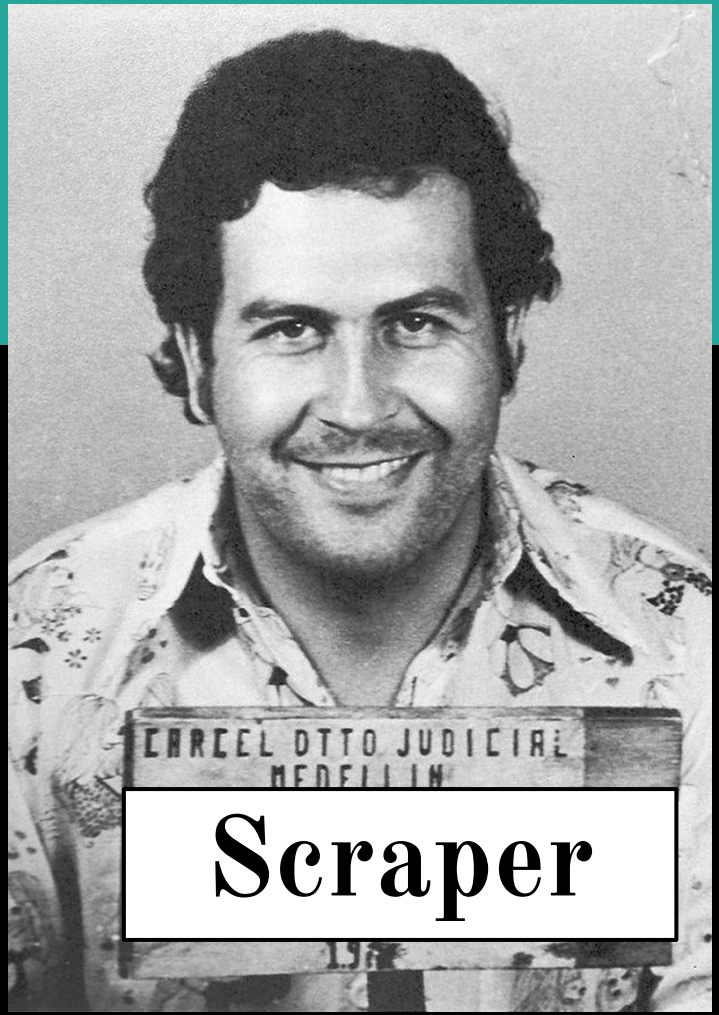
Rosanna &
Konstantin



Scraper

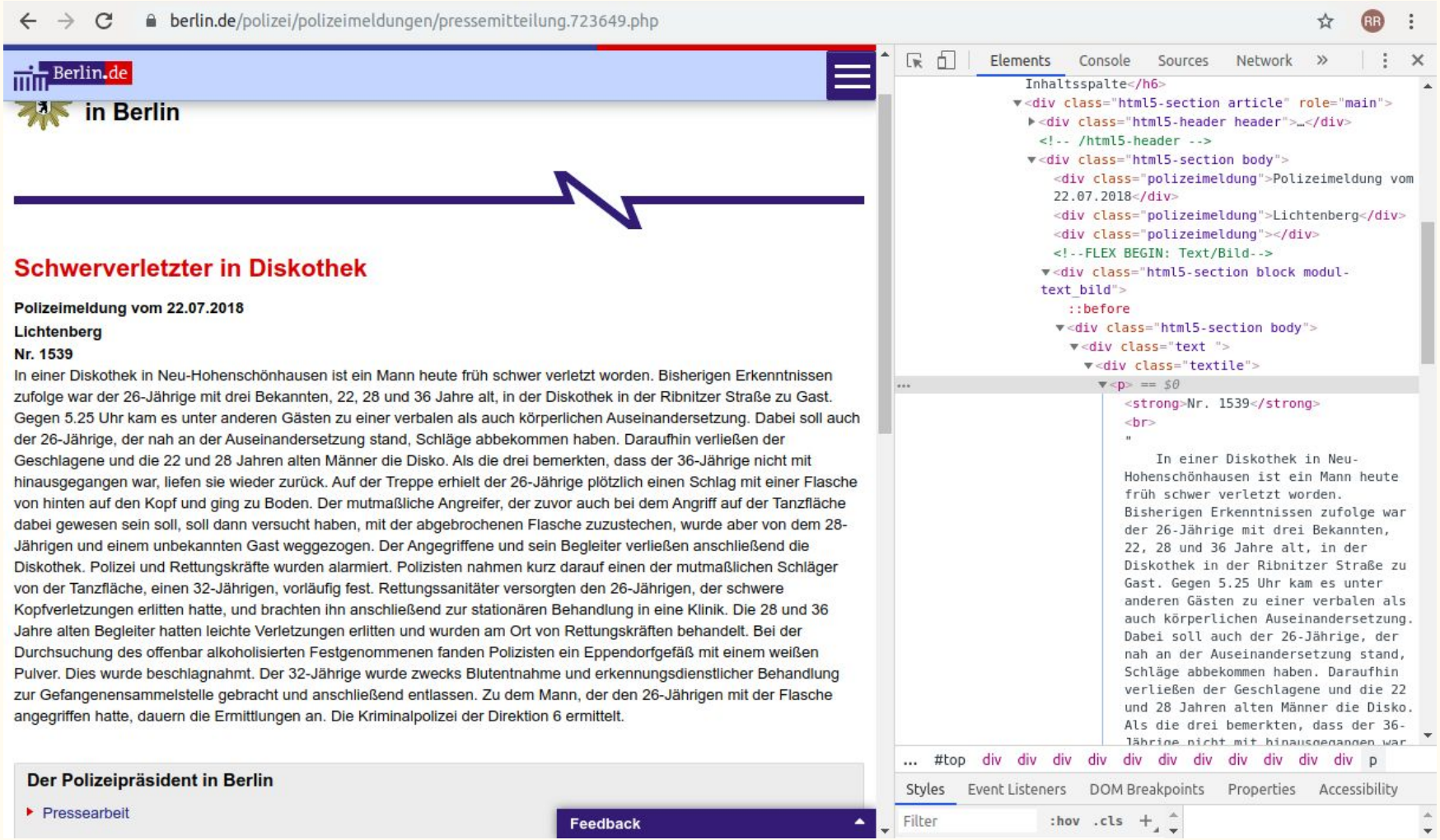
Ziel

- Scraping aller bis Mitte November ‘19 vorhandener Polizeimeldungen auf berlin.de

- ohne Framework, Schritt für Schritt, “händisch”

Seitenaufbau

- Archiv 2014-2019 (www.berlin.de/polizei/polizeimeldungen/archiv/):
  - pro Jahr bis zu 53 Unterseiten
- pro Unterseite des Archivs (...meldungen/archiv/2019/?page_at_1_0=19) :
  - jeweils 50 Links zu Meldungen
  - Bsp. Archiv 2018: 47 Unterseiten, d.h. ~47*50=2350 Links zu Meldungen
- typische Meldung:
  - Überschrift
  - Datum
  - Bezirk
  - Nr.
  - Beschreibung


