„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Was war passiert:

Den Twitter-Account @SWPde gab es seit Januar 2009

28735 Tweets / 3650 Tage (10 Jahre, der Einfachheit halber ohne Schaltjahre angesetzt)
→ im Schnitt 7,9 Tweets pro Tag

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Kurze Stichprobe:

01.11.2018 - 10 Tweets
02.11.2018 - 3 Tweets
03.11.2018 - 8 Tweets

Joah, kommt hin, ~ 7 pro Tag.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Aber:
Seit mindestens 21.01.2019 im Status „geschützter Account“

Bedeutet:

- keine Retweets möglich
- Follower werden ist nur nach Bestätigung möglich (die nicht mehr gewährt wurde)

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Außerdem deutlich gesunkene Tweet-Frequenz:

21.01.2019 - 3 Tweets
22.01.2019 - 1 Tweet
23.01.2019 - 1 Tweet
24.01.2019 - 0 Tweets
25.01.2019 - 1 Tweet
26.01.2019 - 2 Tweets
27.01.2019 - 0 Tweets
28.01.2019 - 0 Tweets
29.01.2019 - 0 Tweets
30.01.2019 - 1 Tweet
31.01.2019 - 2 Tweets
01.02.2019 - 1 Tweet
02.02.2019 - 1 Tweet
03.02.2019 - 0 Tweets
04.02.2019 - 0 Tweets
05.02.2019 - 3 Tweets
06.02.2019 - 3 Tweets
07.02.2019 - 2 Tweets
08.02.2019 - 0 Tweets
09.02.2019 - 0 Tweets
10.02.2019 - 0 Tweets
11.02.2019 - 1 Tweet
12.02.2019 - 0 Tweets
13.02.2019 - 0 Tweets
14.02.2019 - 1 Tweet
15.02.2019 - 3 Tweets
16.02.2019 - 1 Tweet
17.02.2019 - 1 Tweet
18.02.2019 - 2 Tweets
19.02.2019 - 2 Tweets
20.02.2019 - 1 Tweet
21.02.2019 - 1 Tweet
22.02.2019 - 3 Tweets
24.02.2019 - 1 Tweet

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Dann am 25.02.2019, 10:23: Ein Abschiedstweet!

https://twitter.com/SWPde/status/1099962974012354566

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

5 Minuten danach (10:28): Accountlöschung.

Ganz. dumme. Idee.

Der Account hatte ~ 13k Follower

(Cave: Es sind so einige Bots dabei - das sollte mal jemand genauer analysieren.).

So einen Account wirft man nicht einfach weg.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

25.02.2019, 13:37

Per FB fragt die Piratenpartei Ulm bei der SWP nach, und erhält folgende PR-Blafasel-Bullshit-Antwort:

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 10:09

Meine ersten Bot-Tests unter meinem eigenen Account haben soweit funktioniert ...

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 10:09

Meine ersten Bot-Tests unter meinem eigenen Account haben soweit funktioniert …
… und mir gegen 11:13 einen Shadowban von Twitter eingebracht. Ohne mein Follower zu sein, waren Tweets von mir, und in denen ich ge-@-mentioned wurde, nicht über die Suche auffindbar. Möp.
Mag vielleicht auch daran liegen, dass mein Account zwar auch schon seit Mitte 2010 registriert ist, aber nie eine Telefonnummer hinterlegt hatte.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 11:04

Irgendwer bei der SWP (Marketingmensch?) hat wohl bemerkt, dass das mit der Accountlöschung blöd war, und hat sie rückgängig gemacht.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

