



www.dinacon.ch
#DINAcon20



B E K B

B C B E





Automatisierung und Scraper

DINAcon 2020
23.10.2020
Stefan Oderbolz

1 Scraper - Was ist das?

Ausgangslage

- Es ist Mitte März, **aktuelle, maschinenlesbare** Daten zu SARS-CoV-2 sind rar
- Im [openZH/covid_19 GitHub Repository](#) entstehen CSVs mit **händisch zusammengetragenen** Zahlen
- Kantone publizieren meist direkt über ihre **Webseiten**

Was ist ein Scraper?

- Engl. *to scrape*: kratzen/schaben
- Idee: Daten von Webseiten oder PDFs zu *kratzen*, und diese dann strukturiert ablegen
- **Ein Scraper ist ein Programm, das Daten aus semi-strukturierten Quellen** holt und zur Verfügung stellt

2 Realisierung der Scraper

Wie sind die Scraper implementiert?

- Das Projekt wurde als Hackathon initiiert
 - Interessierte haben eigene Scraper entwickelt
 - Verschiedene Programmiersprachen (Shell, Python, Ruby)
 - Verschiedene Infrastruktur (eigene Server, morph.io)
- Scraper basieren auf Unix-Philosophie
 - **«Write programs that do one thing and do it well»**
 - Der Output eines Programms ist der Input des nächsten (Pipes)
 - Modularer Aufbau, so dass einzelne Komponenten ausgetauscht bzw. in verschiedenen Programmiersprachen implementiert werden können

Automatisierung mit GitHub Actions

- GitHub Actions ermöglicht beliebige Programme laufen zu lassen
 - Zeitsteuerung (cron)
 - Integriert ins GitHub Ökosystem
- Scraper laufen alle 20 Minuten
 - Ein Scraper für jeden Kanton + FL
 - Neue Daten werden direkt ins Repository committet
 - Fehlermeldungen landen in einem Slack-Channel
 - Die Daten werden validiert und auf Ausreisser geprüft



Automatisierung mit GitHub Actions

Search or jump to... Pull requests Issues Codespaces Marketplace Explore

openZH / covid_19 Unwatch 27 Star 364 Fork 159

<> Code Issues 10 Pull requests 1 Actions Projects Wiki Security Insights Settings

Update dashboard screenshot master GitHub Action Scraper 848b9e7 Artifacts 1 Re-run jobs

Run scrapers on: schedule 1

- run_scraper (AG)
- run_scraper (AI)
- run_scraper (AR)
- run_scraper (BE)
- run_scraper (BL)
- run_scraper (FR)
- run_scraper (GE)
- run_scraper (GL)
- run_scraper (GR)
- run_scraper (JU)
- run_scraper (LU)

run_scraper (NW) succeeded 12 minutes ago in 2m 13s

Search logs

- Run npm ci 2s
- Remove broken apt repos 0s
- Install dependencies 55s
- Scrape new data 3s

