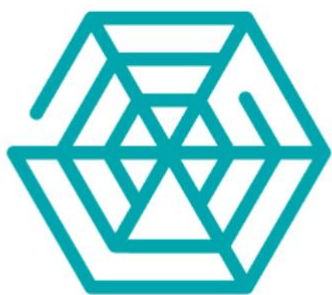


HANDLEIDING WEBSCRAPER.IO

Amy Hermsen, Floris Rexwinkel, Joas Kuiper & Kellen Gortemaker



Web Scraper

0.5.0 RELEASE

Inhoudsopgave

Inleiding Geschreven door Amy	3
Doel van deze handleiding	3
Leerdoelen	3
Wat is webscraping? Geschreven door Amy	4
Journalistieke toepassingen	4
Voordelen ten opzichte van handmatig data verzamelen	4
Beperkingen en aandachtspunten	4
Wat is Webscraper.io? Geschreven door Amy	5
Waarom kiezen voor Webscraper.io?	5
Mogelijkheden van Webscraper.io Geschreven door Joas	6
Kernfunctionaliteiten	6
Toepassingen in onderwijs en werk	6
Webscraper.io in de journalistiek	6
Schaalbaarheid en integratie met tools	7
Grenzen en Aandachtspunten Geschreven door Joas	8
Technische grenzen	8
Juridische aandachtspunten	8
Ethische overwegingen: kun je scrapen wat je mag scrapen?	9
Installatie en Download Geschreven door Floris	10
Basisbegrippen Geschreven door Floris	12
Sitemap	12
Selectors	12
Pagination	12
Exporteren	12
Stap-voor-stap voorbeeldscrape (Browser-Extensie) Geschreven door Kellen	13
Web Scraper extensie stap voor stap uitgelegd	13
Nadelen Web Scraper extensie	22
Stap-voor-stap voorbeeldscrape (F12 versie) Geschreven door Joas	23
Instructies	23
Problemen en Fouten Geschreven door Floris	29
Conclusie Geschreven door Kellen	30

Inleiding

Geschreven door Amy

In een tijd waarin bijna alle informatie online beschikbaar is, speelt webscraping een steeds belangrijkere rol bij het verzamelen en analyseren van data. Journalisten gebruiken webscraping om grote hoeveelheden online gegevens te verzamelen en te analyseren. Bijvoorbeeld om trends in nieuwsberichten te ontdekken, prijzen of woningaanbiedingen te vergelijken of om transparantieonderzoek te doen naar politieke partijen, bedrijven of overheidsinstanties. Webscraping maakt het mogelijk om gegevens automatisch te verzamelen zonder dat elke webpagina handmatig bezocht hoeft te worden. Dit bespaart niet alleen tijd, maar biedt ook nieuwe inzichten en verhalen die anders verborgen zouden blijven.

Webscraper.io is een toegankelijke tool waarmee journalisten, onderzoekers en studenten zonder programmeerkennis webdata kunnen verzamelen. De tool biedt zowel een browserextensie als een F12-versie, waardoor hij geschikt is voor zowel kleinschalige onderzoeksprojecten als grotere datajournalistieke analyses. Deze handleiding helpt je stap voor stap op weg om met Webscraper.io te werken, van het begrijpen van de basisprincipes tot het uitvoeren van je eerste webscraping.

Doel van deze handleiding

Het doel van deze handleiding is om je wegwijs te maken in webscraping en het gebruik van Webscraper.io. We leggen uit wat webscraping is, hoe en wanneer je het kunt toepassen, welke mogelijkheden de tool biedt en waar de grenzen liggen, zowel technisch als juridisch en ethisch. Daarnaast vind je een praktische stap-voor-stap uitleg om zelf een eenvoudige scrape uit te voeren.

Leerdoelen

Na het doorlopen van deze handleiding kun je:

1. Uitleggen wat webscraping is en waarvoor het gebruikt kan worden.
2. Het verschil beschrijven tussen de browserextensie en de F12-versie van Webscraper.io
3. Zelf een eenvoudige webscraping uitvoeren en de resultaten exporteren.
4. De mogelijkheden én grenzen van webscraping benomen, met aandacht voor ethiek en regelgeving.

Wat is web scraping?

Geschreven door Amy

Web scraping is een techniek waarmee informatie automatisch van websites wordt gehaald. In plaats van gegevens handmatig te kopiëren, wordt een digitaal script of een tool, zoals Webscraper.io, gebruikt om gestructureerd data te verzamelen. De verzamelde informatie kan vervolgens worden opgeslagen in een bestand (bijvoorbeeld CSV of Excel) en verder worden geanalyseerd.

Journalistieke toepassingen

Hoewel er in de tutorial is geoefend met een eenvoudige dataset die bestaat uit webshopproducten, kan dezelfde techniek in de journalistiek veel bredere en betekenisvollere toepassingen hebben. Bijvoorbeeld:

- Onderzoek naar politieke besluitvorming: het scrapen van gegevens van Kamerleden om patronen of belangenconflicten zichtbaar te maken.
- Analyse van publieke uitgaven: het verzamelen van data uit overheidsportalen of aanbestedingssites om te onderzoeken waar belastinggeld naartoe gaat.
- Trendanalyse in media of sociale netwerken: het scrapen van nieuwsartikelen of reacties om te zien welke onderwerpen het meeste aandacht krijgen.
- Consumentenonderzoek en reviews: het verzamelen van product- of bedrijfsreviews om inzicht te krijgen in publieke opinies of reputaties.
- Transparantieprojecten: het bundelen van data over subsidies of milieumeldingen om verhalen te vertellen die anders verborgen blijven.

Meer over de toepassing van Webscraper.io is te lezen in het hoofdstuk *Mogelijkheden van Webscraper.io*.