berlin.de/polizei/polizeimeldungen/pressemitteilung.723649.php
Berlin.de
in Berlin
Schwerverletzter in Diskothek
Polizeimeldung vom 22.07.2018
Lichtenberg
Nr. 1539
In einer Diskothek in Neu-Hohenschönhausen ist ein Mann heute früh schwer verletzt worden. Bisherigen Erkenntnissen zufolge war der 26-Jährige mit drei Bekannten, 22, 28 und 36 Jahre alt, in der Diskothek in der Ribnitzer Straße zu Gast. Gegen 5.25 Uhr kam es unter anderen Gästen zu einer verbalen als auch körperlichen Auseinandersetzung. Dabei soll auch der 26-Jährige, der nah an der Auseinandersetzung stand, Schläge abbekommen haben. Daraufhin verließen der Geschlagene und die 22 und 28 Jahren alten Männer die Disko. Als die drei bemerkten, dass der 36-Jährige nicht mit hinausgegangen war, liefen sie wieder zurück. Auf der Treppe erhielt der 26-Jährige plötzlich einen Schlag mit einer Flasche von hinten auf den Kopf und ging zu Boden. Der mutmaßliche Angreifer, der zuvor auch bei dem Angriff auf der Tanzfläche dabei gewesen sein soll, soll dann versucht haben, mit der abgebrochenen Flasche zuzustechen, wurde aber von dem 28-Jährigen und einem unbekannten Gast weggezogen. Der Angegriffene und sein Begleiter verließen anschließend die Diskothek. Polizei und Rettungskräfte wurden alarmiert. Polizisten nahmen kurz darauf einen der mutmaßlichen Schläger von der Tanzfläche, einen 32-Jährigen, vorläufig fest. Rettungssanitäter versorgten den 26-Jährigen, der schwere Kopfverletzungen erlitten hatte, und brachten ihn anschließend zur stationären Behandlung in eine Klinik. Die 28 und 36 Jahre alten Begleiter hatten leichte Verletzungen erlitten und wurden am Ort von Rettungskräften behandelt. Bei der Durchsuchung des offenbar alkoholisierten Festgenommenen fanden Polizisten ein Eppendorfgefäß mit einem weißen Pulver. Dies wurde beschlagnahmt. Der 32-Jährige wurde zwecks Blutentnahme und erkennungsdienstlicher Behandlung zur Gefangenensammelstelle gebracht und anschließend entlassen. Zu dem Mann, der den 26-Jährigen mit der Flasche angegriffen hatte, dauern die Ermittlungen an. Die Kriminalpolizei der Direktion 6 ermittelt.
Der Polizeipräsident in Berlin
Pressearbeit
Feedback
Elements
Console
Sources
Network
Inhaltsspalte</h6>

…
<!-- /html5-header -->

Polizeimeldung vom 22.07.2018
Lichtenberg

<!--FLEX BEGIN: Text/Bild-->

::before

<p> == $0
Nr. 1539

"
In einer Diskothek in Neu-Hohenschönhausen ist ein Mann heute früh schwer verletzt worden. Bisherigen Erkenntnissen zufolge war der 26-Jährige mit drei Bekannten, 22, 28 und 36 Jahre alt, in der Diskothek in der Ribnitzer Straße zu Gast. Gegen 5.25 Uhr kam es unter anderen Gästen zu einer verbalen als auch körperlichen Auseinandersetzung. Dabei soll auch der 26-Jährige, der nah an der Auseinandersetzung stand, Schläge abbekommen haben. Daraufhin verließen der Geschlagene und die 22 und 28 Jahren alten Männer die Disko. Als die drei bemerkten, dass der 36-Jährige nicht mit hinausgegangen war
... #top div div div div div div div div div div div p
Styles
Event Listeners
DOM Breakpoints
Properties
Accessibility
Filter
:hov .cls

Strategie

1. URLs zu allen Archiven
2. für alle Jahre/Archive: URLs zu allen Unterseiten
3. für alle Unterseiten: URLs zu allen Meldungen
4. für jede URL: Meldung mit
    a. Überschrift
    b. Datum
    c. Ort
    d. Zwischentitel (wie Nr.)
    e. Beschreibung
    f. URL
    g. [HTTP-Statuscode]

Erste Versuche

- einzelne Anfragen mit Python, requests, lxml in Konsole
- xpath locator: HTML-Elemente finden & Daten extrahieren
- Jupyter Notebook
- Funktionen auf kleinen Mengen ausprobieren
- Ergebnisse als .csv abspeichern
- pro Meldung ein Dictionary
- Liste von Dictionaries in JSON speichern
- häufige Antwort:
    - HTTP-Statuscode 429 (*Too Many Requests*) & leeres Dictionary

