

A Practical Approach to Web Scraping

Erik Hekman

Beyond Notebooks: Research Software Summer School



Agenda

1. Web Scraping in the Social Sciences and Humanities
2. Personal Examples
3. Legality and Ethics of Web Scraping
4. Live coding 1: requests & BeautifulSoup
5. Live coding 2: pagination & building a dataset
6. Live coding 3: dynamic pages & AI-assisted scraping
7. Practical tips



What is web Scraping?



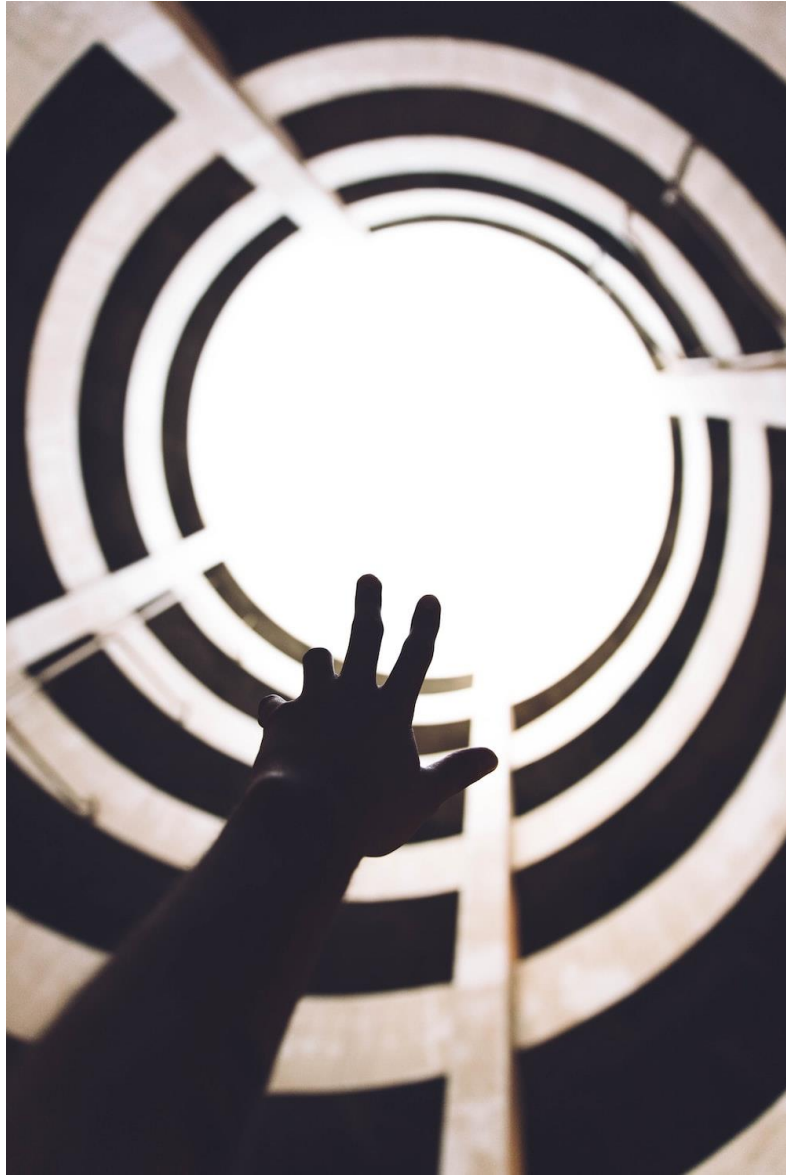
What is web Scraping?

- Web Scraping is defined here as using technology tools for automatic extraction and organization of data from the Web for the purpose of further analysis of this data (Krotov and Tennyson 2018).
- ... scraping is often closer to an art than a science and can take years of practice to master (Possler et al. 2019).



Types of Data

- Structured, semi-structured, and unstructured quantitative and qualitative data distributed in the form of Web pages, HTML tables, Web databases, emails, tweets, blog posts, etc.



Why Web Scraping?