Voordelen ten opzichte van handmatig data verzamelen

Één van de grootste voordelen van web scraping is de efficiëntie, met behulp van deze tool kunnen in korte tijd grote hoeveelheden data worden verzameld. Daarnaast is de nauwkeurigheid ook een voordeel, automatische scraping voorkomt menselijke fouten bij het kopiëren van data. Het derde voordeel is de herhaalbaarheid, scrapers kunnen periodiek opnieuw draaien, bijvoorbeeld om verandering in prijzen of politieke standpunten te volgen. Het laatste voordeel van web scraping is de verwerking in excel, door de tabellen die ontstaan kunnen journalisten gemakkelijk sorteren, filteren of visualisaties maken.

Beperkingen en aandachtspunten

Web scraping heeft ook een aantal beperkingen, hierover wordt in het hoofdstuk *Grenzen en aandachtspunten* van de handleiding de diepte in gegaan. Het draait om technische-, juridische- en ethische beperkingen. Technisch gezien staan niet alle websites scraping toe, of soms hebben ze complexe structuren die scraping bemoeilijken. Als het draait om de juridische beperkingen gaat het vaak over inhoud die auteursrechtelijk beschermd is of valt onder privacywetgeving. Ethisch ontstaat de vraag of het verzamelen en analyseren van (openbare) informatie over individuen privacy schendt of schadelijk kan zijn.

Wat is Webscraper.io?

Geschreven door Amy

Webscraper.io is een gebruiksvriendelijke tool waarmee je zonder programmeerkennis informatie van websites kunt verzamelen. De tool is ontworpen als een visuele webscraper, wat betekent dat je letterlijk kunt aanwijzen welke elementen van een webpagina je wilt verzamelen zoals titels, namen, links, prijzen of recensies.

Voor journalistieke toepassingen is Webscraper.io bijzonder geschikt omdat het snel inzicht kan geven in grote hoeveelheden online data. Denk aan het analyseren van honderden nieuwsartikelen, het vergelijken van politieke websites of het monitoren van veranderingen in openbare registers.

Waarom kiezen voor Webscraper.io?

Er bestaan verschillende tools om webdata te verzamelen, maar Webscraper.io onderscheidt zich door zijn toegankelijkheid:

- Het werkt rechtstreeks in de Google Chrome-browser, zonder dat je code hoeft te schrijven.
- Het gebruikt een visuele interface waarmee je eenvoudig elementen selecteert op een webpagina.
- Het biedt mogelijkheden om data te exporteren naar Excel, CSV of JSON, zodat je de informatie direct verder kunt analyseren.
- Voor gevorderde gebruikers is er een F12-versie, waarmee grotere en geautomatiseerde scrapingprojecten kunnen worden uitgevoerd.

Voor beginnende datajournalisten is Webscraper.io dus een laagdrempelige manier om te leren werken met data als bron voor verhalen.

Mogelijkheden van Webscraper.io

Geschreven door Joas

Webscraper.io is dus een browserextensie voor Chrome die het mogelijk maakt om gestructureerd data van websites te verzamelen, zonder dat er echt sprake is van programmeerkennis. Gebruikers kiezen zelf welke informatie ze willen scrapen (zoals tekst, links of afbeeldingen) en hoe ze door een website navigeren (bijvoorbeeld door te klikken op 'volgende pagina'). De verzamelde data kan vervolgens makkelijk worden geëxporteerd naar bestandsformaten zoals CSV of XLSX om te gebruiken voor analyse.

Kernfunctionaliteiten

Hieronder staan de belangrijkste functies van Webscraper.io:

- **Visuele sitemap builder:** Maak en beheer scraping-instructies via een gebruiksvriendelijke interface.
- **Flexibele data selectie:** Mogelijkheid om diverse datatypes te verzamelen, inclusief content van dynamische (AJAX) pagina's en data verspreid over meerdere pagina's.
- **Eenvoudig exporteren:** De verkregen data is direct klaar voor gebruik in spreadsheetprogramma's.

Toepassingen in onderwijs en werk

Doordat Webscraper.io op verschillende manieren ingezet kan worden, is het een handige tool in diverse sectoren:

1. **Onderzoek:** Voor het verzamelen van openbare gegevens voor academische studies, marktanalyse, of concurrentieonderzoek.
2. **Leadgeneratie:** Het systematisch verzamelen van openbare contactinformatie voor potentiële klanten, rekening houdend met privacywetgeving en websitevoorwaarden.
3. **E-commerce:** Voor prijsmonitoring, inventarisbeheer, of het analyseren van productrecensies van online winkels.
4. **Journalistiek:** Het verzamelen van publieke webdata voor verhalen en data-onderzoek.

Webscraper.io in de journalistiek

Voor datajournalisten is Webscraper.io dus een handig middel om datasets samen te stellen uit openbare webbronnen die niet als gestructureerde data beschikbaar zijn. Dit stelt journalisten in staat om originele en impactvolle verhalen te vertellen. Er zijn verschillende mogelijkheden te bedenken en hieronder worden er een aantal uitgelicht.

Economische analyse

Verzamel prijzen van essentiële producten (zoals voedsel of energie) over verschillende regio's of periodes. Analyseer deze data om trends in inflatie, regionale prijsverschillen of de impact van marktmacht te visualiseren. Dit helpt bij het inzichtelijk maken van bredere economische patronen en hun effecten op de samenleving.

Consumenten inzichten

Scrape product- of dienstrecensies om het publieke sentiment, veelvoorkomende klachten of tevredenheidstrends te analyseren. Deze kwalitatieve data, omgezet in kwantitatieve inzichten, kunnen verhalen opleveren over consumentenervaringen en productkwaliteit.

Politiek en bestuurlijk onderzoek

Scrape openbare profielen, stemgedrag, declaraties, of notulen van politici, beleidsmakers of lobbyisten. Analyse van deze datasets kan leiden tot verhalen over:

- **Transparantie en verantwoording:** Onderzoek van financiële belangen, nevenfuncties of aanwezigheid bij vergaderingen.
- **Beleidsanalyse:** Patronen in besluitvorming of de invloed van externe partijen op beleid.

