

The background of the slide is a photograph showing a close-up of a hand using a wooden scraper to peel off layers of white paint from a surface. The paint is chipped and flaking, revealing a lighter color underneath. The hand is positioned on the right side of the frame, with the scraper angled towards the left.

Web Scraping

Wat, waarom, waarmee, en hoe?

SAI avondconferentie 8 maart 2018
Prof. dr. Seppe vanden Broucke

Overzicht

- Introductie: wat is web scraping, use cases
- Technologie: programmeertalen, pakketten, commerciële tools
- Beheersaspecten: automatisering, onderhoud, vermijden en ontwijken, legale bekommernissen
- Q & A



Seppe vanden Broucke

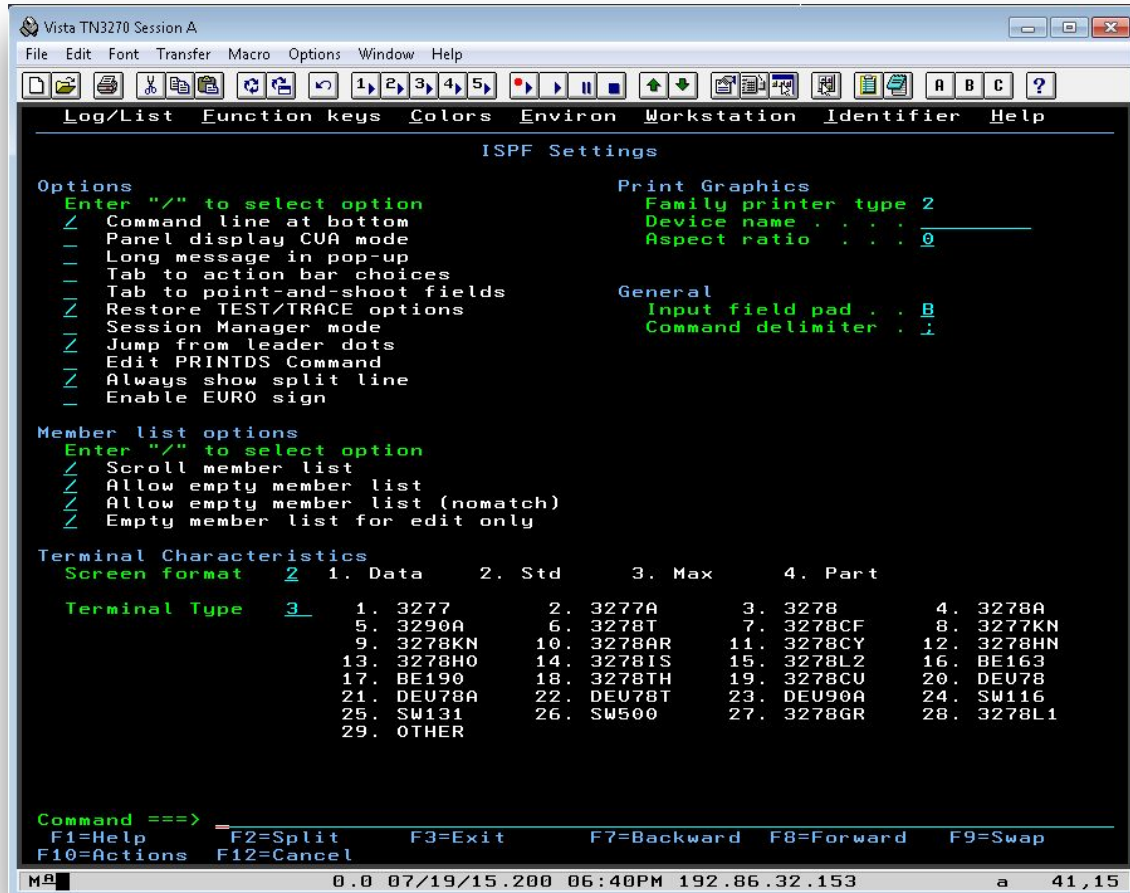


- Assistant professor
 - Decision Sciences and Information Management, KU Leuven (Belgium)
 - PhD in Applied Economics, KU Leuven, Belgium in 2014
 - Titel: Advances in Process Mining: Artificial Negative Events and Other Techniques
- Research: data en process science, profit-driven analytics, fraud analytics
- Contact:
 - www.seppe.net
 - seppe.vandenbroucke@kuleuven.be
 - www.dataminingapps.com

Introductie

Wat is web scraping?

- Web scraping = web harvesting = web data extraction = web data mining
 - Het bouwen van een **software agent** om **gegevens** van het web te **downloaden, ontleden, en organiseren** in een **geautomatiseerde opstelling**
 - Of: in plaats van zelf in een web browser rond te surfen en interessante gegevens te kopiëren-plakken: laat het door een computer doen
- Van ongestructureerde data op het web -> naar gestructureerde data



Wat is web scraping?

- Web scraping is zo oud als het Internet zelf
- En zelfs eerder: “screen scraping”
- Van “scraping” naar “robotic process automation”?



Ui

UiPath Terminal Wizard

Save & Exit

Recording

Stop

Set Text

Get Text

Wait Text

Actions

Type

Refresh

Show Recorded Actions

Show Field Properties

Options

```

>> welcome to the best known public AS/400| <<

System name: PUB1      subsystem : QINTER      devicename: QPADEV001C
Username/Benutzer: _   Password :              (enter blind)

/.....\
| Rechenzentrum Kreuznach |
|                           |
| Holger L. Scherer        |
| Jungstrasse 25-27        |
| 55543 Bad Kreuznach      |
| Germany                  |
| support400@RZKH.de       |
| Tel. +49-671-9203029     |
| iSeries-Hosting, ERP-Software, |
| advanced data processing  |
\...../

*** IMPORTANT NOTE ***

* Server PUB1 will be closed soon.
  Please create save files and store
  them offline.

* english AS/400 forum: www.code400.com

welcome|

powered by www.RZKH.de