- Overcoming the struggle to find data for research (Munzert et al., 2015)
- provide social researchers with access to data that may otherwise have been difficult or impossible to obtain
- [Web] data provides a granular and real-time representation of numerous processes, relationships, and interactions in the social space (Krotov & Tennyson, 2018).
- Post-API age

A Simple Process

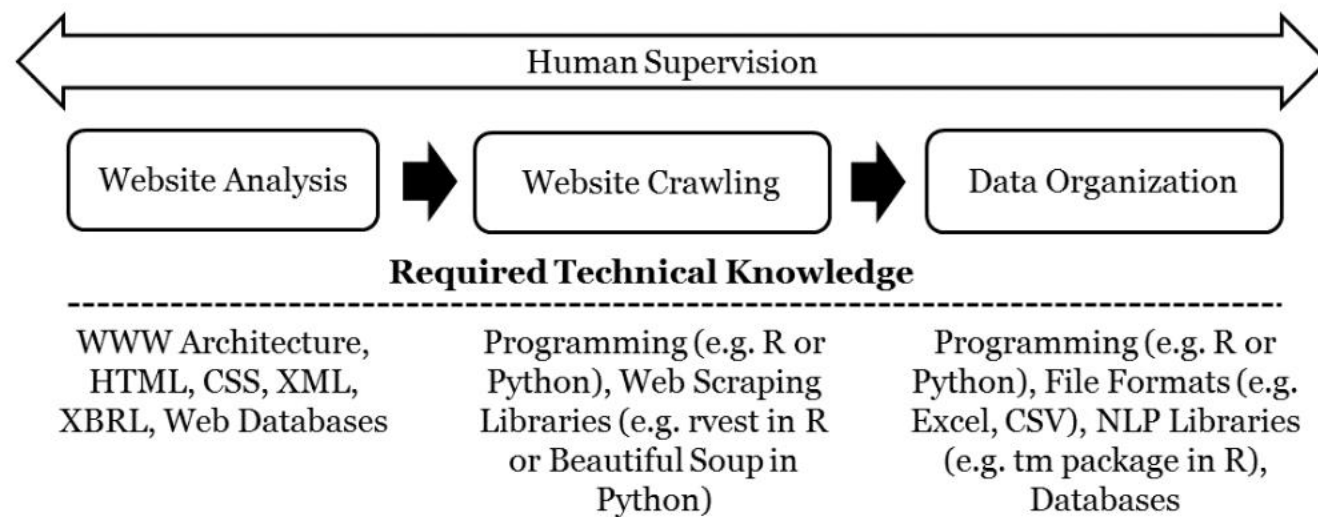


Figure 1: Web Scraping (Adapted from Krotov and Tennyson 2018)



Yes, tools exist!

- Most social scientists do not yet possess the skills to evaluate, re-run, adapt web scraping code
- As trends like the “computational social science” movement (Lazer et al. 2009) and cross-disciplinary collaboration increase, this will change.
- “... proprietary software (e.g., NVivo’s NCapture, Tweetreach, Helium Scraper, Qualtrics) is less technically demanding, these options do not tend to provide the researcher with the same level of reach, flexibility, or transparency.” (Luscombe et al., 2022, p. 3)

Table 2 Available web scraping libraries

Scraping library	Programming language	Particulars
Requests	Python	Useful to acquire raw HTML of a page (static content) High level of abstraction, easy to use in prototypical scraping Able to access APIs and post to forms Unable to access dynamically-generated context
Beautiful Soup	Python	Easy to learn and use (well documented) Ability to automatically detect encodings Much slower than other libraries
Lxml	Python	Very fast and useful for simple extractions Limited in the number of available features
Selenium	Python	Easily interface with dynamically-generated content; websites that generate dynamic content using Javascript requires a browser environment Enables automation of mouse interactions Unable to leverage proxies easily
Scrapy	Python	Amenable to use of proxies and VPNs Useful for complex scraping tasks Highly efficient (i.e. very fast) given its implementation
Request	Javascript	Useful to acquire the raw HTML of a page (static content) Unable to handle dynamically-generated content
Cheerio	Javascript	Requires numerous dependencies
Osmosis	Javascript	HTML parsing, DOM, and request features Fewer dependencies compared to other Javascript libraries
Puppeteer	Javascript	Chrome Browser automation software A Javascript-based implementation of Python's Selenium
Apify	Javascript	Complete web scraping framework
rvest	R	An R-based library with features similar to BeautifulSoup
RSelenium	R	An R-based implementation of Python's Selenium
Kimurai	Ruby	Scrapes dynamically-generated content Supports proxy/user-agent rotation, request delays, default headers
Goutte	PHP	PHP implementation of conventional scraping features

Libraries

The Essence of Web Scraping

Algorithm 1 Programmatic Pseudocode of a Generic Web Scraping Application.