Het is hierbij cruciaal om de verzamelde data goed te contextualiseren: *Welke maatschappelijke vragen kunnen met deze data worden beantwoord en welk verhaal wil de journalist ermee vertellen?*

Schaalbaarheid en integratie met tools

Webscraper.io is ontworpen om zowel kleine als grote hoeveelheden data te verwerken. Het kan efficiënt data verzamelen van een paar producten tot duizenden rijen aan informatie, waardoor handmatige dataverzameling, die tijdrovend en foutgevoelig is, wordt geautomatiseerd. De geëxporteerde data is te combineren met een veelheid aan analyse- en visualisatietools, zoals Excel/Google Sheets, data-analyse software (bijv. Python) en visualisatietools (bijv. Flourish).

Door Webscraper.io in te zetten kunnen gebruikers, met name datajournalisten, dus zelfstandig en makkelijk toegang krijgen tot een grote hoeveelheid aan webdata om deze vervolgens te kunnen gebruiken in betekenisvolle inzichten en verhalen.

Grenzen en Aandachtspunten

Geschreven door Joas

Hoewel Webscraper.io een zeer handig hulpmiddel is, zijn er belangrijke technische, juridische en ethische overwegingen waarmee rekening moet worden gehouden bij het verzamelen van webdata. Het is erg belangrijk om verantwoord te werk te gaan.

Technische grenzen

Websites gebruiken vaak beveiligingsmaatregelen tegen geautomatiseerde scraping, zoals captcha's, IP-blokkades of complexere detectiemechanismen. Deze kunnen het functioneren van de scraper beïnvloeden. Daarnaast kan het versturen van veel verzoeken in korte tijd de servers van een website overbelasten, wat kan leiden tot tijdelijke blokkades en een verminderde beschikbaarheid voor andere gebruikers. Daarom wordt geadviseerd om scraping-activiteiten te vertragen ('throttling') en pauzes in te bouwen in de sitemap. Websites met veel dynamische inhoud (AJAX), vereiste loginprocedures, of complexe interacties kunnen ook lastiger zijn om te scrapen. Hoewel Webscraper.io hiervoor vaak oplossingen biedt, is niet elke websitestructuur even gemakkelijk te benaderen.

Juridische aandachtspunten

Bij het scrapen van data zijn verschillende juridische aspecten van belang. Ten eerste is het essentieel om het robots.txt-bestand van een website te respecteren. Dit bestand geeft aan welke delen van de site niet door webcrawlers bezocht of geïndexeerd mogen worden. Het negeren hiervan wordt niet alleen als onethisch beschouwd, maar kan ook juridische gevolgen hebben. Daarnaast hebben veel websites gebruiksvoorwaarden die het geautomatiseerd verzamelen van data expliciet verbieden. Het schenden van deze voorwaarden kan leiden tot juridische stappen, waaronder de uitsluiting van de website. Een andere belangrijke overweging is auteursrecht. Veel online content is beschermd, en het hergebruiken of publiceren van grote hoeveelheden gescrapete, auteursrechtelijk beschermde data zonder toestemming kan leiden tot inbreuk op het auteursrecht. Tot slot is de Algemene Verordening Gegevensbescherming (AVG/GDPR) een belangrijk aandachtspunt wanneer de gescrapete data persoonsgegevens bevat (zoals namen, e-mailadressen, of gevoelige informatie). De verwerking van dergelijke gegevens moet voldoen aan de AVG, wat betekent dat er een rechtmatige grondslag moet zijn, de data adequaat beveiligd moeten worden en de betrokkenen geïnformeerd moeten worden over de verwerking.

Ethische overwegingen: kun je scrapen wat je mag scrapen?

De discussie rondom webscraping draait vaak om de grens tussen wat technisch mogelijk is en wat ethisch verantwoord of juridisch toegestaan is. Zelfs als iets technisch mogelijk is en juridisch in een grijs gebied valt, is het belangrijk om rekening te houden met de ethische kant van het verzamelen en gebruiken van de data.

Praktijkvoorbeelden en ethische vragen

Hieronder worden aan de hand van de drie besproken perspectieven een aantal voorbeelden gegeven van scrape mogelijkheden.

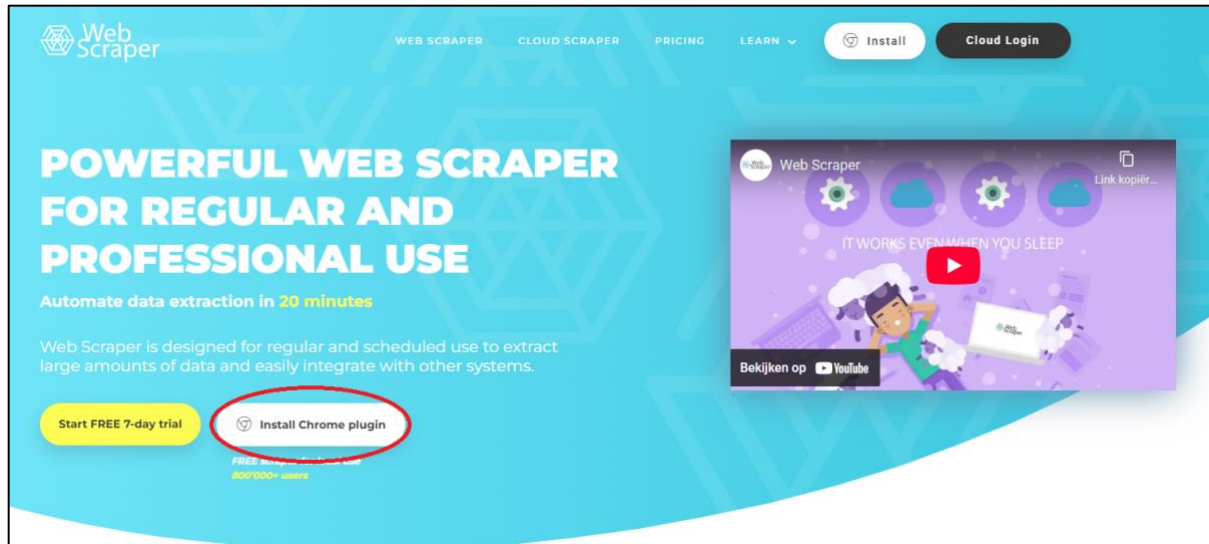
- **Scrapen van sociale media profielen:** Technisch vaak mogelijk van openbare profielen. Juridisch verbieden gebruiksvoorwaarden het vaak en de AVG is van toepassing bij persoonsgegevens. Ethisch is het de vraag of het verzamelen en analyseren van gedetailleerde informatie over individuen, zelfs als deze openbaar is, de privacyverwachting schendt of schadelijk kan zijn (bijv. voor het bepalen van politieke voorkeur).
- **Scrapen van prijzen van kleine ondernemers:** Vaak technisch goed te doen. Juridisch minder snel in strijd met de AVG, maar wel potentieel met gebruiksvoorwaarden. Ethisch is het belangrijk om te overwegen of vergelijking met grote ketens de kleine ondernemer kan benadelen als de data uit context wordt getrokken of verkeerd wordt geïnterpreteerd.
- **Scrapen van overheidsdata:** Technisch veelal openbaar en toegankelijk. Juridisch vaak aangemoedigd (open data beleid). Ethisch ligt de discussie bij de interpretatie en presentatie van de data. Zorgvuldigheid is geboden om geen verkeerde conclusies te trekken of individuen onterecht aan de schandpaal te nagelen. Het doel is transparantie en verantwoording, niet het creëren van desinformatie.