Scraping Strategie

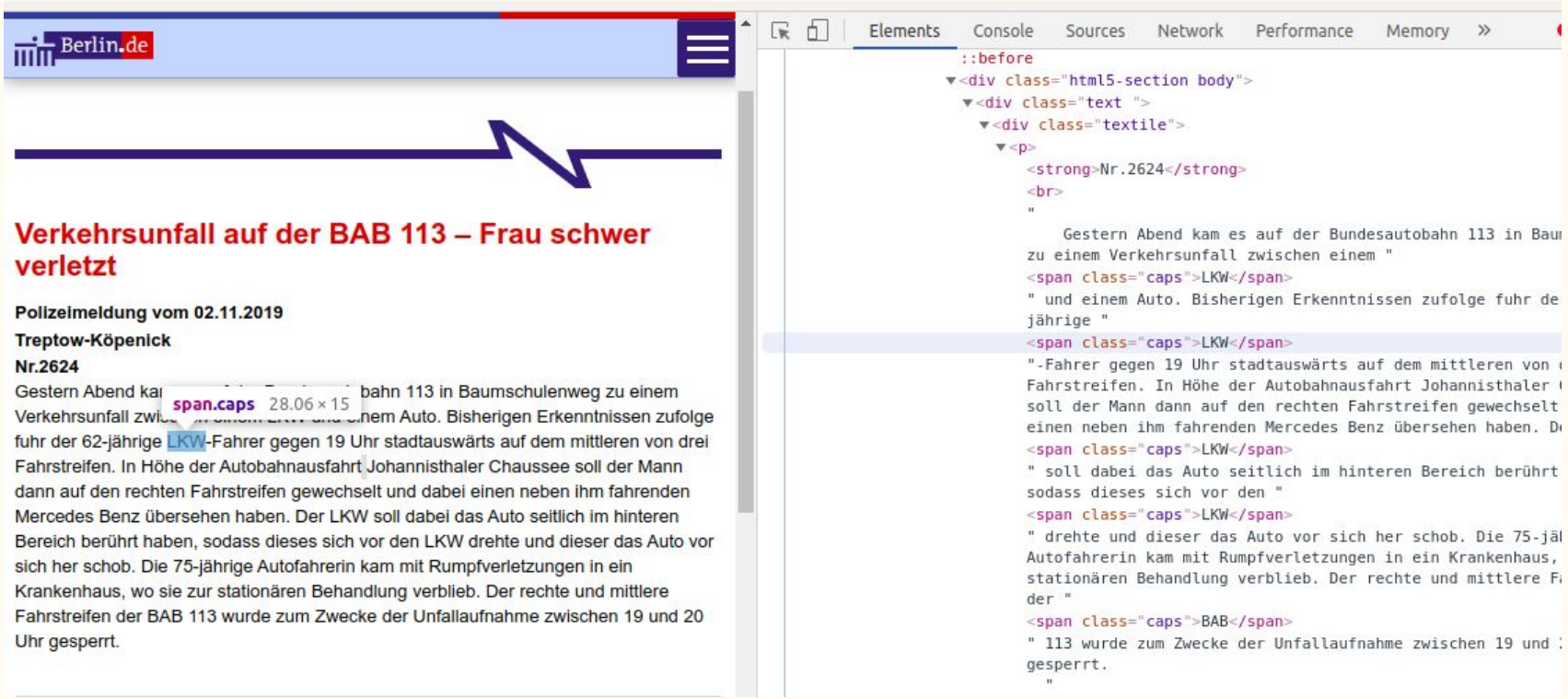
- Konvertierung in Script
- Graceful Scraping
  - Rate Limiter: Decorator für Requests
  - Adaptive Scraping Timeouts
  - Jeder Server hat ~~Gefühle~~ Status Codes
- Wiederverwendung von Daten
  - Lokales auslesen von Statischen Daten
  - Verschiedene Modi: Refresh, Live, Test
- Logging
  - Auswerten
  - Verbessern

Ausnahmen

- Gründe u.a.:
  - Formatierungsfehler auf Website
  - ungewöhnliche Inhalte
  - Fall nicht auf einen Bezirk beschränkt
- Ausnahmebehandlung
  - erforderlich?
  - machbar?

- Beispiele...