Input: set of URLs to scrape, u
 set of patterns to match, p

Output: set of scraped data, d

```
1: import the scraping library
2: initialize the scraper
3: for each URL,  $u$ , in  $u$  do
4:   for each pattern,  $p$ , in  $p$  do
5:     use scraper to download the website content from  $u$ 
6:     extract the pattern  $p$  from that content
7:     add extracted data to  $d$ 
8:   end for
9: end for
```

Constant battle between Website and Scholar

Table 3 Eight defensive strategies to block web scraping

Defensive strategy	Explanation
1. Defining robots.txt	Explicit definition of how a website should be crawled by specific or all bots
2. Banning IP	When a large number of requests come from a specific IP address that specific IP is blocked from future requests
3. Rate-Limiting IP Requests	When an IP address sends requests at a rate quicker than humanly possible (milliseconds), requests are either returned slowly, rejected, or the IP is banned
4. User-Agent Blocking	When submitting a request, the “user-agent” identifies what browser is making the request; a bot would typically have a blank user-agent field
5. Banning by Navigation-Based Detection, e.g., reCAPTCHA	Sophisticated analyses of the history of a requester’s navigation through a website filters out bot-based navigation. Most popular is Google’s reCAPTCHA v3
6. Requiring Email Verification	An email verification step requires users to link their session to an email address
7. Requiring Mobile Verification	A mobile verification step requires users to link their session to a mobile phone number
8. Requiring an API key	For websites exposing a machine-friendly API, requiring a key to identify users is an effective strategy to limit resource usage

Outsmart the System

Table 4 Workarounds to defensive solutions

Defensive strategy	Scraper solution	Explanation
1. Banning IP	Rotating IP Addresses	A bot can spoof the requesting IP address by relying on a pool of IP addresses; a few requests are made and then the IP is switched preventing IP tracking
2. Rate-Limiting IP Requests	Rate-Limiting Requests	A bot can implement a delay to its requests to mimic a human requester; at times, a random amount of delay is used since a regular and identical interval of requests is also detectable
3. User-Agent Blocking	Spoofing and Rotating User-Agent	A scraper can set the user-agent to mimic a given browser (e.g. Chrome) and rotate to other browsers on each request (e.g. Mozilla Firefox)
4. Banning by Navigation-Based Detection, e.g., reCAPTCHAs	Spoofing human-like cookies; paid CAPTCHA solving services; human-mimicking scraping library (e.g., Selenium)	The navigation history of a bot differs dramatically from human-based navigation; the cookies from a human-type session can be used to mask a bot-type session
5. Requiring Email Verification	Pool of emails; use of “burner” email addresses	Either a pool of email addresses can be rotated through, or a paid service of temporary/“burner” emails can be used
6. Requiring Mobile Verification	Use of “burner” mobile phone number service (e.g., Twilio)	As it is expensive to maintain active phone numbers, a one-time use of a “burner” mobile number & messaging service cheaply circumvents this strategy
7. Requiring an API key	Rotating pool of API keys	As long as API keys are free, a pool of keys can be generated linking to “different users”



Web Scraping in the Social Sciences and Humanities

- Offers (new) opportunities for research
- Different types of structured and unstructured data
- A (relatively) simple process
- Tools versus code
- Defensive strategies and solutions
- Computational Social Sciences/Digital Humanities

Personal Examples

Erik Hekman

A Practical Approach to Web Scraping

Kickstarter

kickstarter.com

Discover Start a project

KICKSTARTER

Search

Arts Comics & Illustration Design & Tech Film Food & Craft Games Music Publishing

Bring a creative project to life.

ON KICKSTARTER:

231,559 projects funded	\$7,020,433,956 towards creative work	84,861,070 pledges
----------------------------	--	-----------------------

FEATURED PROJECT

A VERY Klaus CHRISTMAS
BY GRANT MORRISON & DAN MORA
BOOM! RELEASE PRE-ORDER

RECOMMENDED FOR YOU

Locked Lairs & The Puzzles That Ke...
537% funded
By THEORYCRAFT

Midwinter Vampires, a blood-suckin...
2,436% funded
By Matthew David

Umbral Core - Dark Fantasy 2.5D...

THE 10 BEST Restaurants in Lo

tripadvisor.com/Restaurants-g186338-London_England.html

THE 10 BEST Restaurants in London

ReviewTripsAlertsSign inCart

LondonHotelsThings to doRestaurantsFlightsVacation RentalsVacation PackagesCruisesRental Cars

Europe > United Kingdom (UK) > England > London > London Restaurants

Best Restaurants in London, England

Restaurants in London

Reserve: 12/12/20228:00 PM2 guestsFind a restaurantClear search

View map

COVID-19

☐ Restaurants taking safety measures

Establishment Type

☒ Restaurants

☐ Quick Bites

☐ Dessert

☐ Coffee & Tea

Show more

Delivery Available



Sponsored

Sticks'n'Sushi Canary Wh...

528

Japanese, Sushi, As...

\$\$\$ - \$\$\$

Order Online



Bonoo Indian Tapas

1,915

Wine Bar, Indian, C...

\$\$\$ - \$\$\$

Order Online



Devine Restaurant Coffe...