Voordat er gestart wordt met scrapen, is het belangrijk om niet alleen de technische haalbaarheid te overwegen, maar ook de juridische kaders te respecteren en een ethische afweging te maken. Stel daarom de volgende vragen: Waarom wil ik deze data scrapen? Welk nut dient het? Hoe ga ik om met eventuele persoonsgegevens? En wat zijn de potentiële gevolgen van mijn scraping-activiteiten voor de website-eigenaar en de betrokken personen? Door bewust te zijn van deze vragen houden we scrapen ethisch verantwoord.

Installatie en Download

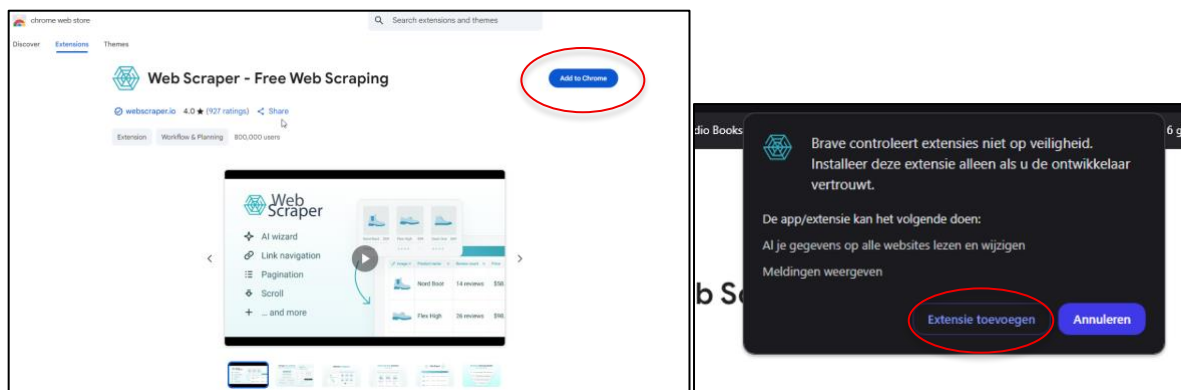
Geschreven door Floris

Om webscraper te gebruiken moet je eerst de extensie downloaden van de site. Zodra je site op komt (webscraper.io) kan je beginnen met de extensie downloaden door op de "install chrome plugin" te drukken.



Nadat je op deze knop hebt geklikt wordt je naar de chrome webstore gebracht. Hier kan je op de blauwe knop "Add to Chrome" of, afhankelijk van je taalinstellingen, "Aan Chrome toevoegen" drukken.

Hierna krijg je nog een korte pop-up waar je de installatie kan bevestigen.



Webscraper.io is nu geïnstalleerd in je browser.

Om webscraper vervolgens te openen kan je op het puzzelstukje naast je zoekbalk klikken. Als je hierop klikt krijg je een lijst met je huidige geïnstalleerde extensies. Webscraper is de extensies met het blauwe spinnenweb. Om makkelijker toegang te krijgen tot Webscraper kan je de extensie aan je taakbalk vastmaken door op het punaise symbool naast de naam van de extensie te klikken.

Om meer uit je Webscraper extensie te halen kun je inloggen op de website van Webscraper. Dit kan je doen door in te loggen met een google account, microsoft account, of door een account aan te maken bij Webscraper zelf. Door in te loggen krijg je toegang tot je scrape opdrachten die je in het verleden hebt gemaakt met een makkelijke interface. Ook kun je scrape's van andere mensen importeren met behulp van de community site maps.

Basisbegrippen

Geschreven door Floris

Om te beginnen met Webscraper is het handig een aantal begrippen te kennen. Deze termen komen vooral voor bij de meer geavanceerde vorm van Webscraper.

Sitemap

Een sitemap is de blauwdruk van je scrapingtaak. Hierin beschrijf je wát je wilt ophalen en hóe de scraper zich door de website moet bewegen.

Selectors

Selectors zijn de datapunten die je wilt meenemen van de site naar je Excel bestand. Dit kunnen titels, prijzen, afbeeldingen of beoordelingen zijn.