```

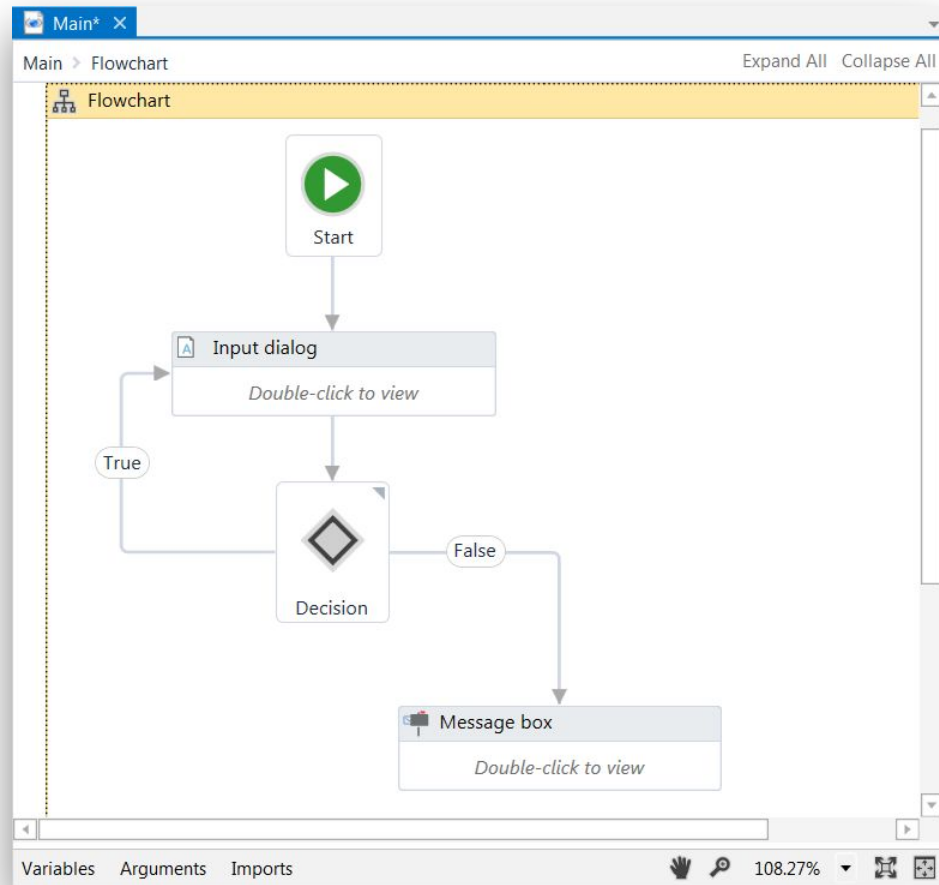
Recorded Actions

Field Properties

RowStart	3
ColStart	15
Text	PUB1
LabeledBy	System name:
FollowedBy	
Index	1
Editable	False

Connected

Cursor Position: Row 4, Column 22



Design Studio - 9.6.1 - \\vmware-host\Shared Folders\...\epo.org.robot - Minimal Execution (Direct)

File Edit View Debug Tools Settings Window Help

My Projects

- 9.6.1
 - Examples
 - Tutorials
 - amazon_deals.robot
 - amazon_deals.type
 - captcha.robot
 - demo3.robot
 - demo5.robot
 - epo.org-demo1.robot
 - epo.org-demo2.robot
 - epo.org.robot
 - objectdb.py
 - Patent.py
 - PatentQuery.type

Shared Projects

- Management Console (localhost)

Databases

- Management Console (localhost)
- Development Database
- Design Studio Databases
- Development Database

Design

epo.org.robot x demo3.robot x PatentQuery.type x Patent.type x

For Each Tag

Click WIDE TOUCHPAD

Extract Id

Extract Title

Extract Status

Extract Applicants

Set Named Tag

Extract Filing Date

Extract Classification

Extract Screenshot

Store in Database

About this file - European Patent Register

https://register.epo.org/application?number=EP15722891

Patentamt
European
Patent Office
Office européen
des brevets

European Patent Register

out European Patent Register Other EPO online services

Search Quick search Advanced search Search results Help

About this file: WO2015171731

Refine search ST36 Espacenet Submit observations Report error

WO2015171731 - TABLE WITH ELECTRICAL PORTS [Right-click to bookmark this link]

Status The international publication has been made
Database last updated on 21.11.2015

Most recent event 20.11.2015 PCT data prior to European publication

Applicant(s) For all designated states
Apple Inc.
1 Infinite Loop
Cupertino, CA 95014 / US
[NIP]

Inventor(s) 01 / MANDON, Kully
c/o Apple Inc.
1 Infinite Loop

Variables

PatentQuery

Patent

Type: Patent

id WO2015171731

title:

status:

applicants:

View

html body div div[1] div[5] div div[1] table tbody tr[0] td b font

Ready. https://register.epo.org/application?number=EP15722891

Web scraping: waarom

- Het web bevat een weelde aan data
 - Ideale voedingsbodem voor big data en analytics
 - Maar: ongestructureerd
 - Web browsers zijn goed in het weergeven van deze ongestructureerde “soep”, maar hoe krijgen we dit gestructureerd?
 - Voor aanleggen en verrijken van gegevens, automatiseren van processen, binnenhalen van open data bronnen, ...

Web scraping: waarom

- “Waarom niet via een API?”
 - Twitter, Facebook, LinkedIn, and Google
 - Ga eerst voor deze optie
- Maar
 - Website biedt geen API aan
 - De API is kostelijk (en publieke website niet)
 - De API is “rate limited”: slechts n verzoeken per seconde, per dag, per gebruiker, ...
 - De API stelt niet alle gewenste gegevens bloot

Web scraping: waarom

- “Moeten we meteen aan het programmeren gaan?”
 - Voor eenmalige, beperkte projecten vermoedelijk niet



WIKIPEDIA
De vrije encyclopedie

Artikel Overleg Lezen

Lijst van Belgische regeringen

Tot 1995 was België een unitaire staat. In dat jaar werd de derde fase afgerond van een staats hervorming die was ingezet in 1970 en werd België formeel een federale staat. De o Belgische regeringen sinds de Belgische Revolutie. De kleur bij de nummers geeft aan van welke partij de premier is.

Nr.	Regering	Premier	Partijen	Van	Tot	Dagen
-	Voorlopig Bewind	Louis de Potter	Unionistisch	24 september 1830	25 februari 1831	154
1	De Gerlache	Etienne de Gerlache	Unionistisch	26 februari 1831	23 maart 1831	25
2	Lebeau I	Joseph Lebeau	Liberalen	23 maart 1831	21 juli 1831	120
3	De Muelenaere	Felix de Muelenaere	Unionistisch	26 juli 1831	17 september 1832	419
4	Goblet d'Alviella	Albert Goblet d'Alviella	Unionistisch	20 oktober 1832	1 augustus 1834	650
5	De Theux de Meylandt I	Barthélémy de Theux de Meylandt	Unionistisch	4 augustus 1834	6 april 1840	2072
6	Lebeau II	Joseph Lebeau	Liberalen	10 april 1840	13 juni 1841	360
7	Nothomb	Jean-Baptiste Nothomb	Unionistisch	13 juni 1841	19 juni 1845	1528
8	Van de Weyer	Sylvain Van de Weyer	Unionistisch	30 juli 1845	2 maart 1846	215
9	De Theux de Meylandt II	Barthélémy de Theux de Meylandt	Katholieken	31 maart 1846	12 augustus 1847	499
10	Rogier I	Charles Rogier	Liberalen	12 augustus 1847	28 september 1852	1874
11	De Brouckère	Henri de Brouckère	Liberalen	31 oktober 1852	2 maart 1855	852