333

Cafe, European, British

\$

Order Online



Hibox

231

Lebanese, Mediterranean...

\$

Order Online

See all

Outdoor Seating Available



THE BEAR

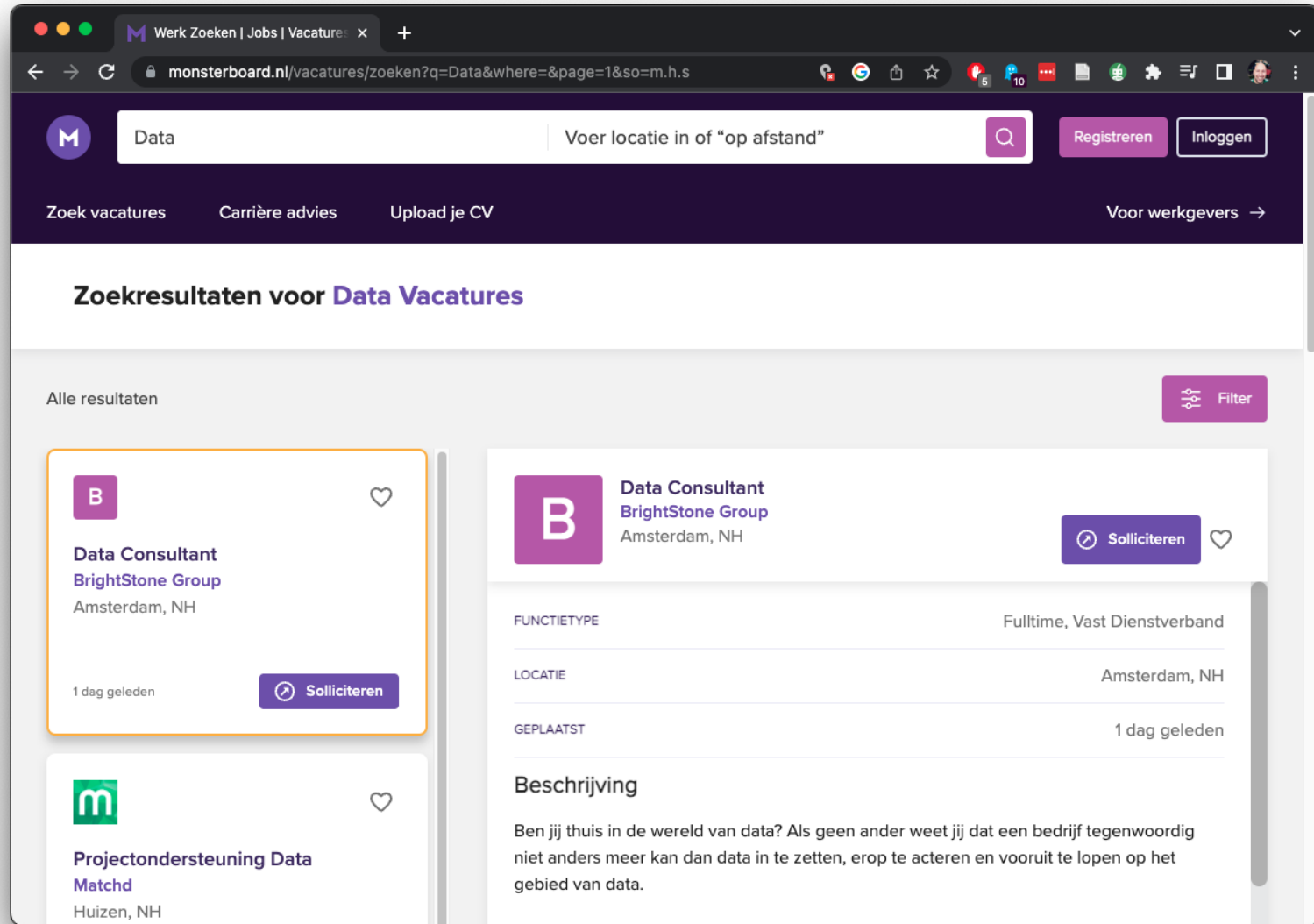






PIZZA BAR - ALL SUPERFAST

See all



Werk Zoeken | Jobs | Vacatures x +

monsterboard.nl/vacatures/zoeken?q=Data&where=&page=1&so=m.h.s

M Data Voer locatie in of "op afstand" Registreren Inloggen