Pagination

Als je wilt dat Webscraper over meerdere pagina's data verzamelt, kan je gebruik maken van pagination. Je geeft bij Webscraper aan op welke manier de pagina verdeeld is, dit kan een pagina zijn die meer producten of items toevoegt aan de pagina des te verder de gebruiker naar beneden scrollt. Het kan echter ook zijn dat de site uit meerdere pagina's bestaat, in dit geval kan je aangeven op welke knop Webscraper moet drukken om verder te gaan.

Exporteren

Het is mogelijk je uiteindelijke dataset in meerdere vormen te exporteren. Je krijgt de mogelijkheid om een CSV, XLSX, of JSON bestand te downloaden. CSV en XLSX zijn de belangrijkste en zijn beide te openen met Microsoft Excel. JSON is een Java script versie die meer functioneert als een programmeertaal.

Stap-voor-stap voorbeeldscrape (Browser-Extensie)

Geschreven door Kellen

Web Scraper extensie stap voor stap uitgelegd

Als journalist en data-analist heb je vaak te maken met grote hoeveelheden informatie die online staan. Denk hierbij aan lijsten van politici, organisaties, bedrijven, producten of zelfs reviews van producten. Die data handmatig verzamelen kan enorm veel tijd kosten.

De browser-extensie Webscraper (gratis) haalt, zonder dat je hoeft te coderen, gestructureerde data van webpagina's. De extensie maakt gebruik van AI Wizard Alpha! die voorgestelde velden op een webpagina detecteert. Hierdoor kun je met een paar klikken een hele website laten uitlezen en de resultaten opslaan in een bestand dat je vervolgens in Excel kunt analyseren. Zelfs zonder kennis over HTML computer codes.

In dit hoofdstuk leer je stap voor stap hoe je de extensie gebruikt. We doen dit met een voorbeeld dat voor journalisten relevant kan zijn. In dit geval, het verzamelen van gegevens van alle Tweede Kamerleden (2025).

Na afloop kun je:

- De extensie vastmaken aan de werkbalk
- Data scrapen van de meeste websites
- De gescrapte data converteren naar een Excelbestand
- Begrijpen hoe dit nuttig kan zijn voor datajournalistiek en data-analyses

Browsers

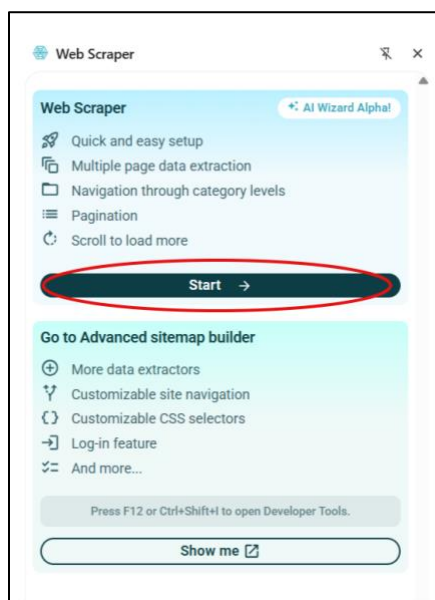
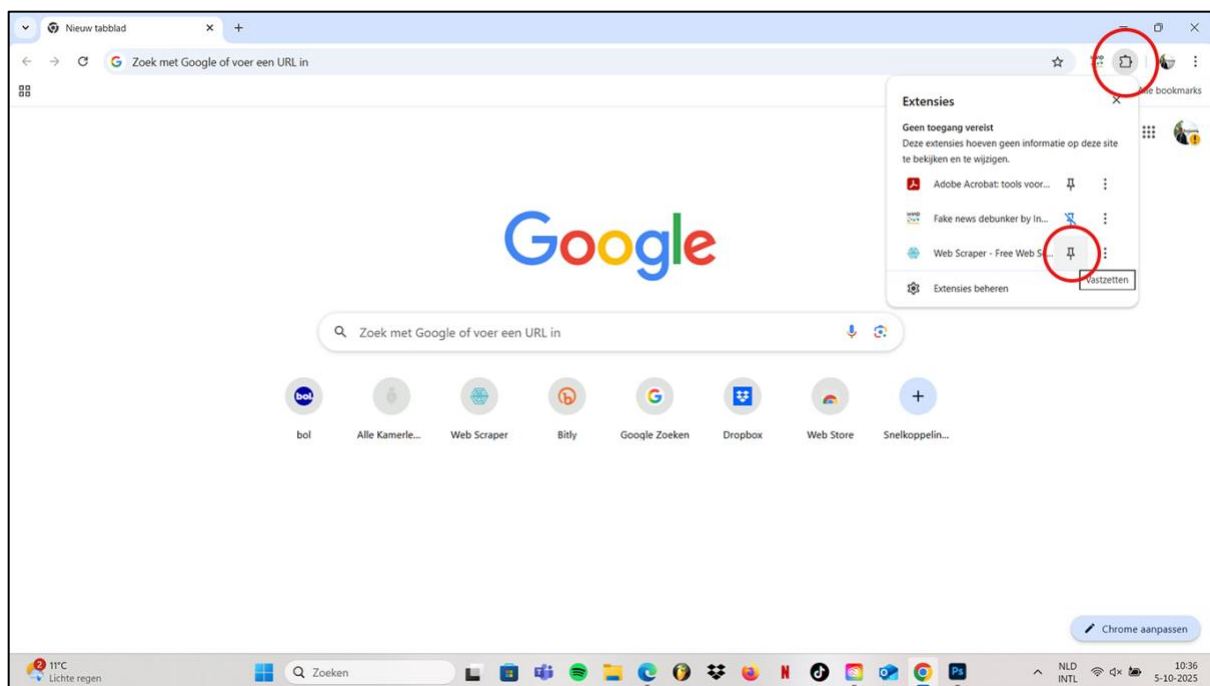
In deze handleiding gebruiken we Google Chrome als voorbeeld browser, omdat deze veel gebruikt wordt. De Webscraper extensie werkt in de meeste moderne browsers zoals Chrome, Firefox en Edge. Op dezelfde manier is de extensie toe te voegen en te gebruiken in de verschillende browsers. De interface en functies van de extensie zijn vrijwel identiek. Dat geldt ook voor de manier waarop je de extensie installeert.