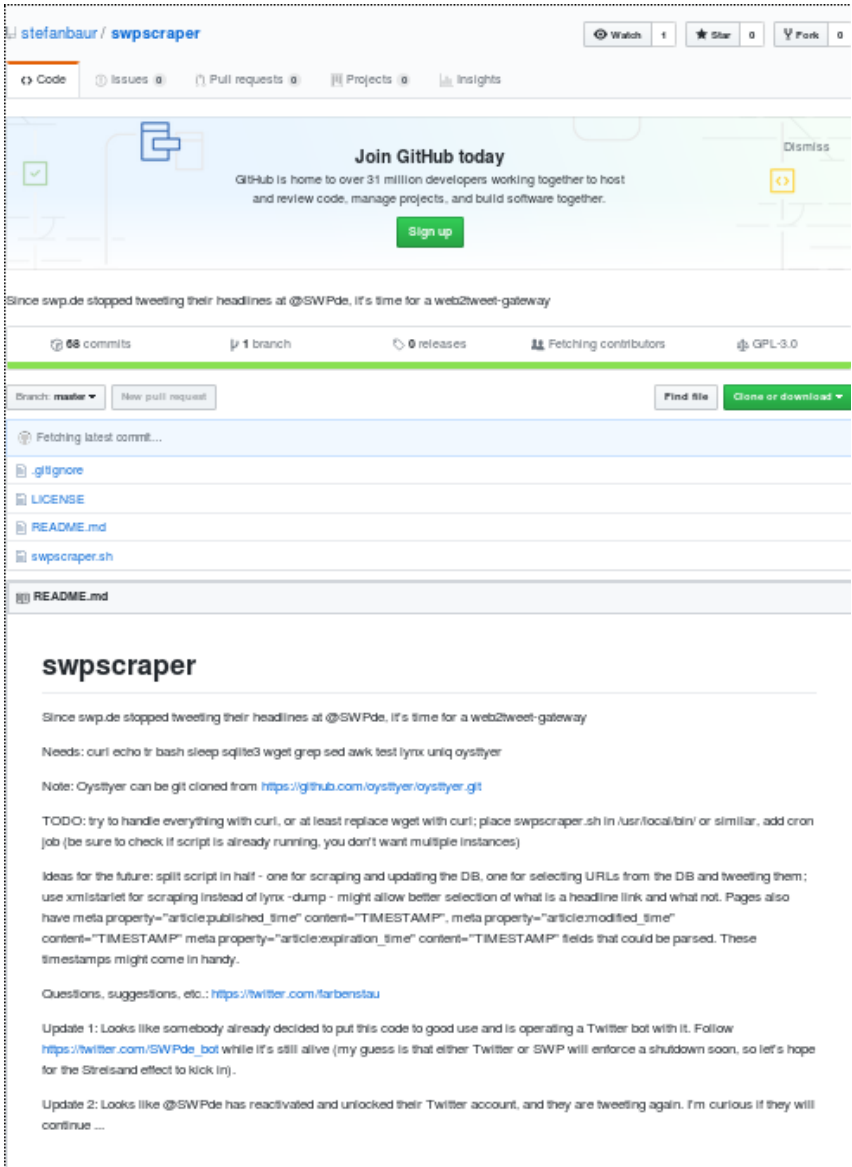
26.02.2019, 11:28

Geschützt ist der Account nach wie vor, aber kurze Zeit später auch das nicht mehr. Neu getwittert wird trotzdem nicht.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 11:56

Ich veröffentliche den Source meines Bots auf github:
https://github.com/stefanbaur/swpscraper/