Zoek vacatures Carrière advies Upload je CV Voor werkgevers →

Zoekresultaten voor Data Vacatures

Alle resultaten Filter



Data Consultant
BrightStone Group
Amsterdam, NH

1 dag geleden

Solliciteren



Projectondersteuning Data
Matchd
Huizen, NH



Data Consultant
BrightStone Group
Amsterdam, NH

Solliciteren

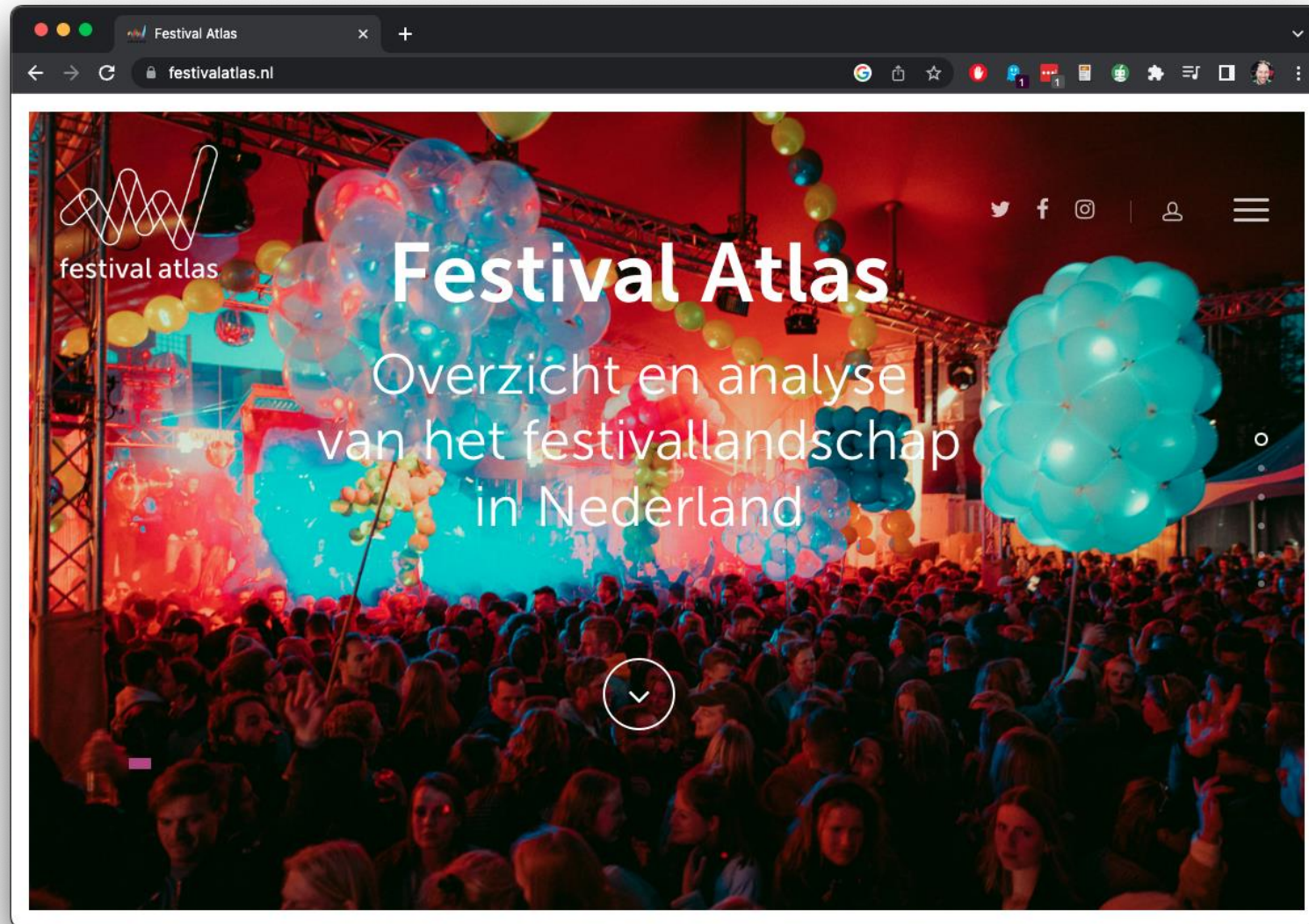
FUNCTIETYPE Fulltime, Vast Dienstverband