	A	B	C	D	E	F	G
1	Nr.	Regering	Premier	Partijen	Van	Tot	Dagen
2	-	Voorlopig Bewind	Louis de Potter	Unionistisch	24 september 1830	25 februari 1831	154
3	1	De Gerlache	Etienne de Gerlache	Unionistisch	26 februari 1831	23 maart 1831	25
4	2	Lebeau I	Joseph Lebeau	Liberalen	23 maart 1831	21 juli 1831	120
5	3	De Muelenaere	Felix de Muelenaere	Unionistisch	26 juli 1831	17 september 1832	419
6	4	Goblet d'Alviella	Albert Goblet d'Alviella	Unionistisch	20 oktober 1832	1 augustus 1834	650
	5	De Theux de Meylandt I	Barthélémy de Theux de Meylandt	Unionistisch	4 augustus 1834	6 april 1840	2072

Voorbeelden

- “We Feel Fine, Jonathan Harris en Sep Kamvar
 - Web scraper voor blogs om op zoek te gaan naar zinnen beginnende met “I feel”



Voorbeelden

- “Using big data to predict suicide risk among Canadian youth”, SAS
 - Web scraper blogs, Twitter, en andere sociale media

Team members Horne, plus Tim Trussell, Manager Presales Specialist, Data Sciences, both of whom have health care backgrounds, and data scientists Marie Soehl and Jos Polfliet, who did the programming, collected 2.3 million tweets and used text mining software to identify 1.1 million of them as likely to have been authored by 13 to 17 year olds in Canada by building a machine learning model to predict age, based on the open source PAN author profiling dataset. Their analysis made use of natural language processing, predictive modelling, text mining, and data visualization.

Voorbeelden

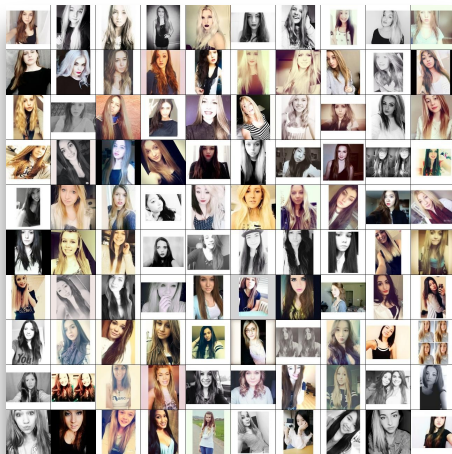
- “The Billion Prices Project: Using Online Prices for Measurement and Research”
 - <http://www.nber.org/papers/w22111>

Web scraping was used to collect a data set of online price information which was used to construct a robust daily price index for multiple countries.

Voorbeelden

- “Predicting selfie success”
 - <http://karpathy.github.io/2015/10/25/selfie/>

Train a deep learning model based on scraped images from Tinder and Instagram together with their “likes” to predict whether an image would be deemed “attractive”.



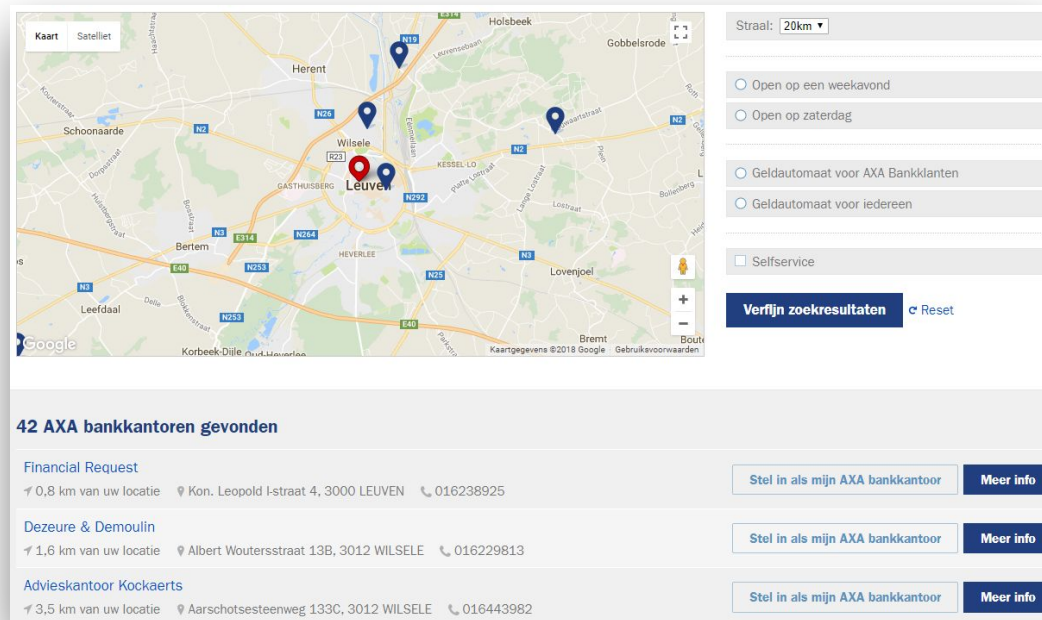
Voorbeelden

- Lyst
 - <http://talks.lystit.com/dsl-scraping-presentation/>

A London based online fashion marketplace, scraped the web for semi-structured information about fashion products and then applied machine learning to present this information cleanly and elegantly for consumers from one central website.

Voorbeelden

- Competitor analytics



The screenshot displays a web application for competitor analytics. At the top, a map of Leuven, Belgium, is shown with a red pin at the city center and several blue pins indicating the locations of AXA bank branches. To the right of the map, there are filters for the search radius (set to 20km) and opening hours (options for weekends and Saturdays). Below the map, a list of 42 AXA bank branches is displayed. Each entry includes the branch name, distance from the user's location, address, and phone number. There are also buttons to 'Stel in als mijn AXA bankkantoor' (Set as my AXA branch) and 'Meer info' (More info) for each branch.

42 AXA bankkantoren gevonden

Branch Name	Distance	Address	Phone Number	Actions
Financial Request	0,8 km van uw locatie	Kon. Leopold I-straat 4, 3000 LEUVEN	016238925	Stel in als mijn AXA bankkantoor Meer info
Dezeure & Demoulin	1,6 km van uw locatie	Albert Woutersstraat 13B, 3012 WILSELE	016229813	Stel in als mijn AXA bankkantoor Meer info
Advieskantoor Kockaerts	3,5 km van uw locatie	Aarschotsesteenweg 133C, 3012 WILSELE	016443982	Stel in als mijn AXA bankkantoor Meer info

Voorbeelden

- HR analytics

The San Francisco based hiQ startup specializes in selling employee analyses by collecting and examining public profile information, for instance from LinkedIn (who was not happy about this but was so far unable to prevent this practice following a court case).