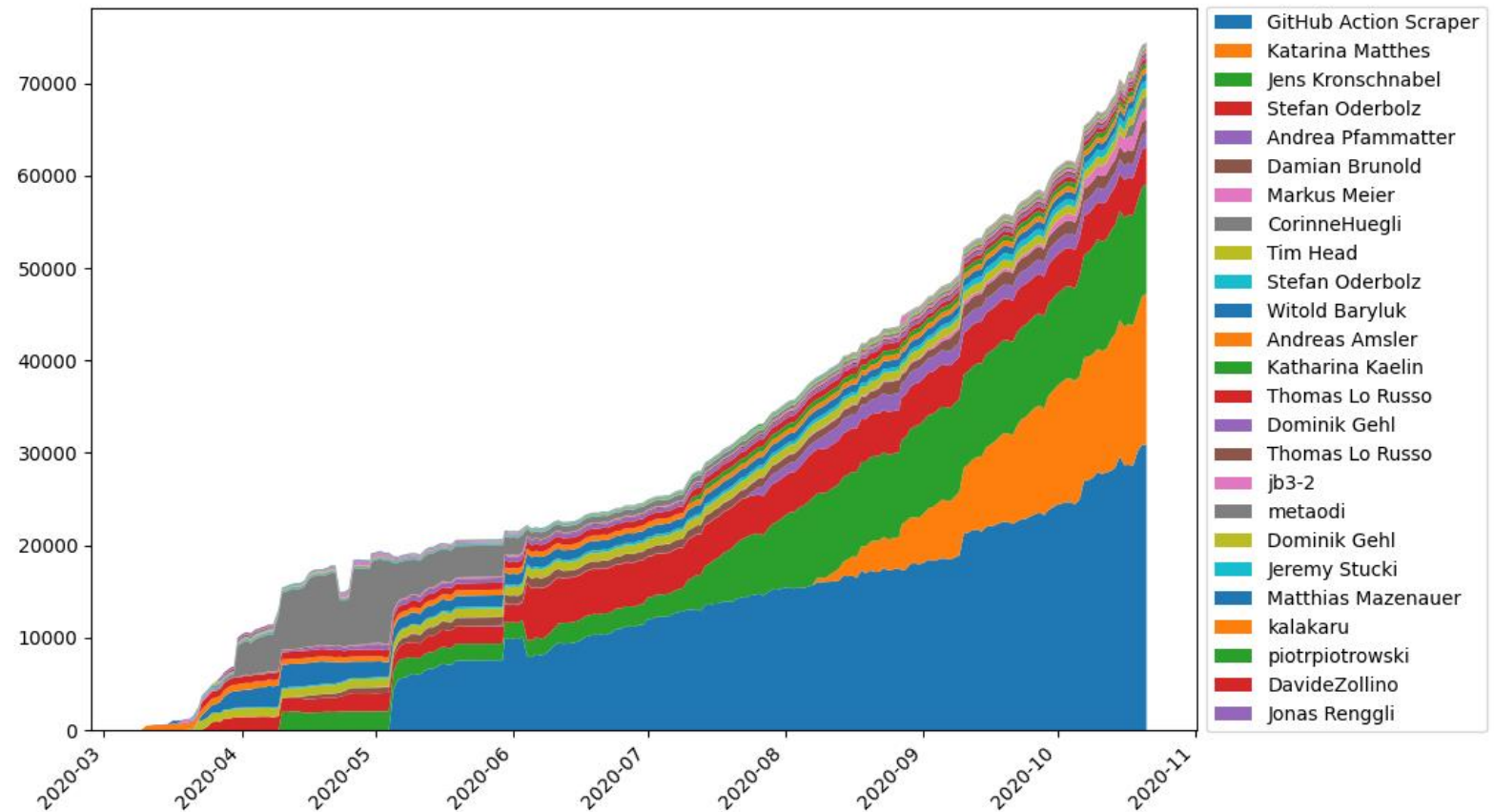
```
1 ▶ Run ./scrapers/run_scraper.sh
8 Populating database from CSV COVID19_Fallzahlen_Kanton_NW_total.csv...
9 Run the scraper...
10 {'date': '2020-03-11', 'time': '', 'abbreviation_canton_and_fl': 'NW', 'ncumul_tested': '', 'ncumul_conf': 4, 'new_hosp': '', 'current_hosp': 2, 'current_icu': '', 'current_vent':
    '', 'ncumul_released': '', 'ncumul_deceased': '', 'source': 'http://www.nw.ch/coronastatistik', 'current_isolated': '', 'current_quarantined': ''}
11 Successfully updated field 'ncumul_conf' of NW: 4 (2020-03-11).
12 Successfully updated field 'current_hosp' of NW: 2 (2020-03-11).
13 Successfully updated field 'source' of NW: http://www.nw.ch/coronastatistik (2020-03-11).
14 {'date': '2020-03-12', 'time': '', 'abbreviation_canton_and_fl': 'NW', 'ncumul_tested': '', 'ncumul_conf': 5, 'new_hosp': '', 'current_hosp': 5, 'current_icu': '', 'current_vent':
    '', 'ncumul_released': '', 'ncumul_deceased': '', 'source': 'http://www.nw.ch/coronastatistik', 'current_isolated': '', 'current_quarantined': ''}
15 Successfully updated field 'ncumul_conf' of NW: 5 (2020-03-12).
16 Successfully updated field 'current_hosp' of NW: 5 (2020-03-12).
17 Successfully updated field 'source' of NW: http://www.nw.ch/coronastatistik (2020-03-12).