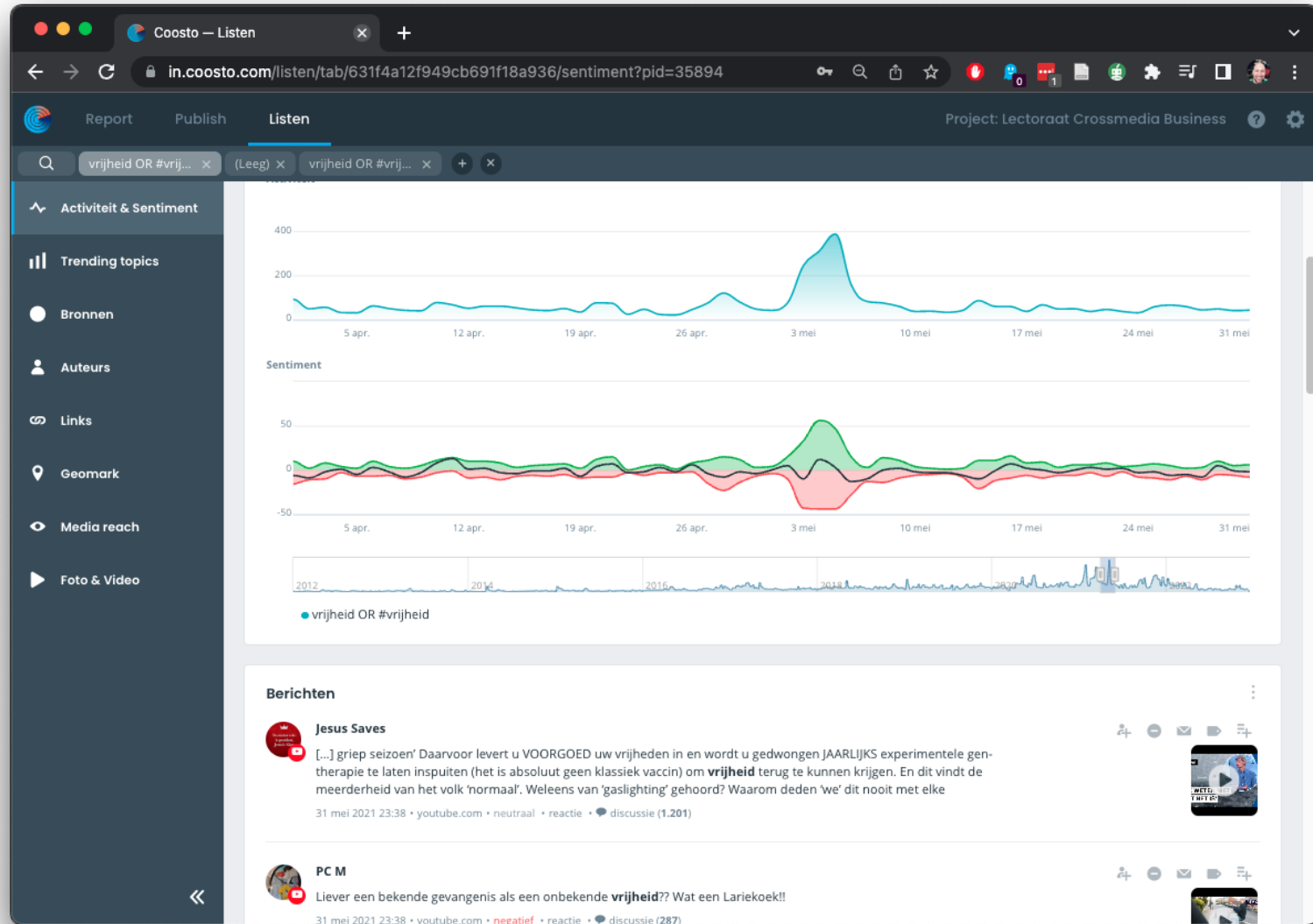
LOCATIE Amsterdam, NH

GEPLAATST 1 dag geleden

Beschrijving

Ben jij thuis in de wereld van data? Als geen ander weet jij dat een bedrijf tegenwoordig niet anders meer kan dan data in te zetten, erop te acteren en vooruit te lopen op het gebied van data.





Search | WIRED

wired.com/search/?q="artificial+intelligence"&sort=score+desc

17342Off

WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

SIGN IN

SUBSCRIBE

Search stories from WIRED

Q "artificial intelligence"

SEARCH

4,353 stories about "artificial intelligence"

Sort by Relevance



DEPTH OF FIELD

A New Trick Lets Artificial Intelligence See in 3D

BY WILL KNIGHT

02.07.2022

MOST POPULAR



SECURITY

Attackers Keep Targeting the US Electric Grid

ANDY GREENBERG

Titeloverzicht - BNNVARA

bnnvara.nl/catalogus/programmas

BNNVARA

Menu

Sluit je aan!

Titeloverzicht

A B C D E F G H I J K L M N O P Q R S T U



#BOOS

-



#IkRedHetNietMeer

-



101Barz

-



26.000 gezichten

-



Home | Radio Dabanga x +

dabangasudan.org/en

العربية

Independent news from the heart of Sudan

DABANGA

Dabanga TV
Watch live →

HOME ALL NEWS EDITORIAL GUIDELINES 2021 COUP BROADCASTS ABOUT US CONTACT US

NEWS HEADLINES

December 11 - 2022 **DELLING / WAD EL MAHI**
Situation 'worse than dire' for West Kordofan displaced

The influx of people fleeing violence in the West Kordofan areas of Lagawa and Tema to the newly established camp for the displaced near Delling in northern South Kordofan continues...