„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 12:00

... und tweete das auch.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

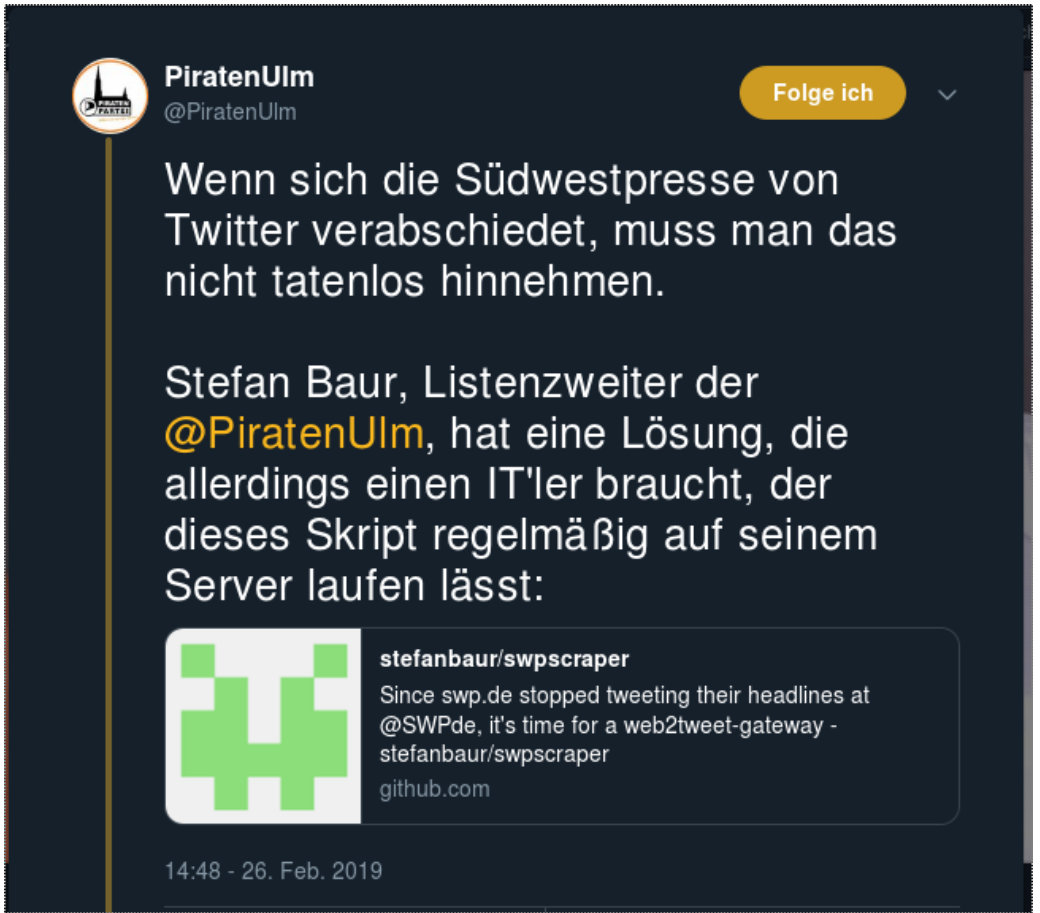
26.02.2019, 13:37

Die Piraten retweeten den Link auf mein Github-Repo:

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 14:48

Die Piraten retweeten den Link auf mein Github-Repo erneut, ohne @-Mention, damit mein Shadowban sich nicht auswirkt:



„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, ??:??

$JEMAND - vermutlich aus dem Piratenumfeld, oder irgendein Follower von mir, aufgrund meines Tweets und meiner durch den Shadowban eingeschränkten Reichweite - registriert @SWPde_bot auf Twitter und setzt meinen Bot darauf an.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 15:32

Erster Bot-Tweet

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 15:55

Der Bot beweist Humor:

https://youtu.be/xJC1dUs9lBM?t=14

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 18:17 – 26.02.2019, 18:29

Backoff-Mechanismus im Botskript gefixt und ...

Wartezeit zwischen den Tweets auf zwischen 60 und 180 Sekunden erhöht (alt: 60s-120s) sowie …

Kategorien "Panorama" und "Sport" im Source des Scrapers ausgeschlossen ...

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 18:32

... aber wohl leider zu spät für den Botbetreiber. Bot spotzt (Rate Limiting) –
→ Genau 2 Stunden nach erstem Tweet – insgesamt 60 Bot-Tweets bisher

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 18:47

Wartezeit zwischen den Tweets auf mindestens 120 Sekunden erhöht (alt: 60s)

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 18:52

Backoff-Mechanismus tat immer noch nicht so, wie er sollte.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 19:29

Bot ist gegen Artikel 13:

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 19:47

Bot hat Anwandlungen von HAL:

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

26.02.2019, 20:27

Bot traurig (wieder Rate Limiting):
→ Wieder grob 2 Stunden später – nur 13 neue Bot-Tweets seit letzter Unterbrechung

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 09:01

Bot nimmt neuen Anlauf (Neuer Tag, neues Glück …):

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 10:45

Scraper achtet nun darauf, dass Seiten, die nur Videos, nur Bilder, oder gar keine Bilder enthalten, nicht getweeted werden

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 13:01

Festgestellt, dass twidge nur 140 Zeichen beherrscht ...

