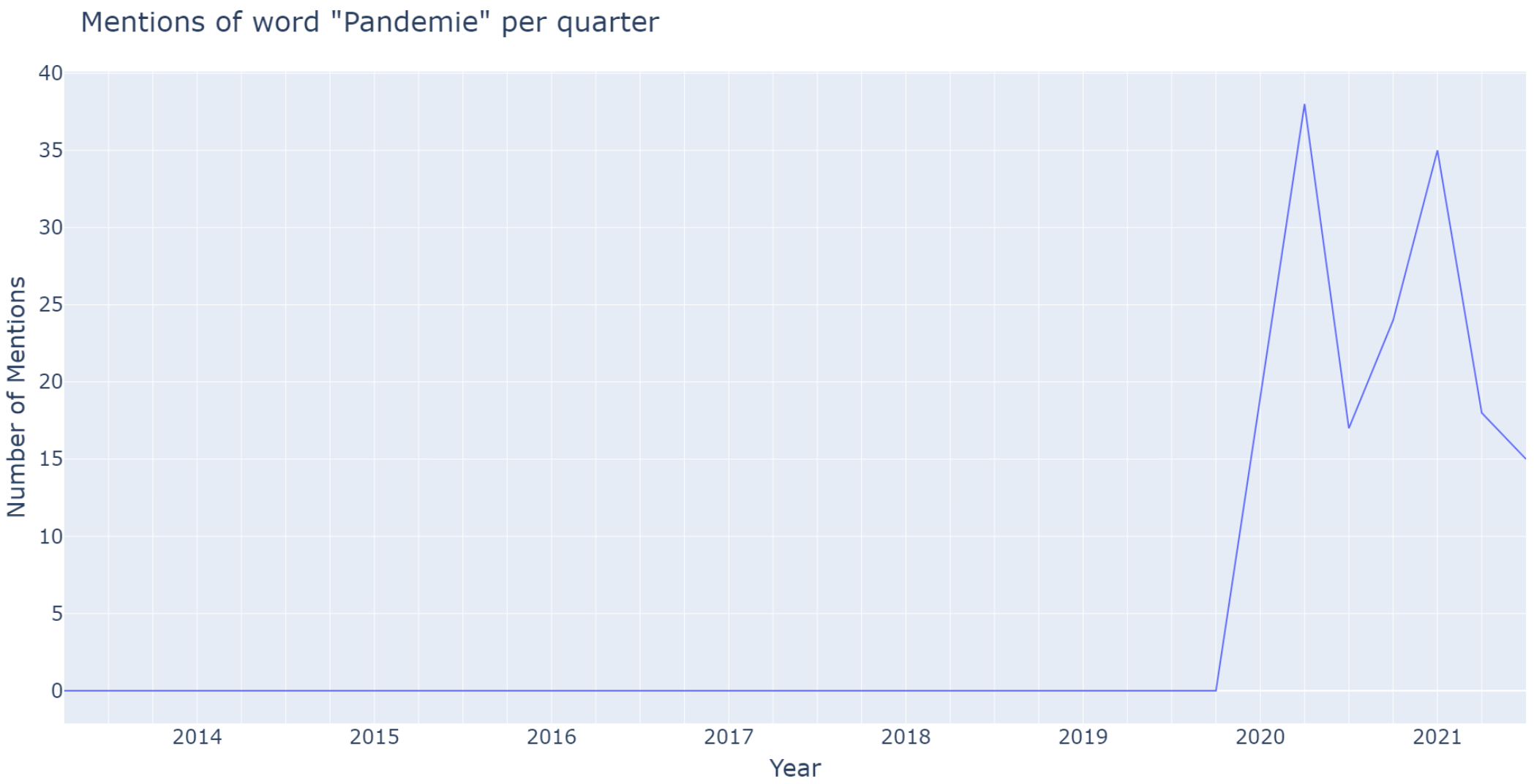
Insights


Beschreibungslänge nach Kategorie
Politik
Wirtschaft
Terrorismus
Naturkatastrophe
Kunst und Kultur
Sport
50
49
48
47
46
44
Kategorie
0
10
20
30
40
50
Beschreibungslänge

Insights


Mentions of word "Bundestagswahl" per quarter
Number of Mentions
0
5
10
15
20
25
2014
2015
2016
2017
2018
2019
2020
2021
Year

Insights


Mentions of word "Pandemie" per quarter
Number of Mentions
40
35
30
25
20
15
10
5
0
2014
2015
2016
2017
2018
2019
2020
2021
Year

Fragen?

johnny.kessler@studium.uni-hamburg.de

https://www.linkedin.com/in/johnny-kessler/

github.com/jhKessler/Tagesschau-Analysis