December 11 - 2022 **KHARTOUM / OMDURMAN / PORT SUDAN**
Sudan doctors: 50+ injured as Dec 8 Marches of the Millions suppressed

More than 50 protesters were wounded during the Marches of the Millions in Khartoum and Omdurman on Thursday, according to a field report on Friday by the Central Committee

HIGHLIGHTED

Call to review procedures, laws, that ensure access to legal abortion by rape victims in Sudan

December 11 - 2022 **KHARTOUM**

LISTEN TO BROADCASTS

December 12 - 2022 06:30 - 07:30
Morning news 12 December

December 11 - 2022 17:30 - 18:30
Evening news 11 December

December 11 - 2022 06:30 - 07:30
Morning news December

December 10 - 2022 17:30 - 18:30
Evening news 10 December

<https://www.dabangasudan.org/en/all-news>

WhatsApp

web.whatsapp.com

Get notified of new messages
Turn on desktop notifications >

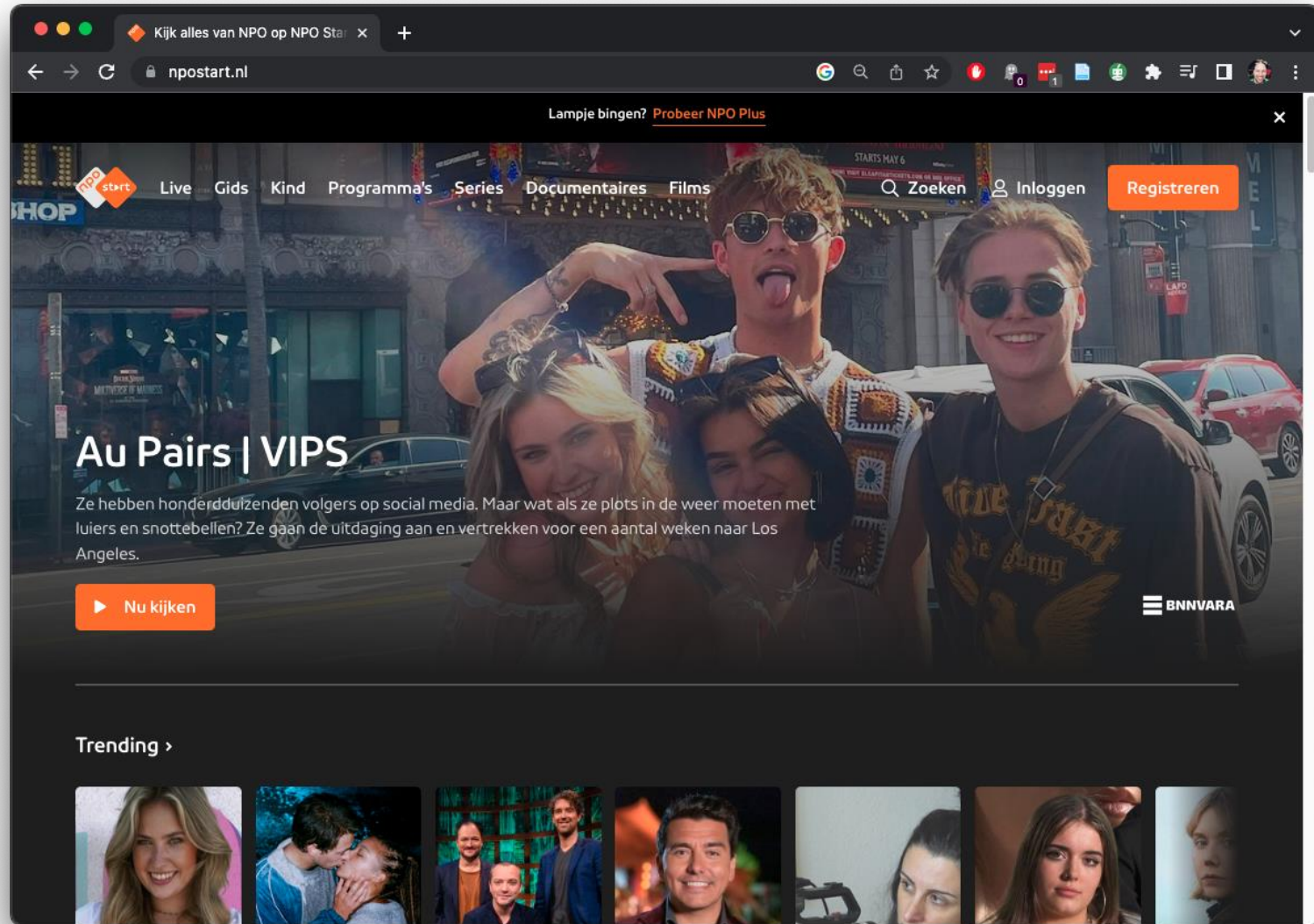
Search or start new chat

WhatsApp Web new

Now send and receive messages without keeping your phone online.
Use WhatsApp on up to 4 linked devices and 1 phone at the same time.