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 13:19

Bot spotzt wieder (evtl. hat der Betreiber das Update von 10:45 nicht eingespielt?):
→ Dieses Mal gut 4 Stunden - insgesamt 32 Bot-Tweets seit letzter Unterbrechung

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 14:47

Code auf oysttyer als Commandline-Twitter-Client umgestellt. Das ist auch bescheuert (kein Return Code im Fehlerfall), aber anders bescheuert, und kann 280-Zeichen-Tweets senden.
Außerdem scheint es Rate Limits selbst zu erkennen und vorher abzubremsen, anstatt wie twidge blindlings hineinzurennen.

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 14:50

Bot nimmt neuen Anlauf:

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 14:58

Ein erster neuer Newstweet von @SWPde (Auffällig: reiner Link, kein Text dazu.):

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

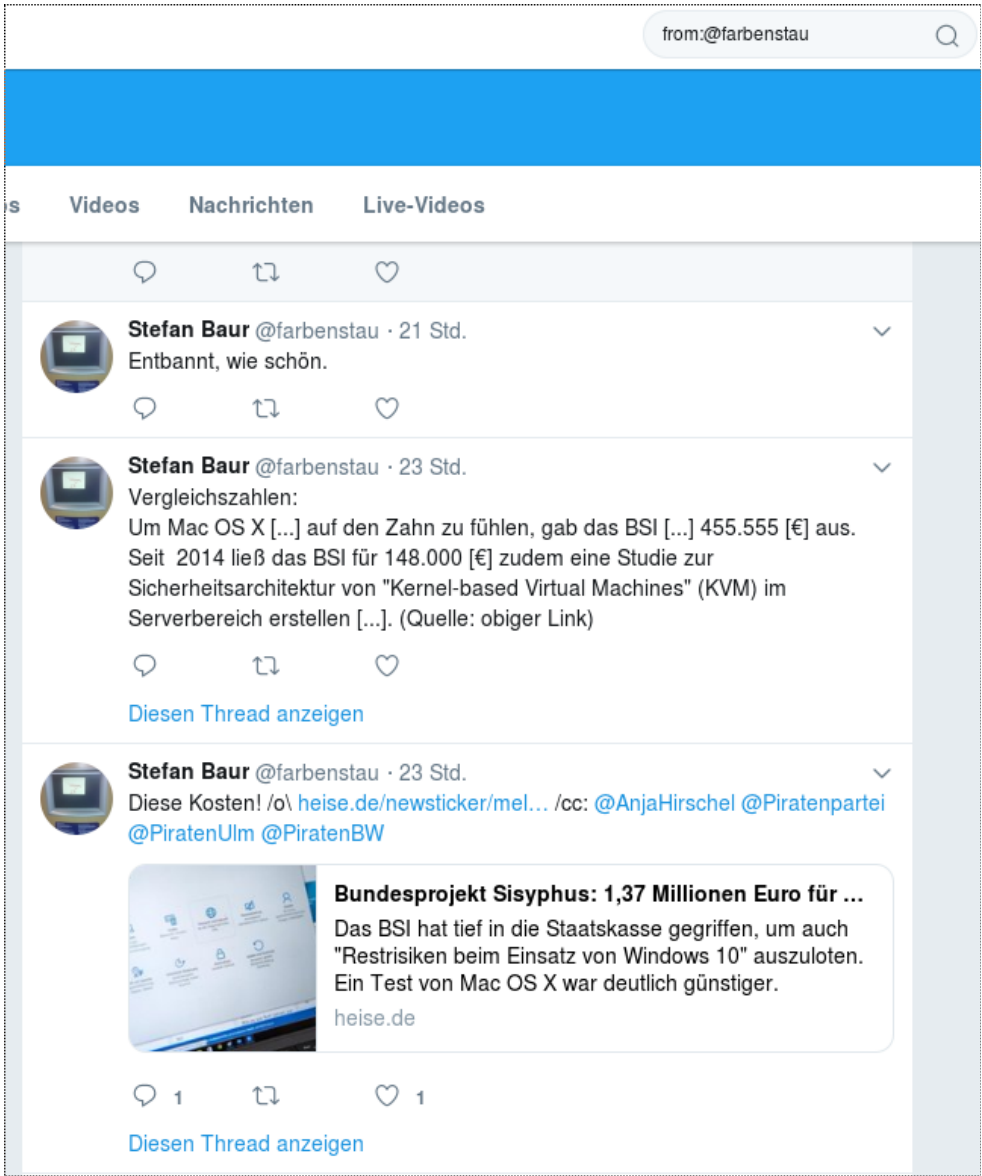
27.02.2019, 15:35

Bot stellt fest, dass er arbeitslos ist:
→ insgesamt 11 Bot-Tweets seit letzter Unterbrechung

