Wirksames Wundermittel Web-Scraping

mit Python, BeautifulSoup und Selenium Webdriver

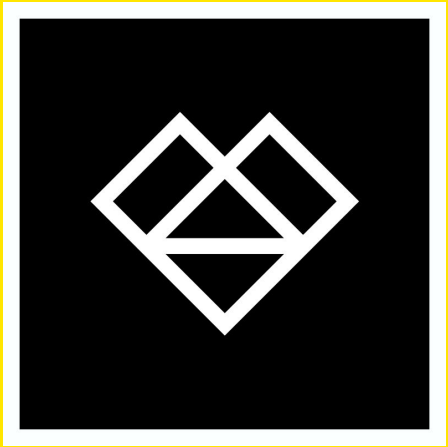
Dieter Föttinger

df@roeschke.net

Fullstack Developer
Frontend/Backend
Apps/Mobile
DevOps
Art
C#, PHP, Javascript, Python, u.v.a ;)
Digital seit 1980
ZX-81, VC-20, VC-64, TI99/4A, Amiga, Atari, Apple…
(fu**ing 39 years!)

roeschke&roeschke GmbH

digital. event. marketing.



seit 2006

roeschke&roeschke GmbH

Team 12/16
https://roeschke.net

Works:
https://www.opernball-nuernberg.de
https://www.ball-der-unternehmer.de
https://www.juradirekt.com
https://indag.de
...

Inhalt

- Allgemeines zu Digitalisierung / Web Scraping
- Warum Python?
- Was wir brauchen / Python installieren
- BeautifulSoup / Pandas
- Demo einfaches Scraping mit BeautifulSoup
- Web Automation mit Selenium + Webdriver installieren
- Demo komplexeres Scraping mit Selenium
- Demo Formular ausfüllen mit Selenium
- Ausblick: Web Crawling mit scrapy
- Fragen (Q&A)

Digitalisierung

Der Begriff **Digitalisierung** bezeichnet ursprünglich das Umwandeln von analogen Werten in digitale Formate. Die so gewonnenen Daten lassen sich informationstechnisch verarbeiten, ein Prinzip, das allen Erscheinungsformen der Digitalen Revolution (die heute zumeist gemeint ist, wenn von *Digitalisierung* die Rede ist) im Wirtschafts-, Gesellschafts-, Arbeits- und Privatleben zugrunde liegt.

API

Eine **Programmierschnittstelle** (auch **Anwendungsschnittstelle**, genauer *Schnittstelle zur Programmierung von Anwendungen*), häufig nur kurz **API** genannt (von englisch *application programming interface*, wörtlich ‚**Anwendungsprogrammierschnittstelle**'), ist ein Programmteil, der von einem Softwaresystem anderen Programmen zur Anbindung an das System zur Verfügung gestellt wird. Im Gegensatz zu einer Binärschnittstelle (ABI) definiert eine Programmierschnittstelle nur die Programmanbindung auf Quelltext-Ebene. Zur Bereitstellung solch einer Schnittstelle gehört meist die detaillierte Dokumentation der Schnittstellen-Funktionen mit ihren Parametern auf Papier oder als elektronisches Dokument.

Standard -> REST-API / GET | POST | PATCH/UPDATE | DELETE (REpresentational State Transfer)
Cutting Edge -> GraphQL

Web Scraping

Web Scraping, Web Harvesting oder Web-Datenextraktion ist Daten-Scraping, das zum Extrahieren von Daten von Websites verwendet wird. Web Scraping-Software kann direkt über das Hypertext Transfer Protocol oder über einen Webbrowser auf das World Wide Web zugreifen. Während Web-Scraping manuell von einem Software-Benutzer durchgeführt werden kann, bezieht sich der Begriff in der Regel auf automatisierte Prozesse, die mit einem Bot oder Web-Crawler implementiert werden. Hierbei handelt es sich um eine Form des Kopierens, bei der bestimmte Daten gesammelt und aus dem Web kopiert werden, normalerweise in eine zentrale lokale Datenbank oder Kalkulationstabelle, um sie später abzurufen oder zu analysieren.

Web Scraping

Neuere Formen des Web Scraping beinhalten das Abhören von Datenfeeds von Webservern. Beispielsweise wird JSON häufig als Transportspeichermechanismus zwischen dem Client und dem Webserver verwendet.

JSON

Die JavaScript Object Notation, kurz **JSON** [ˈdʒeɪsən], ist ein kompaktes Datenformat in einer einfach lesbaren Textform zum Zweck des Datenaustauschs zwischen Anwendungen.