Ausnahme I


Berlin.de
Verkehrsunfall auf der BAB 113 – Frau schwer verletzt
Polizeimeldung vom 02.11.2019
Treptow-Köpenick
Nr.2624
Gestern Abend ka bahn 113 in Baumschulenweg zu einem Verkehrsunfall zwi nem Auto. Bisherigen Erkenntnissen zufolge fuhr der 62-jährige LKW-Fahrer gegen 19 Uhr stadtauswärts auf dem mittleren von drei Fahrstreifen. In Höhe der Autobahnausfahrt Johannisthaler Chaussee soll der Mann dann auf den rechten Fahrstreifen gewechselt und dabei einen neben ihm fahrenden Mercedes Benz übersehen haben. Der LKW soll dabei das Auto seitlich im hinteren Bereich berührt haben, sodass dieses sich vor den LKW drehte und dieser das Auto vor sich her schob. Die 75-jährige Autofahrerin kam mit Rumpfverletzungen in ein Krankenhaus, wo sie zur stationären Behandlung verblieb. Der rechte und mittlere Fahrstreifen der BAB 113 wurde zum Zwecke der Unfallaufnahme zwischen 19 und 20 Uhr gesperrt.
span.caps 28.06 × 15
Elements
Console
Sources
Network
Performance
Memory
::before

<p>
Nr.2624

" Gestern Abend kam es auf der Bundesautobahn 113 in Bau zu einem Verkehrsunfall zwischen einem "
LKW
" und einem Auto. Bisherigen Erkenntnissen zufolge fuhr de jährige "
LKW
"-Fahrer gegen 19 Uhr stadtauswärts auf dem mittleren von Fahrstreifen. In Höhe der Autobahnausfahrt Johannisthaler soll der Mann dann auf den rechten Fahrstreifen gewechselt einen neben ihm fahrenden Mercedes Benz übersehen haben. D
LKW
" soll dabei das Auto seitlich im hinteren Bereich berührt sodass dieses sich vor den "
LKW
" drehte und dieser das Auto vor sich her schob. Die 75-jäl Autofahrerin kam mit Rumpfverletzungen in ein Krankenhaus, stationären Behandlung verblieb. Der rechte und mittlere F der "
BAB
" 113 wurde zum Zwecke der Unfallaufnahme zwischen 19 und gesperrt. "

Ausnahme II

Mutmaßlicher Vergewaltiger im Harz festgenommen

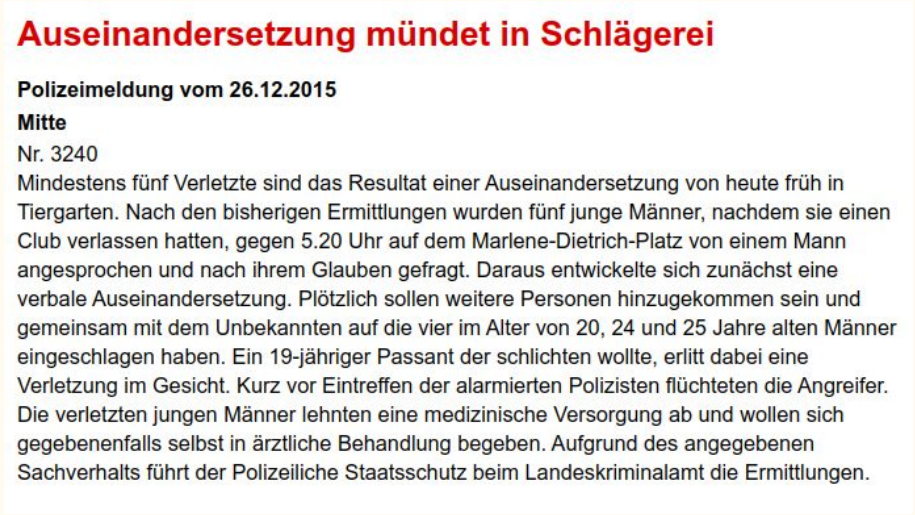
Polizeimeldung vom 30.12.2015
Gemeinsame Meldung Polizei und Staatsanwaltschaft Berlin