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

27.02.2019, 17:54

Mein Shadowban ist aufgehoben – ob es nun daran lag, dass ich eine Handynummer und neue Mailadresse hinterlegt habe, oder ob er automatisch expired, weiß keiner ...



„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Funktionsprinzip Bot

- Hole die swp.de-Startseite und extrahiere alle „Hidden Links“ – was nicht „Hidden Link“ ist, gehört sehr wahrscheinlich zum Newsticker, den wollen wir nicht (zu hohe Updatefrequenz, keine Bilder)
- Sortiere Dubletten aus der Linkliste aus
- Folge jedem der übriggebliebenen Links, wenn er nicht zu Panorama oder Sport gehört, und extrahiere den Namen aus dem <title> - Tag
- Prüfe weiterhin, ob die Tags enthalten sind, die auf eine Bildergalerie oder ein Video hindeuten - wenn ja, überspringe dieses Ergebnis
- Prüfe weiterhin, ob einer der Tags enthalten ist, der auf mindestens ein anwesendes Bild im Artikel hindeutet - wenn nein, überspringe dieses Ergebnis
- Kürze den Titel notfalls so, dass er in (280-24) Zeichen passt - Links werden per Twitter-URL-Shortener auf 23 Zeichen zurechtgestuzt, 1 Zeichen brauchen wir als trennendes Blank
- Kombiniere Titel + Link zu einem Tweet und übergib ihn an den Client

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Bells and Whistles

- Es sind randomisierte Wartezeiten hinterlegt, sowohl beim Scrapen (für Startseite und Linkfollow), als auch beim Twittern, damit man nicht so schnell als Bot erkannt wird (und vor allem keinen Tweetstorm auslöst)
- Der User-Agent-String wird beim Scrapen auf gängige User Agents gesetzt, damit seitens SWP nicht z.B. auf wget oder Lynx im User-Agent-String gefiltert werden kann - wenn sie ernsthaftes Blocken betreiben wollten, müsste man vielleicht dazu übergehen, die Seiten mit allen Bildern etc. herunterzuladen, oder einen echten Browser fernsteuern und dann die Seiten speichern lassen, damit es echter wirkt.
- Wer will, kann über $PREFACE noch einen Text definieren, der vor jeden Tweet gesetzt wird, z.B. #SWPde oder #Bot - die Länge wird beim Kürzen des Titels automatisch mit berücksichtigt.
- Um keinen Tweetsturm auszulösen, wenn das Skript das erste Mal läuft, wird bei einer leeren Datenbank die Datenbank nur befüllt, ohne die Einträge zu tweeten - nur, wenn die Datenbank nicht leer ist, werden die Deltas getweetet
- Beim Neustart wird in der Datenbank nachgesehen, ob es noch Nachrichten gibt, die noch nicht getweetet wurden - diese werden zuerst abgesendet, bevor der nächste Scraping-Lauf beginnt

„Ich wollt' doch nur die Südwest-Presse-Schlagzeilen via @SWPde auf Twitter lesen ... und jetzt habe ich einen Twitter-Bot auf github.“

Verbesserungswürdig

- Der Scraper könnte auf xmlstarlet umgestellt werden. Durch eine bessere Analyse könnten dann feinere Entscheidungskriterien festgelegt werden, welche Artikel getwittert werden sollen und welche nicht.
- Aktuell ist das sehr krude mit Stringfiltern und Black- bzw. Whitelists umgesetzt.
- Wenn man das ganze verteilt und redundant laufen lassen möchte, sollte man sqlite durch eine richtige, spiegelungsfähige Datenbank ersetzen.
- Und natürlich die TODOs aus dem README.md ...