18 {'date': '2020-03-13', 'time': '', 'abbreviation_canton_and_fl': 'NW', 'ncumul_tested': '', 'ncumul_conf': 5, 'new_hosp': '', 'current_hosp': 6, 'current_icu': '', 'current_vent':
    '', 'ncumul_released': '', 'ncumul_deceased': '', 'source': 'http://www.nw.ch/coronastatistik', 'current_isolated': '', 'current_quarantined': ''}
19 Successfully updated field 'ncumul_conf' of NW: 5 (2020-03-13).
20 Successfully updated field 'current_hosp' of NW: 6 (2020-03-13).
21 Successfully updated field 'source' of NW: http://www.nw.ch/coronastatistik (2020-03-13).
22 {'date': '2020-03-14', 'time': '', 'abbreviation_canton_and_fl': 'NW', 'ncumul_tested': '', 'ncumul_conf': 5, 'new_hosp': '', 'current_hosp': 6, 'current_icu': '', 'current_vent':
    '', 'ncumul_released': '', 'ncumul_deceased': '', 'source': 'http://www.nw.ch/coronastatistik', 'current_isolated': '', 'current_quarantined': ''}
23 Successfully updated field 'ncumul_conf' of NW: 5 (2020-03-14).
24 Successfully updated field 'current_hosp' of NW: 6 (2020-03-14).
25 Successfully updated field 'source' of NW: http://www.nw.ch/coronastatistik (2020-03-14).
26 {'date': '2020-03-15', 'time': '', 'abbreviation_canton_and_fl': 'NW', 'ncumul_tested': '', 'ncumul_conf': 8, 'new_hosp': '', 'current_hosp': 6, 'current_icu': '', 'current_vent':
```