Berlin/Niedersachsen
Nr. 3266
Nach intensiven Ermittlungen des Landeskriminalamtes und der Staatsanwaltschaft Berlin nahmen niedersächsische Fahnder gestern Mittag einen mutmaßlichen Vergewaltiger fest. Der 30-Jährige wurde gegen 13.50 Uhr in Herzberg am Harz in einem Reihenhaus angetroffen und in Untersuchungshaft genommen. Er steht im dringenden Verdacht, am 31. Mai eine Frau in Berlin-Wittenau vergewaltigt zu haben.

Erstmeldung Nr. 2317 vom 28. September 2015 – Tatverdächtiger einer Vergewaltigung gesucht
Mit Bildern aus einer Überwachungskamera und einer Videosequenz sucht die Polizei Berlin nach einem derzeit unbekannten Mann, der in Verdacht steht, eine Frau in Wittenau vergewaltigt zu haben. Nach den bisherigen Ermittlungen war der Gesuchte der 37-Jährigen am Sonntag, den 31. Mai 2015 gegen 3.45 Uhr gefolgt, nachdem sie am U-Bahnhof Rathaus Reinickendorf ausgestiegen war, hat sie überwältigt und sich an ihr vergangen.
Der Tatverdächtige hat ein südeuropäisches Aussehen und sprach das Opfer auf serbo-kroatisch an. Er ist 25 bis 35 Jahre alt, trug einen roten Kapuzenpullover, eine Jeans und schwarze Turnschuhe mit weißer Sohle.


```
{'response': 200,
 'headline': 'Mutmaßlicher Vergewaltiger im
Harz festgenommen',
 'published': 'Polizeimeldung vom
30.12.2015',
 'bezirk': '',
 'subheads': ['Gemeinsame Meldung Polizei
und Staatsanwaltschaft Berlin',
  'Berlin/Niedersachsen',
  'Nr. 3266',
  'Erstmeldung Nr. 2317 vom 28. September
2015 – Tatverdächtiger einer Vergewaltigung
gesucht'],
 'article': 'Nach intensiven Ermittlungen
des Landeskriminalamtes und der
Staatsanwaltschaft Berlin nahmen
niedersächsische Fahnder gestern Mittag
einen mutmaßlichen Vergewaltiger fest.[...]
Er ist 25 bis 35 Jahre alt, trug einen roten
Kapuzenpullover, eine Jeans und schwarze
Turnschuhe mit weißer Sohle.',
 'url':
'https://www.berlin.de/polizei/polizeimeldun
gen/pressemitteilung.379061.php'}
```

Ausnahme III



Auseinandersetzung mündet in Schlägerei

Polizeimeldung vom 26.12.2015

Mitte

Nr. 3240

Mindestens fünf Verletzte sind das Resultat einer Auseinandersetzung von heute früh in Tiergarten. Nach den bisherigen Ermittlungen wurden fünf junge Männer, nachdem sie einen Club verlassen hatten, gegen 5.20 Uhr auf dem Marlene-Dietrich-Platz von einem Mann angesprochen und nach ihrem Glauben gefragt. Daraus entwickelte sich zunächst eine verbale Auseinandersetzung. Plötzlich sollen weitere Personen hinzugekommen sein und gemeinsam mit dem Unbekannten auf die vier im Alter von 20, 24 und 25 Jahre alten Männer eingeschlagen haben. Ein 19-jähriger Passant der schlichten wollte, erlitt dabei eine Verletzung im Gesicht. Kurz vor Eintreffen der alarmierten Polizisten flüchteten die Angreifer. Die verletzten jungen Männer lehnten eine medizinische Versorgung ab und wollen sich gegebenenfalls selbst in ärztliche Behandlung begeben. Aufgrund des angegebenen Sachverhalts führt der Polizeiliche Staatsschutz beim Landeskriminalamt die Ermittlungen.


```
{'response': 200,
 'headline': 'Auseinandersetzung mündet in
Schlägerei',
 'published': 'Polizeimeldung vom 26.12.2015',
 'bezirk': 'Mitte',
 'subheads': [],
 'article': 'Nr. 3240\n    Mindestens fünf
Verletzte sind das Resultat einer
Auseinandersetzung von heute früh in Tiergarten.
Nach den bisherigen Ermittlungen wurden fünf
junge Männer, nachdem sie einen Club verlassen
hatten, gegen 5.20 Uhr auf dem
Marlene-Dietrich-Platz von einem Mann
angesprochen und nach ihrem Glauben gefragt.
[...] Kurz vor Eintreffen der alarmierten
Polizisten flüchteten die Angreifer. Die
verletzten jungen Männer lehnten eine
medizinische Versorgung ab und wollen sich
gegebenenfalls selbst in ärztliche Behandlung
begeben. Aufgrund des angegebenen Sachverhalts
führt der Polizeiliche Staatsschutz beim
Landeskriminalamt die Ermittlungen.',
 'url':
'https://www.berlin.de/polizei/polizeimeldungen/
pressemitteilung.428390.php'}
```