<https://www.bloomberg.com/news/features/2017-11-15/the-brutal-fight-to-mine-your-data-and-sell-it-to-your-boss>



Voorbeelden

- Finance

Scouring emails for such nuggets of information is just the tip of the “alternative data” iceberg. Almost everything we do leaves a digital fingerprint, which can be scraped, aggregated and sold to investment firms looking for tradable signals, or to use the jargon — market-beating “alpha”.

<https://www.ft.com/content/08a22da8-b587-11e6-ba85-95d1533d9a62>



Voorbeelden

- Fraud analytics



Voorbeelden

- Web scraping voor...
 - Downloaden en analyseren jaarrekeningen Nationale Bank en Kruispuntbank Ondernemingen voor voorspellen vroeg faillissement
 - Scrapen van job sites om trends binnen data science tooling te analyseren
 - Scrapen nieuwsartikels om sentiment analyse rond Bitcoin te verrichten
 - Scrapen nieuwsartikels om biased versus unbiased nieuwsbronnen te analyseren
 - Scrapen Amazon recensies voor sentiment analyse
 - Scrapen filmsite voor bouwen recommender systeem

Technologie

Bouwstenen

- HTTP client
 - Ondersteuning voor GET, POST en andere methods
 - Ondersteuning voor URL parameters
 - Ondersteuning POST data
 - Ondersteuning wijzigen headers
 - Ondersteuning voor cookies
 - Extra's: meesturen JSON data, HTTP auth, file uploads
 - Op programmeer/eur-vriendelijke wijze

Bouwstenen

GET /docs/index.html HTTP/1.1

Host: www.example.com

Accept: image/gif, image/jpeg, */*

Accept-Language: en-us

Accept-Encoding: gzip, deflate

User-Agent: Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)

Request method, url, versie

Request headers (host is verplicht)

Blanko lijn

HTTP/1.1 200 OK

Date: Sat, 3 Mar 2018 08:56:53 GMT

Server: Apache/2.2.14 (Linux)

Last-Modified: Sat, 20 Nov 2017 07:16:26 GMT

Accept-Ranges: bytes

Content-Length: 44

Connection: close

Content-Type: text/html

Versie, status code en bericht

Response headers

Blanko lijn

Response content

<html><body><h1>It works!</h1></body></html>

Bouwstenen

- HTML parser
 - Niet: reguliere expressies
 - Van HTML soup naar een georganiseerde structuur
 - Opzoeken via CSS selector queries, HTML-boom, XPath
- Andere:
 - Javascript parser
 - Databank
 - ...

Overzicht

- Doe-het-zelf
 - Python
 - Andere: Java, R, JavaScript
- Commerciele tools

Python

- HTTP client
 - urllib (standaard module) (urllib en urllib2 in Python 2): nogal vervelend en “verbose” in gebruik (sommige functies nog steeds handig, echter)
 - httplib2, urllib3: goede alternatieven, maar ook vrij “low level” met heel wat “boilerplate”
 - **requests: de kampioen!**
 - grequests of aiohttp: voor wie het echt asynchronous moet zijn

Python

```
import urllib.parse
import urllib.request

url = 'http://www.webscrapingfordatascience.com/postform2/'

formdata = {
    'name': 'Seppe',
    'gender': 'M',
    'pizza': 'like',
    'haircolor': 'brown',
    'comments': ''
}

data = urllib.parse.urlencode(formdata).encode("utf-8")
req = urllib.request.Request(url, data)
response = urllib.request.urlopen(req)
text = response.read()

print(text)
```

```
import requests

url = 'http://www.webscrapingfordatascience.com/postform2/'

formdata = {
    'name': 'Seppe',
    'gender': 'M',
    'pizza': 'like',
    'haircolor': 'brown',
    'comments': ''
}

text = requests.post(url, data=formdata).text

print(text)
```

Python



- “HTTP for Humans”
- requests
 - Zeer populair, uitstekende documentatie, goede snelheid
 - <http://docs.python-requests.org/en/master/>
 - Uitgebreide functionaliteit: Keep-Alive, sessies voor eenvoudig cookie-beheer, SSL ondersteuning, automatische content decoding, HTTP auth ondersteuning, gzip ondersteuning, proxy ondersteuning, Unicode by default, streaming downloads

Python



- HTML parser
 - BeautifulSoup: een API bovenop een parser-engine (lxml, html5lib, of de ingebouwde html.parser)
 - Vrij makkelijk in gebruik, maar komt wel met enkele gotcha's (selectie op basis van CSS selectors, combinatie van meerdere filters)
 - Overhead zorgt ervoor dat men op termijn vaak rechtstreeks lxml of html5lib gebruikt
 - Maar desalniettemin een populair (en prima) keuze
 - <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Python

```
import requests
from bs4 import BeautifulSoup

url = 'https://en.wikipedia.org/w/index.php?title=List_of_Game_of_Thrones_episodes&oldid=802553687'
html_soup = BeautifulSoup(requests.get(url).text, 'html.parser')

episodes = []

ep_tables = html_soup.find_all('table', class_='wikiepisodetable')
for table in ep_tables:
    headers = []
    for header in table.find_all('th'):
        headers.append(header.text)
    for row in table.find_all('tr'):
        values = []
        for col in row.find_all('td'):
            values.append(col.text)
        if values:
            episode_dict = {headers[i]: values[i] for i in range(len(values))}
            episodes.append(episode_dict)
```

Python

```
{'No.\noverall': '1',  
 'No. in\nseason': '"Winter Is Coming"',  
 'Title': 'Tim Van Patten',  
 'Directed by': 'David Benioff & D. B. Weiss',  
 'Written by': 'April\xa017,\xa02011\xa0(2011-04-17)',  
 'Original air date': '2.22[20]'}
```

```
{'No.\noverall': '2', 'No. in\nseason': '"The Kingsroad"', 'Title': 'Tim Van Patten', 'Directed by': 'David Benioff & D. B. Weiss',  
 'Written by': 'April\xa024,\xa02011\xa0(2011-04-24)', 'Original air date': '2.20[21]'}
```

...