Het belangrijkste verschil zit in de privacyaspecten van de browsers zelf. Google Chrome verzamelt standaard relatief veel gebruikersdata (zoals browsegeschiedenis en gebruik van extensies) om gepersonaliseerde diensten en advertenties te verbeteren. Mozilla Firefox heeft een sterkere focus op privacy. Het blokkeert standaard trackers en deelt minder gegevens met derden. Microsoft Edge zit hier tussenin. Het is gebaseerd op Chromium, maar bevat extra functies voor privacybeheer zoals tracking preventie.

Stap 1: openen Web Scraper extensie

Webscraper is een extensie. Een extensie is een klein programma dat extra functies toevoegt aan je webbrowser (zoals Chrome, Firefox of Microsoft Edge). Je kunt het zien als een hulpmiddel dat je browser slimmer maakt. Extensies kunnen bijvoorbeeld advertenties blokkeren, wachtwoorden beheren of zoals in dit geval data automatisch verzamelen van websites. Om Web Scraper te kunnen gebruiken na het gedownload te hebben moet je een aantal stappen ondernemen.

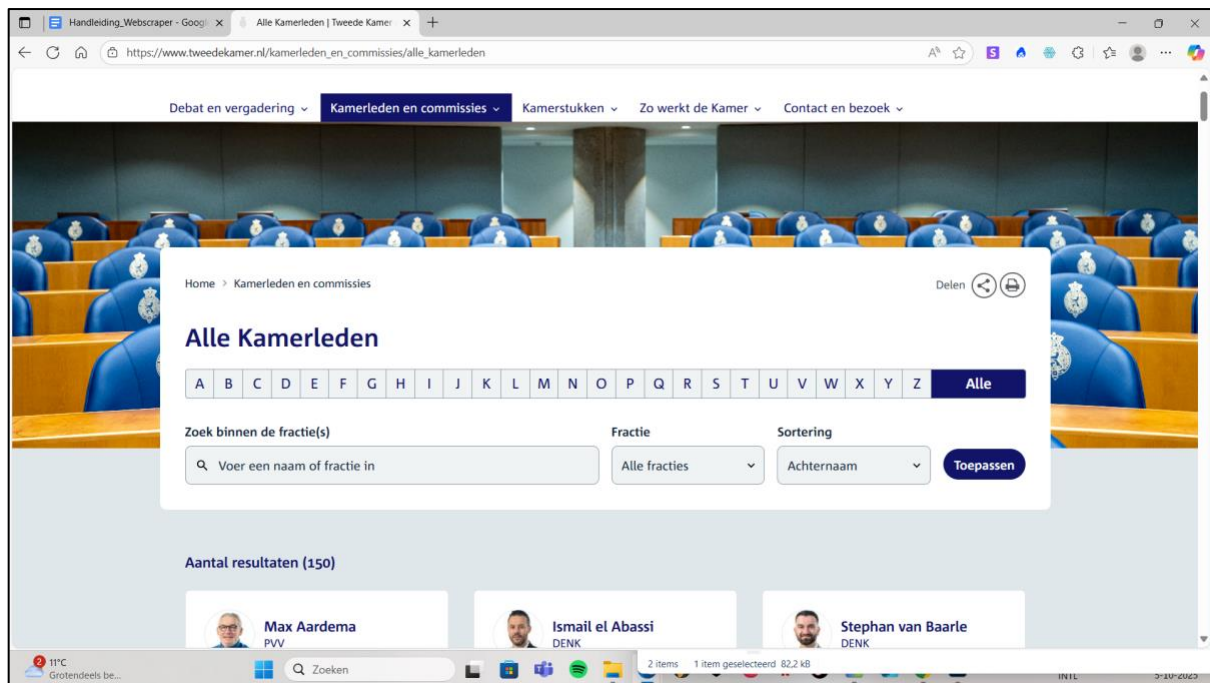
1. Open je browser (Chrome, Firefox of Microsoft Edge) waar je de extensie op hebt gedownload.
2. Klik op Extensies (puzzelstukje) boven aan de browsertoolbar van je browser.
3. Maak de Web Scraper extensie vast aan de browsertoolbar door op de punaise te klikken.
4. Nu de Web Scraper extensie vastzit aan de browsertoolbar kun je hem openen door er op te klikken. Vervolgens klik je op *start*.



Voorbeeld Tweede Kamerleden

De extensie is geopend en nu ben je klaar om te scrapen. De volgende stappen worden uitgelegd aan de hand van een voorbeeld. Hiervoor wordt de pagina Alle Kamerleden gescrapet (2025) en verzamelen we verschillende data van alle Tweede Kamerleden. Hieronder staat de link naar pagina, mocht je de stappen zelf willen proberen aan de hand van het voorbeeld:

https://www.tweedekamer.nl/kamerleden_en_commissies/alle_kamerleden



Op deze pagina staan per Kamerlid gegevens zoals: naam, politieke partij, leeftijd, woonplaats en anciënniteit. Door deze informatie te verzamelen (oftewel te scrapen), kun je de data vervolgens analyseren of gebruiken voor onderzoek. Het voordeel van webscraping is dat je actuele gegevens automatisch kunt verzamelen, zonder dat je alles handmatig hoeft over te typen.

De gegevens over Kamerleden bieden veel mogelijkheden voor onderzoek en analyse. Denk bijvoorbeeld aan vragen als:

- Hoe is de leeftijdsverdeling binnen de Tweede Kamer?
- Hoe lang zitten de huidige Kamerleden gemiddeld in de Kamer?
- Hoe is de vertegenwoordiging per partij verdeeld in de Kamer?

Door deze data te scrapen, kun je dus niet alleen leren hoe webscraping technisch werkt, maar ook hoe je de verkregen informatie kunt toepassen voor maatschappelijke of journalistieke analyses.

