Web Scraping 101

Introductie in Web Scraping

Bèr `berkes` Kessels

Web Scraping 101

Over deze presentatie

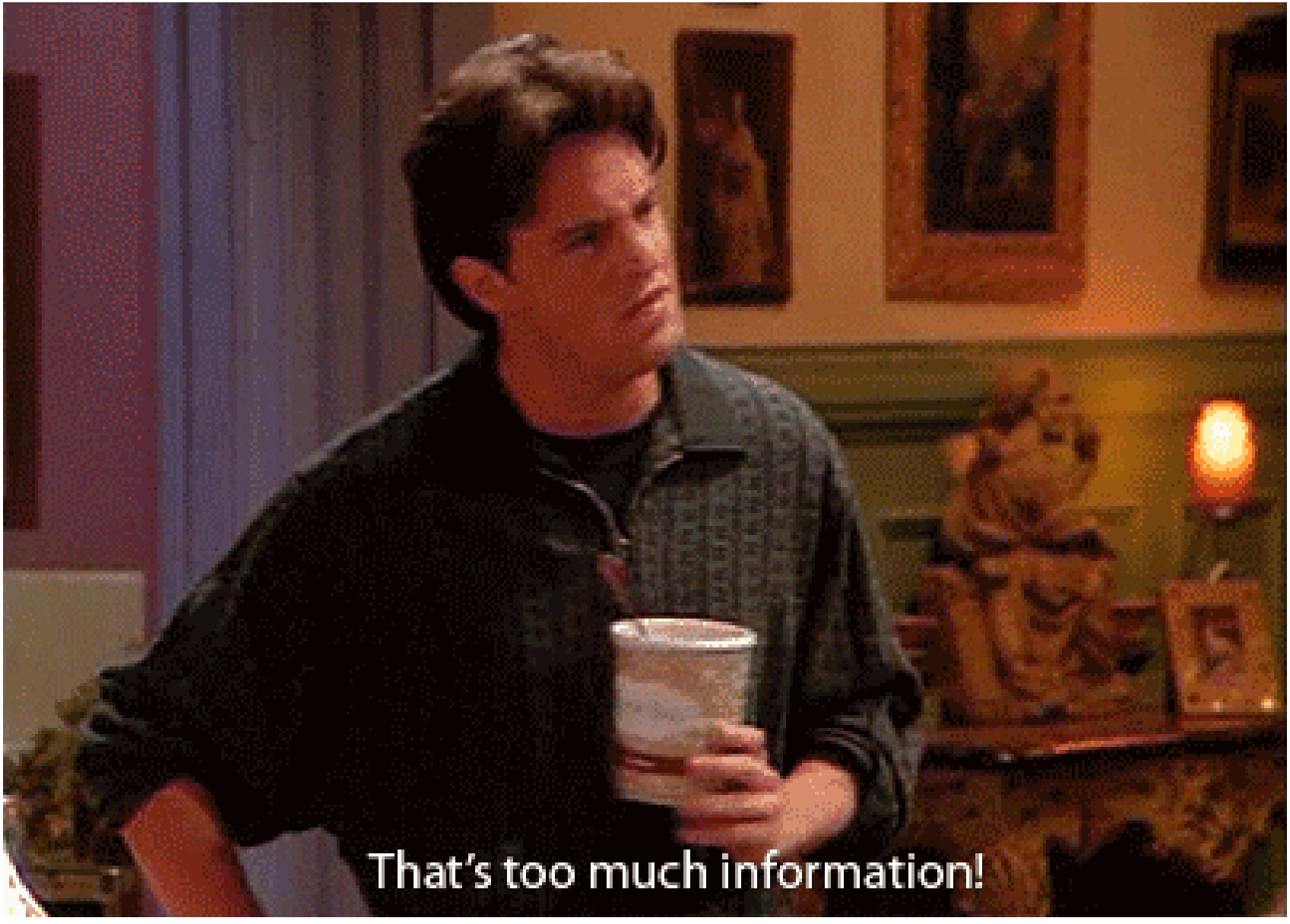
- Is online: berk.es/scraping (github.com/berkes/scraping)
- Bevat alle links

Over mij

- Bèr `berkes` Kessels
- @berkes - LinkedIn, Twitter, Fediverse
- Werkt aan Founder Led Sales en andere
- Blog https://berk.es

Scraping

Web-scraping is een techniek waarbij software automatisch gegevens van websites verzamelt door de HTML-structuur van de pagina te analyseren en specifieke informatie te extraheren. Het wordt vaak gebruikt voor taken zoals het ophalen van nieuwsartikelen, het verzamelen van product-prijzen of het analyseren van concurrenten op het web.



Figuur 1: Too much information

Verschillende opties

Van simpel tot complex.

- One-off - Parse één url.
- Spider - Kruip door een hele site of groep sites.
- Mirror - Als spider, maar bewaar de HTML en assets lokaal.
- Schrijf een gespecialiseerde scraper.

cURL

Figuur 2: Demo

cURL

- Kan alleen maar dingen ophalen
- Kan heel veel HTTP dingen: van headers tot obscure features
- Snapt HTML niet
- Overal beschikbaar
- Kent bijna alle protocollen, niet alleen HTTP(s)

wget

Figuur 3: Demo

wget

- Kent enkel WWW en HTTP(s)
- Heeft HTTP(s) features zoals redirects volgen
- Snapt HTML een beetje
- Heeft opties om links te volgen
- Kan een site mirroren

Onderdelen

- Iets wat de boel ophaalt
- Iets wat de boel parsed
- Iets wat daaruit informatie haalt

Ophalen

Alles wat HTTP kent of kan.

- Mirroren: voordeel: eenmalig. nadeel: diskruimte.

Parsen

- XPATH
- CSS selectors
- Goede HTML (bijna nooit)
- HTML parsers nodig
- Selector vinden en debuggen met developer tools.
- Addons of volledige IDEs om dit te vereenvoudigen.

Xpath en CSS selectors

- Xpath is erg compleet. Kan bijna alles
- CSS selectors zijn eenvoudiger maar ingwikkelder dingen zijn meteen heel moeilijk.
- Balanceeract tussen fragiel en robuust

Selectors

Figuur 4: Demo

Een scraper programmeren

- Veel frameworks en suites voor.
- Oudste vaak Perl.
- Bekendste Scrapy - Python.

Voorbeeld Scrapy

Figuur 5: Demo

Scrapy

- Parrallel
- Tooling: Logging, debugging, scheduling etc.
- Meerdere spiders in één project beheren
- ETL functies (naar database wegschrijven etc)

Voorbeeld Puppetteer

Figuur 6: Demo

Puppetteer

- Eén voorbeeld van een full browser
- Headless (headfull voor debugging)
- Interactie DSL
- Geen scraping framework
- Scrapy kan ook headless browser

Samengevat

- cURL of wget voor simpele requests
- Scrapy voor complexere taken
- Headless browser voor dynamische content

Pres

- berk.es/scraping