☐ Make calls from desktop with WhatsApp for Windows. Get it here.

End-to-end encrypted



Kirin Ichiban

bierista.nl/kirin-ichiban

Bierista trakteert in december: 2 halen 1 betalen >> [Bestel nu](#)

 **BIERISTA**
MEESTER IN BIER

Zoek op bier...

Login

Haal je Bierdiploma

Bierista > Kirin Ichiban



Kirin Ichiban

Kirin Brewery Company (1)

Kirin Ichiban is een populair Japans bier, een Lager zonder uitgesproken smaken en zonder bittere nasmaak. De Pale Lager is vooral dorstlessend en prima doordrinkbaar.

6,4

Haal nu je **bierdiploma!**

Meer info



Een **gediplomeerd** bierdrinker worden?

Details

Brouwerij	Kirin Brewery Company
Bierstijl	Lager
Alcohol	5.0%
Categorie	Soepel en Subtiel
Spijzen	Salades, Kip en gevogelte, Witvis, Jonge en romige kazen
Drinktemp.	3 °C
Volume	33 cl
Drinkmoment	Terras, Barbecue, In goed gezelschap
Gisting	Ondergistend

BIER BUTLER

Powered by Bierista

Mijn bieren

☒ Dit bier heb ik gedronken

Mijn cijfer:

7

Gedronken op:

Star Trek - Season 1 - IMDb

imdb.com/title/tt0060028/episodes?ref_tt_eps_sm

IMDb Menu All Search IMDb IMDbPro Watchlist Sign In EN

Star Trek (1966–1969)
Episode List

Season: 1 OR Year:

Season 1

The Cage 4 Oct. 1988
★ 7.6 (6,362) ☆ Rate
Capt. Pike is held prisoner and tested by aliens who have the power to project incredibly lifelike illusions.

The Man Trap 8 Sep. 1966
★ 7.2 (5,143) ☆ Rate
Dr. McCoy discovers his old flame is not what she seems after crew members begin dying from a sudden lack of salt in their bodies.

Charlie X 15 Sep. 1966
★ 7.0 (4,828) ☆ Rate
Captain Kirk must learn the limits to the power of a 17-year-old boy with the psychic ability to create anything and destroy anyone.

Sign In X
New Customer? [Create account](#)

Star Trek
TV
Episode List
[Explore More](#)

Editorial Lists
Related lists from IMDb editors

- To Boldly Go: The Women of "Star Trek"**
a list of 24 images
updated 24 Oct 2017
- 12 Biggest Emmy Snubs and Surprises of All Time**
a list of 12 images
updated 15 Sep 2016
- "Star Trek" Villains Through the Years**
a list of 12 images
updated 20 May 2016
- Top 15 "Star Trek" Episodes as Ranked by IMDb Users**
a list of 15 images
updated 19 May 2016
- 2016 Oscars In Memoriam**
a list of 84 images
updated 04 Aug 2016

Legality and Ethics of Web Scraping

Erik Hekman

A Practical Approach to Web Scraping

What is your perspective on this?

Erik Hekman

A Practical Approach to Web Scraping



Legal and ethical challenges

- Grey area in the legal field (Snell & Manaldo, 2016)
- No direct legislature that addresses it:
 - Copyright infringement
 - Terms of Use (Breach of Contract)
 - Damage to the Website



Legal and ethical challenges

- Individual privacy
- Organizational privacy
- Right of research subject



Figure 2. Legality and Ethics Framework for Web Scraping

Live coding

Erik Hekman

A Practical Approach to Web Scraping

Practical tips



- Pay for the content
- Be sensible
- Create a queue or list
- Use Selenium or any alternative
- Start locally, then scale up (e.g. a cloud server)
- Save HTML files
- Work in batches (scraping and extracting)
- Try/except for missing elements
- Use Multiprocessing
- `from multiprocessing import Pool`

Happy Web Scraping!

Erik Hekman

A Practical Approach to Web Scraping

Questions?

Erik Hekman

A Practical Approach to Web Scraping