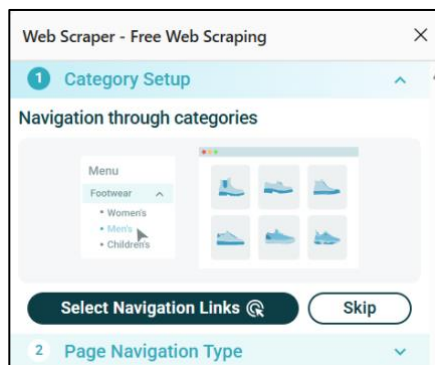
Stap 2: Categorie instelling

De *Category Setup* bepaalt hoe de scraper door verschillende categorieën van een website navigeert voordat hij data gaat verzamelen. Als je bijvoorbeeld alle schoenen wilt scraperen van een webshop, wil je dat de scraper eerst door de categorieën: heren, dames en kinderen bladert, en daarna binnen elke categorie de producten uitleest. In dit menu leer je de scraper waar hij moet klikken om door de categorieën van een site te gaan.

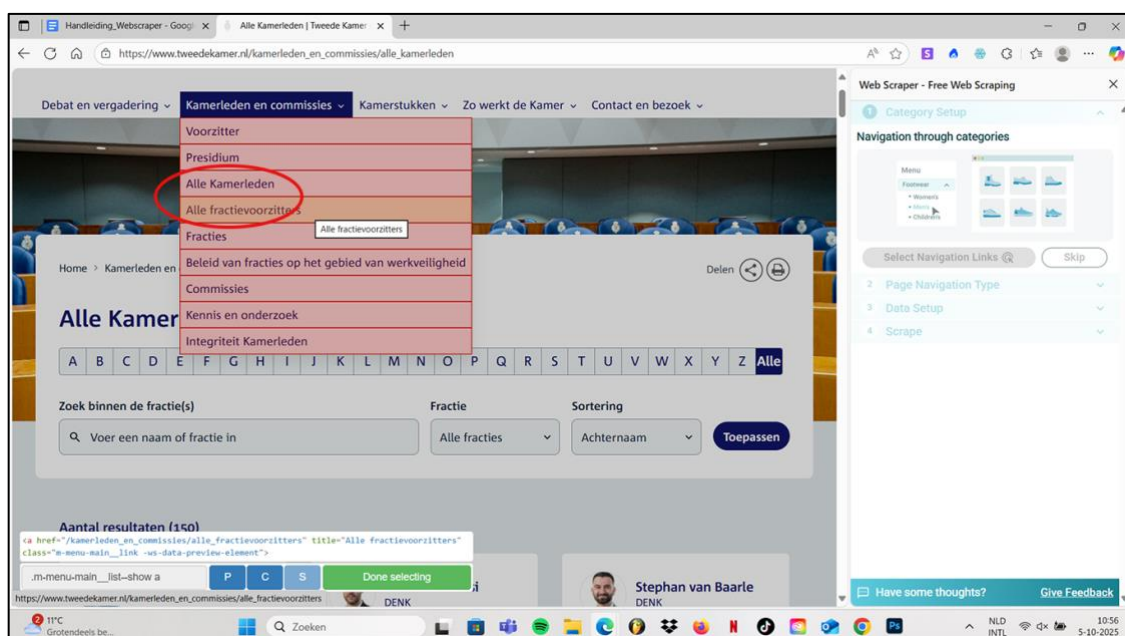
Voor het voorbeeld Alle Kamerleden is dat echter niet relevant. Alle Kamerleden staan al binnen dezelfde categorie op één pagina. Klik daarom op de knop *Skip*.

Stel je wil naast alle Kamerleden toch ook de Voorzitter, Presidium, Alle Fractievoorzitter, etc. meenemen:

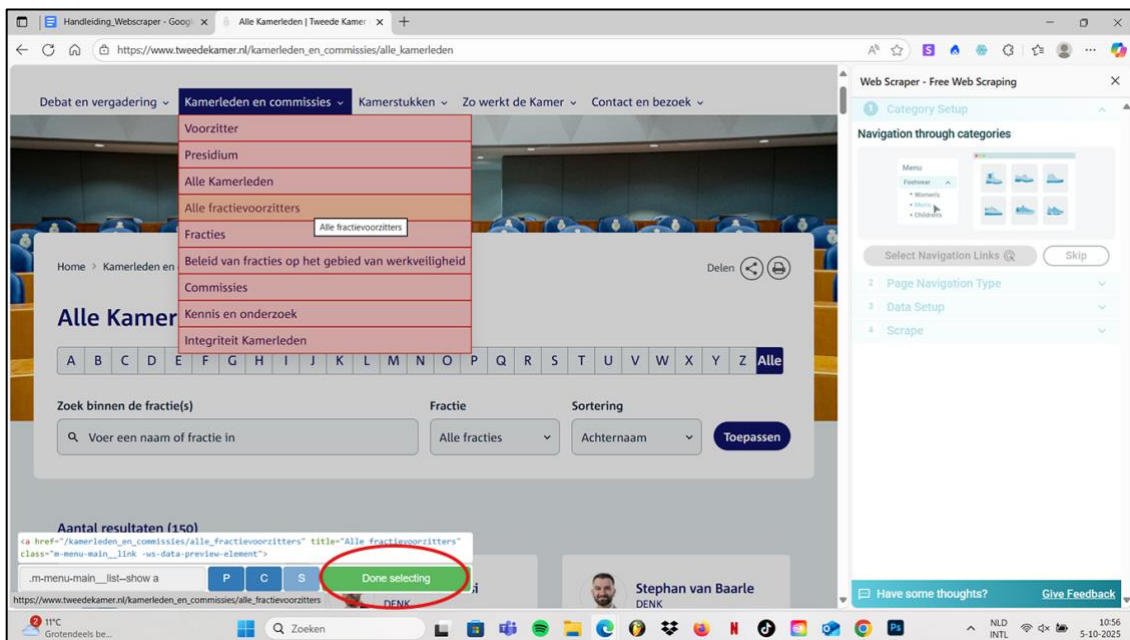
1. Klik op de knop *Select Navigation Links*.



