

# TikTok and News Sites Extraction

## Mastodon

Gruppe 1

March 20, 2023

## Was ist Mastodon?

- free and open-source social network
- sehr ähnlich zu Twitter
  - Man kann 'Toots' posten
  - Es gibt Hashtags, die trenden
  - Man kann anderen Nutzern folgen

## Wichtiger Unterschied zu Twitter: Mastodon ist [dezentral](#).

- Es gibt keinen Hauptserver, den alle User nutzen.
- Stattdessen gibt es Instanzen, auf denen User Accounts anlegen und posten können.
- Instanzen sind isoliert funktionsfähig.
- Können aber auch mit anderen Instanzen interagieren.

Das Programm hat folgende Funktionen:

- Eine Liste von Instanz-Domains in einer Datenbank speichern.
- Für jede der Instanzen Toots fetchen und speichern.
  - Welche Attribute der Toots gespeichert werden, ist auszuwählen.
  - Toots sollen nach verschiedenen Eigenschaften filterbar sein.
- Für jede Instanz Statistiken zur Antwortzeit angeben.

Für das Fetchen der Toots, muss man die IDs von Mastodon verstehen.

bis zu 6 bytes millisecond timestamp + 2 bytes unique sequence data

Entsprechend können Toots wie folgt chronologisch sortiert werden.

- Sortiere nach Größe: Neuere Toots haben längere IDs.
- Sortiere Lexikographisch.

Aus der Toot-ID ist also auch der Postzeitpunkt berechenbar.

# IDs in Mastodon

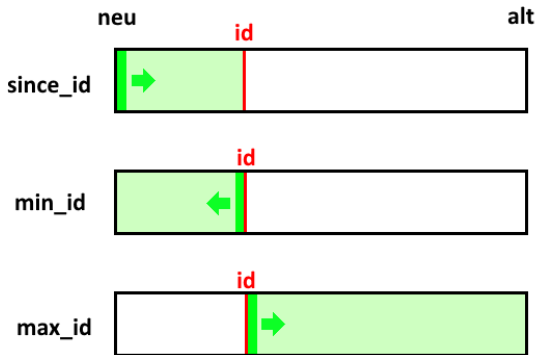
```
query = [  
  {  
    "$project": {  
      "_id": "$_id",  
      "caughtUp": "$caughtUp",  
      "posts": {  
        "$sortArray": {  
          "input": {  
            "$map": {  
              "input": "$posts",  
              "as": "post",  
              "in": {  
                "$mergeObjects": [{  
                  "id_length": {"$strLenCP": "$$post._id"},  
                  "_id": "$$post._id"  
                }]  
              }  
            }  
          },  
          "sortBy": {"id_length": -1, "_id": -1}  
        }  
      },  
    },  
  },  
  {  
    "$project": {  
      "_id": "$_id",  
      "values": ["$caughtUp", {"$first": "$posts._id"}, {"$last": "$posts._id"}]  
    }  
  }  
]
```

# Fetchen von Toots

Der folgende Link gibt 40 Toots aus der Instanz 'mastodon.social'.

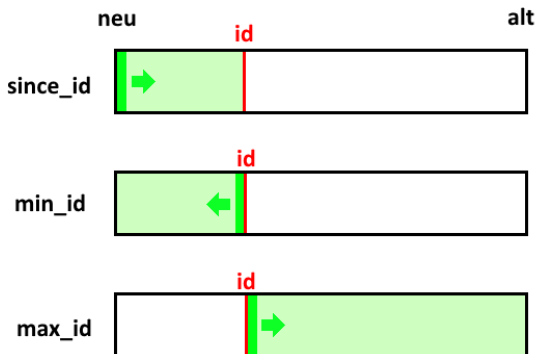
<https://mastodon.social/api/v1/timelines/public?limit=40>

Dazu kann man eine **ID** als Parameter geben.



# Fetchen von Toots

Dazu kann man eine **ID** als Parameter geben.



[https://mastodon.social/api/v1/timelines/public?limit=40&max\\_id=1111](https://mastodon.social/api/v1/timelines/public?limit=40&max_id=1111)

Gibt die neusten Toots, die eine **ID** > 1111 haben, also **älter** sind.

Da Mastodon dezentral ist, hängt das Fetchen von einzelnen Instanzen ab.

- Manche Instanzen brauchen Bewerbungen für die Registrierung.
- Einzelne Attribute der Toots können fehlen.
- 5xx Fehler und unterschiedliche Antwortzeiten.