Python

```
import requests
from bs4 import BeautifulSoup

url = 'https://en.wikipedia.org/w/index.php?title=List_of_Game_of_Thrones_episodes&oldid=802553687'
html_soup = BeautifulSoup(requests.get(url).text, 'html.parser')

episodes = []

ep_tables = html_soup.find_all('table', class_='wikiepisodetable')
# of: ep_tables = html_soup.select('table.wikiepisodetable')

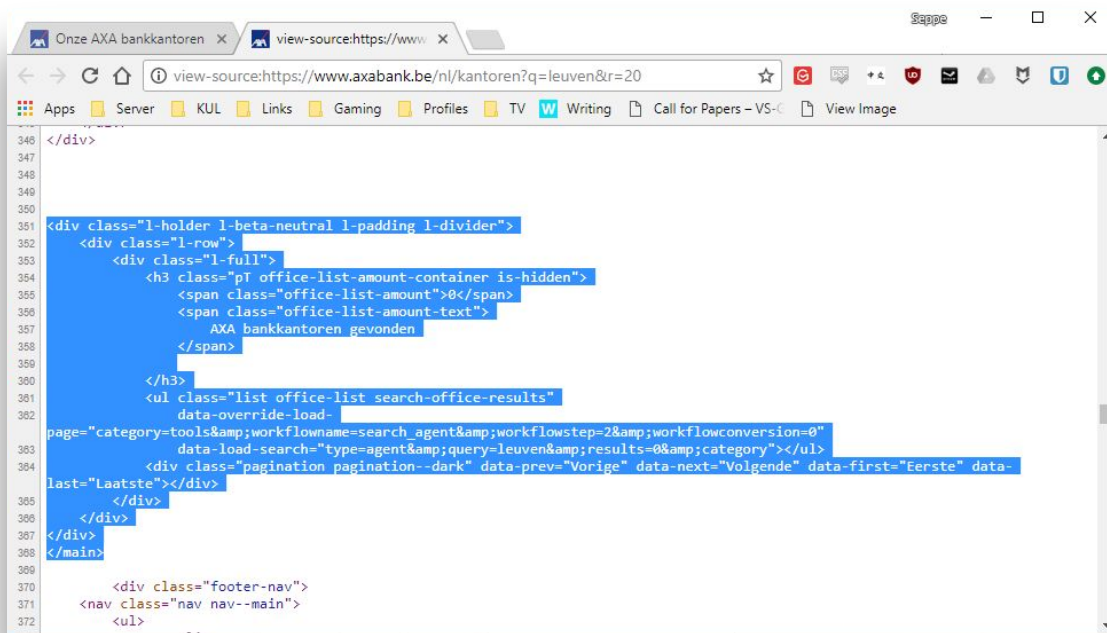
for table in ep_tables:
    headers = []
    for header in table.find_all('th'):
        headers.append(header.text)
    for row in table.find_all('tr'):
        values = []
        for col in row.find_all('td'):
            values.append(col.text)
        if values:
            episode_dict = {headers[i]: values[i] for i in range(len(values))}
            episodes.append(episode_dict)
```

Python

- Meer HTML parsers
 - Reguliere expressies: voor de extra opschoning en extractie (!)
 - parse (<https://pypi.python.org/pypi/parse>): “het omgekeerde van format()”:
 - `search('Age: {d}\n', 'Name: Rufus\nAge: 42\nColor: red\n')`
 - pyquery (<http://pyquery.readthedocs.io/en/latest/>): “opzoeken in jQuery-stijl”
 - parse (<https://parsel.readthedocs.io/en/latest/>): “opzoeken met XPath en CSS selectors”
 - ... in combinatie met BeautifulSoup: voor uitzuiveren gevonden informatie
 - ... of met lxml, html5lib

Python

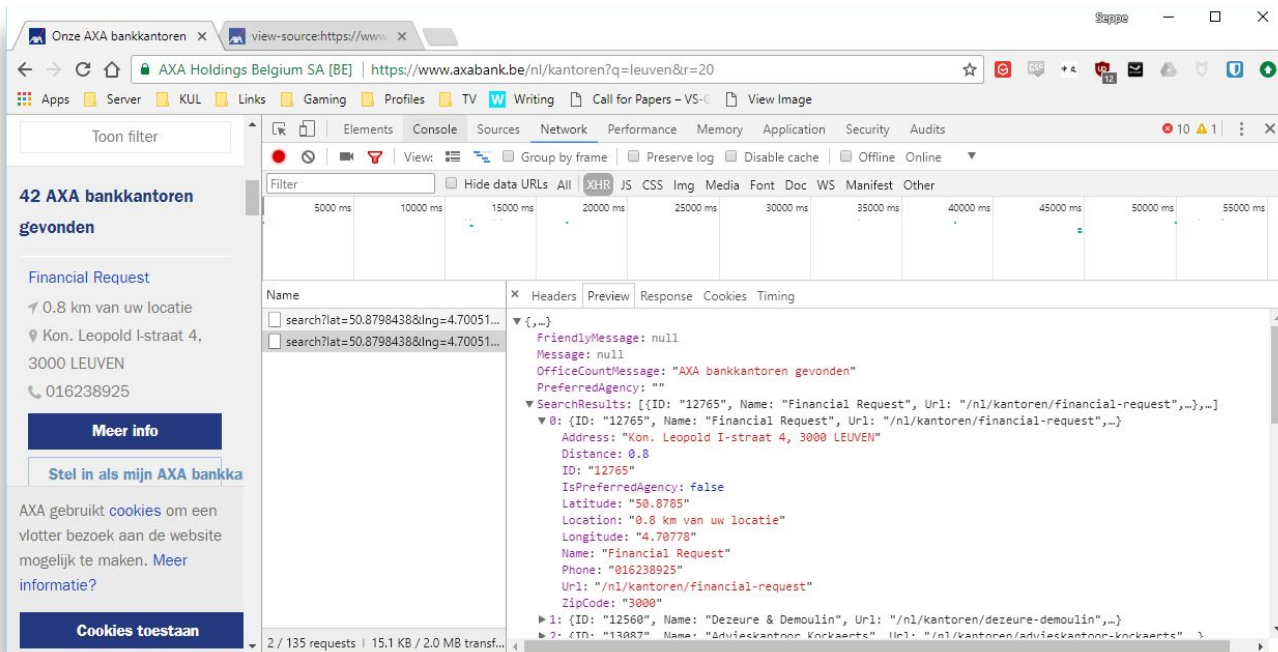
- JavaScript zorgt voor problemen



```
346 </div>
347
348
349
350
351 <div class="l-holder l-beta-neutral l-padding l-divider">
352   <div class="l-row">
353     <div class="l-full">
354       <h3 class="pT office-list-amount-container is-hidden">
355         <span class="office-list-amount">0</span>
356         <span class="office-list-amount-text">
357           AXA bankkantoren gevonden
358         </span>
359       </h3>
360       <ul class="list office-list search-office-results"
361         data-override-load-
362         page="category=tools&workflowname=search_agent&workflowstep=2&workflowconversion=0"
363         data-load-search="type=agent&query=leuven&results=0&category"></ul>
364       <div class="pagination pagination--dark" data-prev="Vorige" data-next="Volgende" data-first="Eerste" data-
365         last="Laatste"></div>
366     </div>
367   </div>
368 </main>
369
370 <div class="footer-nav">
371   <nav class="nav nav--main">
372     <ul>
```

Python

- Of juist niet?