Code:

```
{
  "name": "Dieter",
  "alter": 53,
  "verheiratet": true,
  "beruf": "Developer"
}
```

Automatisierungstools

OSX
Automator

Windows
AutoIt, AutoHotkey

Linux
Actiona

Nachteil: proprietär, plattform abhängig

Warum Python?

Python ist eine moderne, objektorientierte Programmiersprache.

Sie wurde wurde Anfang der 1990er Jahre von Guido van Rossum entwickelt und nach der berühmten Comedy-Truppe "Monty Pythons Flying Circus" benannt (und nicht nach der - ebenfalls berühmten - Schlange).

Warum Python?

Python als Programmiersprache ist aus mehreren Gründen attraktiv. Zum einen lassen sich Programme in Python in der Regel bedeutend schneller entwickeln als in traditionellen Programmiersprachen wie C(++), Pascal oder Java. Zum andern ist es eine echte plattformunabhängige Sprache, die auf fast allen Betriebssystemen vorhanden ist. Neben den Versionen für die "großen" Betriebssysteme Windows, Macintosh und Linux gibt es auch Versionen für OS/2, BeOS und Rasperry Pi etc.
Mit Python lassen sich schnell und einfach kleine Programme entwickeln. Aber auch für große Entwicklungen ist Python aufgrund seiner Objektorientierung gut geeignet. Daher kann Python als »Rapid Application Development«-Tool (RAD) ebenso eingesetzt werden wie als Scriptsprache zur Systemverwaltung.

Warum Python?

Python wurde 2018 zur Programmiersprache des Jahres gekürt und landete im März 2019 auf Platz drei!

Python bietet tausende Libraries, Beispiele und Online Kurse und kann sich damit gerade für Künstliche Intelligenz- und Big-Data-Anwendungen als alternative Programmiersprache lohnen. Und es existiert eine grafische Oberfläche (GUI) namens Tkinter, die auf Mac und PC zum Lieferumfang des Programmierpakets gehört.

Darum Python!

Mit Python geht fast überall alles!

www.meetup.com/de-DE/Python-User-Group-Nuremberg

Python installieren

https://www.python.org/downloads Download Python 3.7.4

! Windows Path

PATH ist die Systemvariable, die das Betriebssystem verwendet, um über die Befehlszeile oder das Terminalfenster nach erforderlichen ausführbaren Dateien zu suchen.

Windows 10 und Windows 8

1. Suchen Sie in der Suche, und wählen Sie dann: System (Systemsteuerung)
2. Klicken Sie auf den Link **Erweiterte Systemeinstellungen**.
3. Klicken Sie auf **Umgebungsvariablen**. Suchen Sie im Abschnitt **Systemvariablen** die Umgebungsvariable PATH, und wählen Sie sie. Klicken Sie auf **Bearbeiten**. Wenn die Umgebungsvariable PATH nicht vorhanden ist, klicken Sie auf Neu.

(OSX Alternative Homebrew, Anaconda, Jupiter Notebook)
Editor: MS Visual Studio Code, Jetbrains PyCharm Community Version

Python PIP

pip ist das Paketinstallationsprogramm für Python. Du kannst pip verwenden, um Pakete aus dem Python-Paketindex und anderen Indizes zu installieren.

cmd/$>
pip install beautifulsoup4
pip install pandas
pip install tabulate
(lxml, requests, re, selenium, matplotlib, twython, json…)

Beispiel Hands on - Intro

> git clone
https://github.com/sherpadee/WebScrapingDemo.git

python hello.py ;)
python tkinterhello.py -> GUI-APP
python twitter.py -> API
python simplescrape.py
-> BeautifulSoup, Internetnutzer Weltweit, json

Beispiel Hands on #nuedigital 1

python test1.py

```
#get page
page =
requests.get("https://nuernberg.digital/festival/programm")

#parse page
soup = BeautifulSoup(page.content, 'html.parser')

print(page.content)
```

Beispiel Hands on #nuedigital 2

python test2.py -> HTML/CSS!

```
#get events details
event_names = soup.select('li.EventList__item .Card__text
h3.EventCard__title')
event_names = [en.get_text() for en in event_names]

#konvert to panda Dataframe
events = pd.DataFrame({"Eventname": event_names})
```