Ausnahme IV

Brennende Fahrzeuge

Polizeimeldung vom 27.12.2014
Treptow-Köpenick/Neukölln
In Adlershof und Rudow haben Unbekannte in der vergangenen Nacht zwei Fahrzeuge in Brand gesetzt. Verletzt wurde in beiden Fällen niemand. Der Polizeiliche Staatsschutz prüft, ob die Taten politisch motiviert waren.

Nr. 3063
Gegen 2.45 Uhr bemerkte ein 25-jähriger Zeuge in der Husstraße einen Feuerschein am Radkasten eines geparkten „Ford“ und alarmierte daraufhin Feuerwehr und Polizei. Die Brandbekämpfer löschten die Flammen. Der Motorraum des „Ford“ brannte vollständig aus.

Nr. 3064
In der Köpenicker Straße gelang es dem Halter eines „VW“ noch vor Eintreffen der Rettungskräfte ein Feuer an seinem Fahrzeug zu löschen. Ein Zeuge hatte die Flammen an dem „Caddy“ gegen 3.10 Uhr bemerkt und sowohl die Rettungskräfte als auch den Besitzer des Autos informiert.


```
{'response': 200,
 'headline': 'Brennende Fahrzeuge',
 'published': 'Polizeimeldung vom 27.12.2014',
 'bezirk': '',
 'subheads': ['Treptow-Köpenick/Neukölln',
'Nr. 3063', 'Nr. 3064'],
 'article': 'In Adlershof und Rudow haben
Unbekannte in der vergangenen Nacht zwei
Fahrzeuge in Brand gesetzt. Verletzt wurde in
beiden Fällen niemand. Der Polizeiliche
Staatsschutz prüft, ob die Taten politisch
motiviert waren. Gegen 2.45 Uhr bemerkte ein
25-jähriger Zeuge in der Husstraße einen
Feuerschein am Radkasten eines geparkten
„Ford“ und alarmierte daraufhin Feuerwehr und
Polizei. Die Brandbekämpfer löschten die
Flammen. Der Motorraum des „Ford“ brannte
vollständig aus. In der Köpenicker Straße
gelang es dem Halter eines „VW“ noch vor
Eintreffen der Rettungskräfte ein Feuer an
seinem Fahrzeug zu löschen. Ein Zeuge hatte
die Flammen an dem „Caddy“ gegen 3.10 Uhr
bemerkt und sowohl die Rettungskräfte als auch
den Besitzer des Autos informiert.',
 'url':
'https://www.berlin.de/polizei/polizeimeldunge
n/pressemitteilung.246955.php'}
```

Deployment / Live

- Installation auf den Server
  - Virtual Env
  - Cronjob immer um 12 an mehreren Tagen
- Verlauf des Scrapings
  - Code-Probleme
  - 500er Probleme: Server Timeout immer um Mitternacht
- Ergebnisse
  - Dauer: ca. 13h
  - Inhalte: 9647 Dokumente
  - Errors: Serverseitig