Python

```
import requests
import json
```

Zelfs geen HTML parser nodig, bedankt JavaScript!

```
url = 'https://www.axabank.be/api/nl/agency/search'
```

```
params = {'lat': '50.8798438', 'lng': '4.700517600000012', 'r': '20'}
```

```
r = requests.get(url, params=params)
data = json.loads(r.text)
```

```
print(data['SearchResults'][:3])
```

```
[{'ID': '12765', 'Name': 'Financial Request', 'Url': '/nl/kantoren/financial-request', 'Location': '0.8 km van uw locatie',
'Address': 'Kon. Leopold I-straat 4, 3000 LEUVEN', 'Phone': '016238925', 'Latitude': '50.8785', 'Longitude': '4.70778', 'Distance':
0.8, 'IsPreferredAgency': False, 'ZipCode': '3000'},
 {'ID': '12560', 'Name': 'Dezeure & Demoulin', 'Url': '/nl/kantoren/dezeure-demoulin', 'Location': '1.6 km van uw locatie',
'Address': 'Albert Woutersstraat 13B, 3012 WILSELE', 'Phone': '016229813', 'Latitude': '50.894', 'Longitude': '4.69993', 'Distance':
1.6, 'IsPreferredAgency': False, 'ZipCode': '3012'},
 {'ID': '13087', 'Name': 'Advieskantoor Kockaerts', 'Url': '/nl/kantoren/advieskantoor-kockaerts', 'Location': '3.5 km van uw
locatie', 'Address': 'Aarschotsesteenweg 133C, 3012 WILSELE', 'Phone': '016443982', 'Latitude': '50.9092', 'Longitude': '4.71277',
'Distance': 3.5, 'IsPreferredAgency': False, 'ZipCode': '3012'}]
```

Python

- Maar het is niet altijd zo simpel
 - Wanneer JavaScript te complex wordt
 - Complexe cookie-mechanismes die te lastig worden om te reverse-engineeren
 - Complexe login-portalen
 - Browser checks
 - Scraping gebaseerd op de “render” van de website (b.v. screenshots)
 - ... dan moeten we een volledige browser aansturen



- Selenium
 - “Web browser automation framework”
 - <https://www.seleniumhq.org/>
 - Meer dan alleen Python: Java, C#, R, Ruby, ...
 - En met verschillende browsers (Chrome, Firefox, ...)
 - Krachtig, maar komt met een serieuze overhead
 - Iets moeilijker onder de knie te krijgen
 - Niet langer HTTP en HTML, maar: “zoek een veld, vul dit in, klik met de muis, ...”

Python

```
from selenium import webdriver
from selenium.webdriver.support.select import Select
from selenium.webdriver.common.keys import Keys

url = 'http://www.webscrapingfordatascience.com/postform2/'

driver = webdriver.Chrome()
driver.implicitly_wait(10)
driver.get(url)

driver.find_element_by_name('name').send_keys('Seppe')
driver.find_element_by_css_selector('input[name="gender"][value="M"]').click()
driver.find_element_by_name('pizza').click()
driver.find_element_by_name('salad').click()
Select(driver.find_element_by_name('haircolor')).select_by_value('brown')
driver.find_element_by_name('comments').send_keys(['First line', Keys.ENTER, 'Second line'])

driver.find_element_by_tag_name('form').submit()
# Of: driver.find_element_by_css_selector('input[type="submit"]').click()

input('Press ENTER to close the automated browser')
driver.quit()
```

Python

- Andere opties
 - Selenium Studio: een IDE voor het bouwen van scrapers (verouderd en vrij onbruikbaar)
 - MechanicalSoup (<http://mechanicalsoup.readthedocs.io/en/stable/>) als vervanging voor het oudere Mechanize: maar ondersteund geen JavaScript
 - Scrapy (<https://scrapy.org/>): uitgebreid pakket voor het schrijven en beheren van crawlers
 - Ondersteuning voor logging, deployment in the “Scrapy Cloud”
 - JavaScript ondersteunt via “Splash”, maar lastiger om aan de praat te krijgen
- requests + BeautifulSoup / lxml / pyquery of Selenium blijven aan te raden keuzes

Andere talen

- Andere talen

- R: rvest of Selenium
- C# en Java: Selenium
- JavaScript: PhantomJS (kan gebruikt worden als “browser” in combinatie met Selenium), Nightmare, SlimerJS, CasperJS
- Doch nu grotendeels vervangen door Puppeteer

(<https://github.com/GoogleChrome/puppeteer>): high-level JavaScript API voor headless Chrome

- <https://intoli.com/blog/running-selenium-with-headless-chrome/>
- <https://pypi.python.org/pypi/pyppeteer>
- <https://github.com/ariya/phantomjs/issues/15344>: Archiving the project: suspending the development
- Gelijkaardig project voor Firefox bestaat
- Development met Selenium vaak met de “gewone Chrome”, dan in productie naar Puppeteer

Andere tools

- Command line tools
 - <https://httpie.org/>
 - <https://curl.haxx.se/>
- Proxy tools
 - <http://www.squid-cache.org/>
 - <https://www.telerik.com/fiddler>
- Caching
 - <http://cachecontrol.readthedocs.io/en/latest/>

Commerciële tools

- Browser plugins
 - <https://portia.scrapinghub.com>
 - <https://www.parsehub.com>
 - Vaak zeer beperkt, goed voor simpele one-off scraping opdrachten, maar kan ook in enkele regels code, geen ondersteuning voor moeilijkere workflows
- Complete pakketten
 - <https://www.kofax.com/data-integration-extraction>
 - <http://www.fminer.com/>
 - <https://dexi.io/>
 - Vaak aantrekkelijk in termen van opstellen workflow
 - Maar: duurder, workflows vereisen evenzeer onderhoud, gebruikte browser-engines niet altijd even robuust, wanneer websites geavanceerde detectie-mechanismes toepassen onmogelijk om te omzeilen, en onderliggend worden vaak selectie-regels aangemaakt die erg precies en dus zeer gevoelig zijn t.o.v. website-wijzigingen
 - **Maturiteit team en visie in overweging nemen!**

Commerciële tools

portia.scrapinghub.com/#/projects/13359/allrecipes.com/recipe/items

allrecipes.com

search by ingredient recipes » videos » holidays » thebuzz » magazine »

RECIPE BOX SHOPPING LISTS MENU PLANNER COOKING SCHOOL Go Pro! Sign In or Sign Up

This page requires that scripting be enabled in your browser to function properly. If you need help enabling scripting, please visit our technical FAQ page.

Home > Recipes > Breakfast and Brunch > Eggs > Frittata

39 Photos

Spinach and Potato Frittata

READY IN 30 mins

★★★★★ Read Reviews (390)

Recipe by CHERYL H.

"This dish is not only delicious, but easy to make. I make this for Saturday family brunches and get together. It's a big hit."