Beispiel Hands on #nuedigital 3

python test3.py

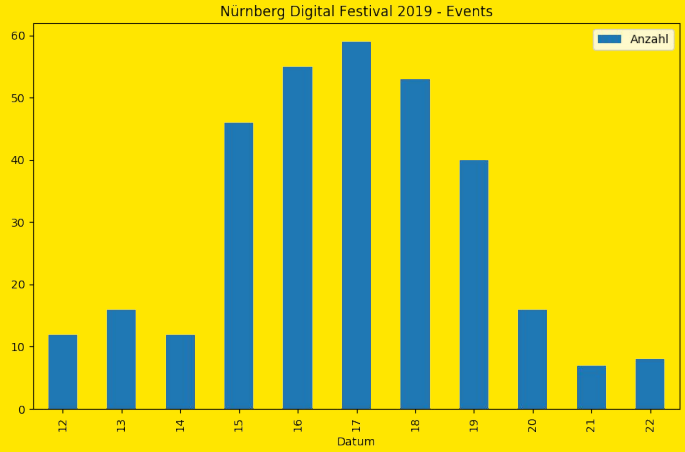
```
#printevents
print(tabulate(events, headers='keys', tablefmt='grid'))

print ("Total: {0} Events".format(len(events)))
```

Beispiel Hands on #nuedigital 4

python nuedigitalevents.py

Mehr Infos, Sortierung, Auswertungen Events Heute, Graph



Web Automatisierung mit Selenium

Selenium ist ein Framework für automatisierte Softwaretests von Webanwendungen, die von einem Programmiererteam der Firma *ThoughtWorks* entwickelt und als freie Software unter der Apache-2.0-Lizenz veröffentlicht wurde. Es ist ein weit verbreitetes Tool und zählt zu den bekanntesten quelloffenen Testwerkzeugen.

Mit Selenium ist es möglich, Interaktionen mit einer Webanwendung aufnehmen zu lassen und diese Tests automatisiert beliebig oft zu wiederholen. Es kann vor allem Entwicklern von Webanwendungen sehr viel Tipparbeit abnehmen – beispielsweise beim Ausfüllen von Webformularen – und macht das Testen dadurch schneller, flexibler und verlässlicher.

Selenium Webdriver

Edge
https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/

Chrome
http://chromedriver.chromium.org/downloads

Firefox
https://github.com/mozilla/geckodriver/releases

Safari
https://webkit.org/blog/6900/webdriver-support-in-safari-10
/usr/bin/safaridriver (Safari Entwicklermodus -> allow)
Achtung: Webdriver Version == Browser Version !

Selenium Locate Elements

- find_element_by_id
- find_element_by_name
- find_element_by_xpath
- find_element_by_link_text
- find_element_by_partial_link_text
- find_element_by_tag_name
- find_element_by_class_name
- find_element_by_css_selector

```
login_form = driver.find_element_by_id('loginForm')
```

Selenium Demo Hands On

python browsertest.py -> check Webdrivers

python unitest.py -> Unit Test Example

python kansas.py -> Gehälter Scrape

Selenium UseCase

Real world customer example

JURADIREKT
Zentrales Vorsorgeregister - Bundesnotarkammer

https://www.vorsorgeregister.de

Web Crawling mit scrapy

https://scrapy.org

```
pip install scrapy
cat > myspider.py <<EOF
import scrapy
class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for title in response.css('.post-header>h2'):
            yield {'title': title.css('a ::text').get()}

        for next_page in response.css('a.next-posts-link'):
            yield response.follow(next_page, self.parse)
EOF
scrapy runspider myspider.py
```

Schlusswort

Hacking with Python?
Don’t walk on the Dark Side, there are no cookies ;-)

Vielen Dank für Eure Aufmerksamkeit!

Fragen? Q&A

Brezen, Drinks

Aufkleber, Blocks

Quatschen/Netzwerken

Links

https://www.python.org/
https://www.meetup.com/de-DE/Python-User-Group-Nuremberg/
https://www.hackerrank.com/domains/python

https://www.crummy.com/software/BeautifulSoup/bs4/doc/
https://selenium-python.readthedocs.io/

https://stackabuse.com/accessing-the-twitter-api-with-python/

https://pythonprogramminglanguage.com/web-scraping-with-pandas-and-beautifulsoup/
https://www.freecodecamp.org/news/better-web-scraping-in-python-with-selenium-beautiful-soup-and-pandas-d6390592e251/
https://towardsdatascience.com/time-series-analysis-and-climate-change-7bb4371021e
https://scrapy.org