3 Hinter den Kulissen

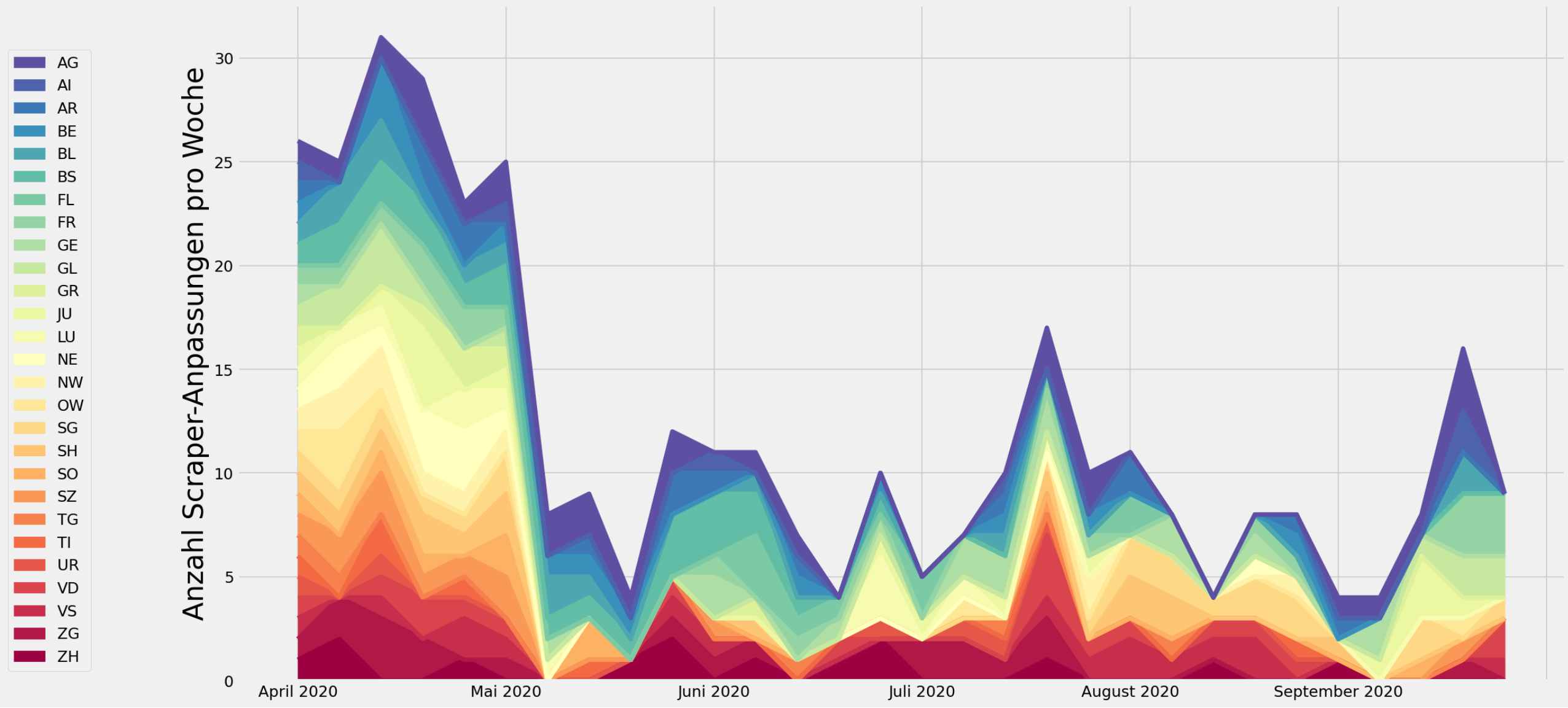
Repository-Statistik

Stand 22.10.2020

- Insgesamt über **16'500 Commits**
 - **55** Contributors
 - **882** Pull Requests
 - **351** Issues
 - **14'229** Scraper-Durchläufe je Kanton

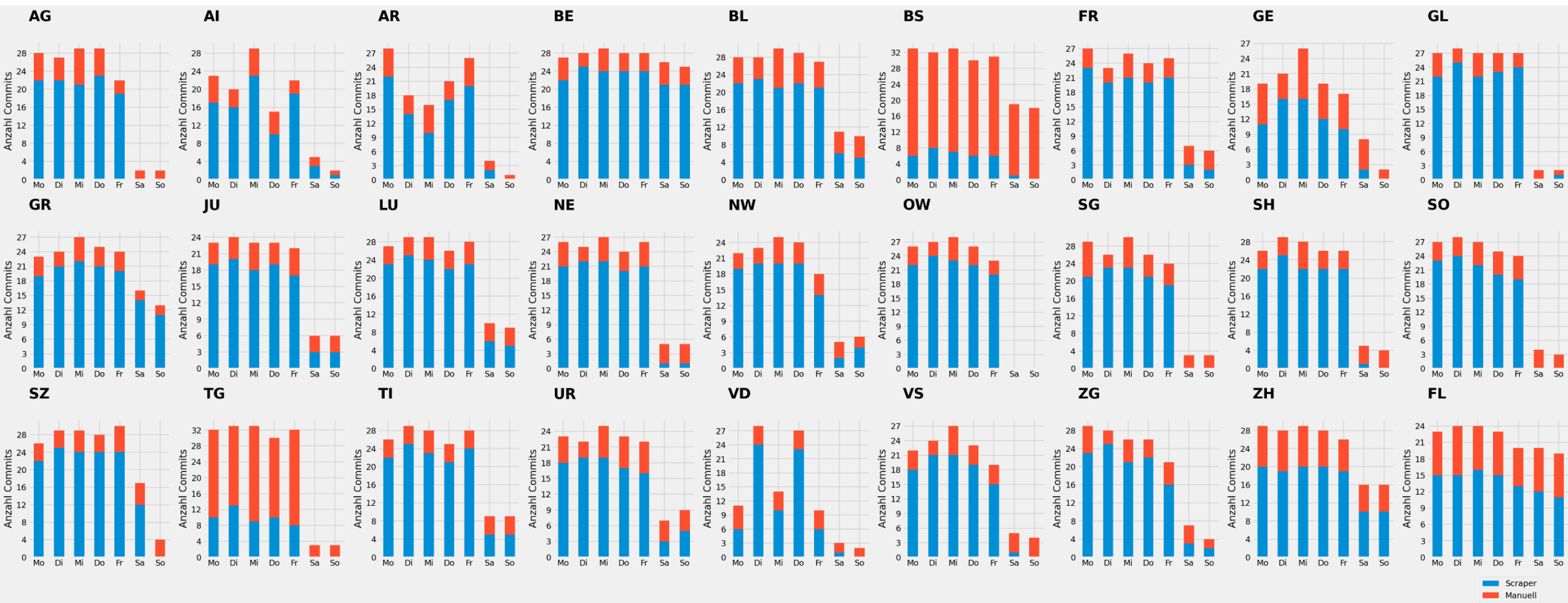


Scraper-Anpassungen



Daten von den Kantonen

Commits pro Wochentag und Kanton in den letzten 20 Wochen



Vielen Dank.