Recipe Box Shopping List Menu Email Print

Ingredients Edit and Save

Original recipe makes 6 servings Change Servings

- ☐ 2 tablespoons olive oil
- ☐ 6 small red potatoes, sliced
- ☐ 1 cup torn fresh spinach
- ☐ 2 tablespoons sliced green onions
- ☐ 1 teaspoon crushed garlic
- ☐ salt and pepper to taste
- ☐ 6 eggs
- ☐ 1/3 cup milk
- ☐ 1/2 cup shredded Cheddar cheese

Check All Add to Shopping List

Watch video tips and tricks

Crustless Spinach Quiche Potato, Sausage, & Egg Breakfast

Menu Planner FREE Trial Start Now

In Season

Tasty Easter Leftovers

Egg Salad Just Like Mom's

Subscribe Today! Only \$7.99

Related Videos

Potato and Pepper Frittata

Bacon and Potato Frittata with Greens

Potato and Chorizo Mini Quiches

Items

Field	Type	Required	Vary
url	text	☒	☒
author	text	☒	☒
cook_time	number	☒	☒
image	image	☒	☒
Ingredients	safe html	☒	☒
name	text	☒	☒
prep_time	number	☒	☒

+ Field

+ Item Save changes Discard Changes

Commerciële tools

The screenshot displays the Design Studio 9.6.1 interface, which is used for creating and executing workflows. The main window shows a workflow diagram with the following steps: For Each Tag, Click VIDEO TOUCHPAD, Extract ID, Extract Title, Extract Status, Extract Applicants, Set Named Tag, Extract Filing Date, Extract Classification, Extract Screenshot, and Store in Database. The workflow is currently executing, and the results are displayed in the 'Variables' panel on the right. The 'Variables' panel shows a table with columns 'PatentQuery' and 'Patent'. The 'Patent' column contains the following data:

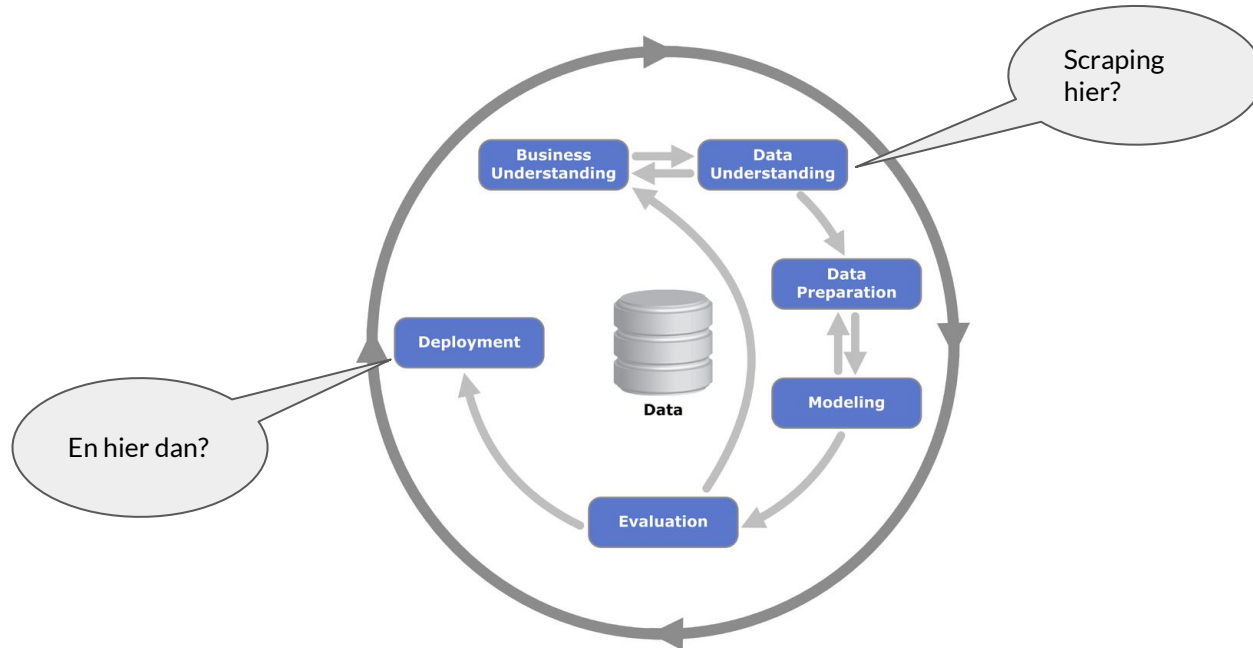
PatentQuery	Patent
id	WO2015171731
title	WO2015171731 - TABLE WITH ELECTRICAL PORTS
status	
applicants	

The 'About this file: European Patent Register' window is also open, showing details for the patent WO2015171731. The 'About this file' section includes the title 'WO2015171731 - TABLE WITH ELECTRICAL PORTS', the status 'The international publication has been made', the database update date '21.11.2015', and the applicant 'INFINITE LOOP CORP. INC. (US)'. The 'Investor(s)' section lists '01 / MANDON, Kelly' and '02 / INFINITE LOOP, Inc.'.

Beheersaspecten

Web scraping als onderdeel data science?

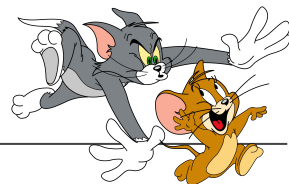
- Model train, of model run?



Web scraping als onderdeel data science?

- Een one-shot project waar scraping van pas kan komen voor verzamelen externe data?
 - Dan is er meestal geen probleem in termen van onderhoud en automatisering
- Of: een predictief model dat in gebruik zal genomen worden?
 - Worden er variabelen gebruikt die verversst moeten worden?
 - Extra druk op het productie-model!
 - Hoe lang blijft de data geldig? Hoe snel moeten we ingrijpen als de data onbeschikbaar wordt? Hoe lang zal het model worden gebruikt? Hoe mission-critical is het model?
 - Data quality: GIGO principle: bouw zoveel als mogelijk cleaning in de scraping-setup
 - Zelfde vragen voor rapportering-gerichte projecten
- Denk niet te snel: “als we maar Facebook’s data hadden”

Het kat-en-muis spel



- Websites gaan bijzonder ver in het afblokken van scrapers
- Ontwijkmechanismes
 - “Fake” zo veel als mogelijk: User Agent, Referer, andere headers, cookies, ... soms diepgaande wijzigingen nodig in code!
 - Maak gebruik van proxy's of de cloud (Google Cloud, AWS): maar niet alle providers houden evenzeer van scrapers!
 - Timing en retry-mechanismen
 - Captcha's: <http://www.deathbycaptcha.com>, OCR systemen, deep learning
 - UI profilers (snelheid van typen, muiskliks, scrollgedrag, ...)
 - Soms is een web server gewoon “raar” (“?a=34&b=35” v.s. “?b=35&a=34”, header volgordes, cookie volgordes, dubbele headers, etc...)

Het kat-en-muis spel

Using CNTK backend

Selected GPU[0] GeForce GTX 980M as the process wide default device.

Train on 1665 samples, validate on 555 samples

Epoch 1/10

32/1665 [.....] - ETA: 36s - loss: 3.0294 - acc: 0.0312

64/1665 [>.....] - ETA: 22s - loss: 5.1515 - acc: 0.0312

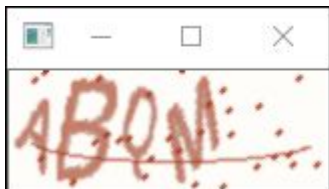
[...]

1600/1665 [=====>..] - ETA: 0s - loss: 7.6135e-04 - acc: 1.0000

1632/1665 [=====>..] - ETA: 0s - loss: 8.3265e-04 - acc: 1.0000

1664/1665 [=====>..] - ETA: 0s - loss: 8.2343e-04 - acc: 1.0000

1665/1665 [=====] - 3s 2ms/step - loss: 8.2306e-04 - acc: 1.0000 - val_loss: 0.3644 - val_acc: 0.9207



Het kat-en-muis spel

- <https://blog.shapesecurity.com/2015/01/22/detecting-phantomjs-based-visitors/>
- <https://intoli.com/blog/making-chrome-headless-undetectable/>
- <http://antoinevastel.github.io/bot%20detection/2017/08/05/detect-chrome-headless.html>

```
if(navigator.plugins.length == 0) {  
    console.log("It may be Chrome headless");  
}  
  
var canvas = document.createElement('canvas');  
var gl = canvas.getContext('webgl');  
  
var debugInfo = gl.getExtension('WEBGL_debug_renderer_info');  
var vendor = gl.getParameter(debugInfo.UNMASKED_VENDOR_WEBGL);  
var renderer = gl.getParameter(debugInfo.UNMASKED_RENDERER_WEBGL);  
  
if(vendor == "Brian Paul" && renderer == "Mesa OffScreen") {  
    console.log("Chrome headless detected");  
}
```

Het kat-en-muis spel



Checking your browser before accessing example.com

This process is automatic. Your browser will redirect to your requested content shortly.

Please allow up to 5 seconds...

[DDoS protection by CloudFlare](#)

Scraping of crawling?

- “We willen een scraper die automatisch op het web op zoek gaat naar informatie over x...”
 - Gewoon links volgen? Maar: moeilijk om op te zetten en te beheren!
 - API's hier een nuttige meerwaarde
 - B.v. voor nieuws: <https://newsapi.org/> en <https://webhose.io/news-api>
 - <http://newspaper.readthedocs.io/en/latest/> en <https://github.com/fhamborg/news-please>
- Elke scraper en crawler is gedoemd om te crashen (site gaat plat, voert wijziging uit, blokkeert de scraper, ...)
 - Vereist grondige codeer-aanpak en structuur

Legale aspecten

- “LinkedIn and Microsoft executives, were understandably not so very happy with this state of affairs. The data belonged to LinkedIn, so they thought, and a cease and desist order was sent out to request hiQ to stop scraping LinkedIn's data, as well as various technical measures were implemented to keep hiQ Labs' bots out.”
- “In 2014, Resultly, another start-up providing a shopping app which constructs a catalog of items for sale by scraping various retailers, accidentally overloaded the servers of the TV retailer QVC, causing outages that, according to QVC, cost them two million dollars in revenue.”
- “In the case of Ticketmaster vs. Riedel Marketing Group (RMG), the latter was web scraping Ticketmaster's site so that it could harvest large quantities of desirable tickets for resale. Ticketmaster argued that RMG had agreed to the terms and conditions of the site but ignored them and the court held that RMG had infringed on Ticketmaster's copyrighted material.”
- “In Ryanair Ltd vs. PR Aviation BV, the European Court found that Ryanair was free to create contractual limits on the use of its database, and the case was hence ruled in its favor.”
- “In 2006, Google got involved in a long-winding legal battle with Belgian media firm Copiepresse. This led to an ugly battle between Copiepresse and Google, ending with the two of them reaching an agreement to include the sites again in 2011.”
- “In Facebook vs. Power Ventures, Facebook also claimed that the defendant has violated the CFAA and the CAN-SPAM Act, a federal law that prohibits sending commercial e-mails with materially misleading information. The judge ruled in favor of the plaintiff.”

Legale aspecten



- Breach of Terms and Conditions (US)
- Copyright or Trademark Infringement (US)
- Computer Fraud and Abuse Act (CFAA) (US)
- Trespass to Chattels (US)
- Robots Exclusion Protocol (industry agreement)
- The Digital Millennium Copyright Act (DMCA), CAN-SPAM Act, ... (US)
- The EU Database Directive of 1996: (2015 ruling by the European Court of Justice greatly strengthened the ability of operators to protect content through contractual terms and conditions when not covered by the database directive)
- The Computer Misuse Act and Trespass to Chattels (UK)
- Computer Misuse Act 1990 (UK)
- General Data Protection Regulation (GDPR) (EU)

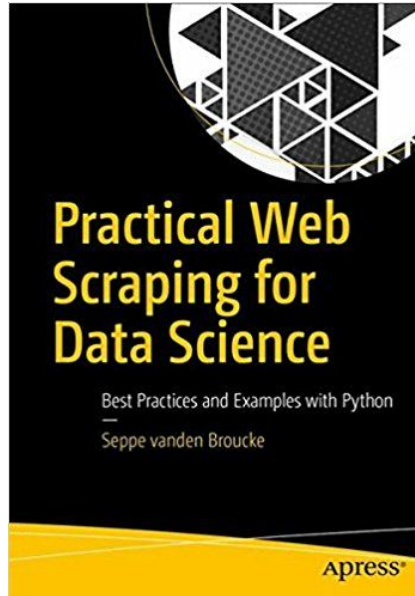
Legale aspecten

- Verkrijg schriftelijke toestemming
- Bekijk de terms of use
- Publieke informatie: geen logins
- Richt geen schade aan
- Let op voor copyright
- Check robots.txt
- Let op met persoonlijke informatie
- ... “het mag, maar geen logins, geen persoonlijke informatie, geen schade, geen beschermde werken, ...”

Uitsmijters

- Ga eerst voor de “nette weg”
- Gebruik established tools (requests, BeautifulSoup, Selenium), commerciële tools minder nuttig
- Diepgaande kennis nodig voor grote, complexe projecten
- Scrapers vereisen onderhoud en beheer
- Neem aan dat het zal crashen
- Kat-en-muis spel
- Wees bewust van de legale grijze zone

Q & A



www.webscrapingfordatascience.com

www.amazon.com/dp/1484235819/