2. Vervolgens klik je op een van de categorieën van het rijtje dat je mee wilt nemen in je dataset. Deze categorie verandert van een geel blokje in een rood blokje. Vervolgens klik je op een andere categorie uit dezelfde rij. Zo leer je de extensie welke knoppen hij moet aanklikken om de categoriepagina's af te gaan voordat hij gaat scraperen.



3. Klik op de knop *Done selecting* als de gehele rij geselecteerd is.



LET OP: Bij dit voorbeeld zul je echter merken dat de categorieën NIET goed worden herkend door de extensie. Dat komt door een designfout in de extensie. Deze website over de Kamerleden beschikt namelijk over een 'dropdownmenu' of 'hover-menu' (net zoals de meeste websites). Zodra je de muis beweegt om op *Done selecting* te klikken, sluit het categoriemenu van de website. Daardoor kan Web Scraper die links niet zien en opslaan binnen de extensie. In deze handleiding kiezen we voor *Skip* om verder te gaan..

Stap 3: Type pagina navigatie

In deze stap geef je aan hoe Web Scraper de data moet verzamelen van de website.

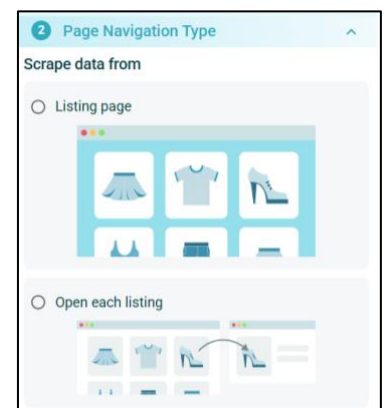
1. Open *Page Navigation Type* door er op te klikken.

Er zijn twee opties: Rechtstreeks van de overzichtspagina (Listing Page) of door elke afzonderlijke pagina van een Kamerlid te openen (Open each listing)

Je kiest voor de optie *Listing page* als je alle gewenste gegevens die je wilt scrapen al zichtbaar zijn op de overzichtspagina. In het voorbeeld met de pagina van de Kamerleden is dat het geval. Het voordeel is dat het scrapen sneller wordt uitgevoerd door de extensie.

Stel er staan meer details per item die je wilt verzamelen die pas zichtbaar zijn nadat je op een item klikt, kies je voor *Open each listing*. Het nadeel is dat het scrapen langer duurt. Dit is niet van toepassing voor het scrapen van data over Kamerleden.

2. Kies voor *Listing page* en klik vervolgens op *continue*.



Stap 4: Data instelling

In deze stap geef je aan welke data je wilt dat de extensie gaat verzamelen en orden je de data alvast in een tabel.

1. Open *Data Setup* door er op te klikken.

Er komt een tabel tevoorschijn met de namen van Kamerleden, foto's, categorie namen, aantal dagen, etc. De extensie heeft alle elementen op de pagina in het schema gezet. Nu moet je de extensie nog duidelijk maken welke elementen er verzameld moeten worden en wat voor soort element het is. De tabel ziet er nog rommelig uit en ga je dus eerst opschonen. Je kan de tabel groter en kleiner maken door de extensie aan de linkerkant vast te pakken en te verslepen.

The screenshot shows the Web Scraper interface with the 'Data Setup' step selected. The table displays 100 out of 150 records. The columns are: data, image, Leefijd, Leefijd.2, data2, data3, Leefijd.3, data4, and data5. The data rows include names like Max Aardema, Ismail el Abbasi, Stephan van Baarle, Mpanzu Bamanga, Thierry Baudet, Bente Becker, Sandra Beckerman, Martin de Beer, and Harry Bevers, along with their respective ages and locations.

2. Verwijder alle rijen waarvan je de data niet nodig hebt. In dit geval geldt dat voor de rij met de afbeeldingen van Kamerleden. Daarnaast zijn er drie verticale rijen met alleen de woorden van de categorie die we nodig hebben ontstaan. Deze mogen ook worden verwijderd. Klik op het kruisje om een rij te verwijderen. Zie afbeelding hieronder.

The screenshot shows the Web Scraper interface with the 'Data Setup' step selected. The table displays 100 out of 150 records. Red 'X' marks are placed over the 'image', 'data2', 'data3', and 'data5' columns, indicating they are to be removed.

Vervolgens geven we de overgebleven rijen de juiste naam. In de eerste verticale rij staan alle namen van de Kamerleden.

LET OP: zorg ervoor dat je eerst alle rijen hebt verwijderd die je niet nodig hebt voordat je de namen verandert. Als je rijen verwijdert nadat je de rijen namen hebt gegeven, veranderen de namen weer.

3. Klik op de bovenste grijze cel van deze rij om de naam van deze rij te veranderen naar 'naam Kamerlid'.

The screenshot shows a web scraper interface with a table of 100 records. The table has the following columns: **data**, **Leeftijd**, **Leeftijd.2**, **Leeftijd.3**, and **data5**. The first row of data is: Max Aardema, 670 dagen, 62 jaar, Drachten, PVV. The 'data' column is highlighted with a red circle.

4. Geef elke verticale rij de bijpassende naam: partij, woonplaats, leeftijd, etc. De tabel moet er zo uit komen te zien:

The screenshot shows the same web scraper interface, but the table has been updated with the following columns: **Naam Kamerlid**, **Anciënniteit**, **Leeftijd**, **Woonplaats**, and **Partij**. The table is highlighted with a red border. The first row of data is: Max Aardema, 670 dagen, 62 jaar, Drachten, PVV.

5. Controleer of alle data die je nodig hebt in de tabel staan en of de namen van de rijen correct zijn. Klik dan op *continue*.

Stap 5: paginering en scrollen

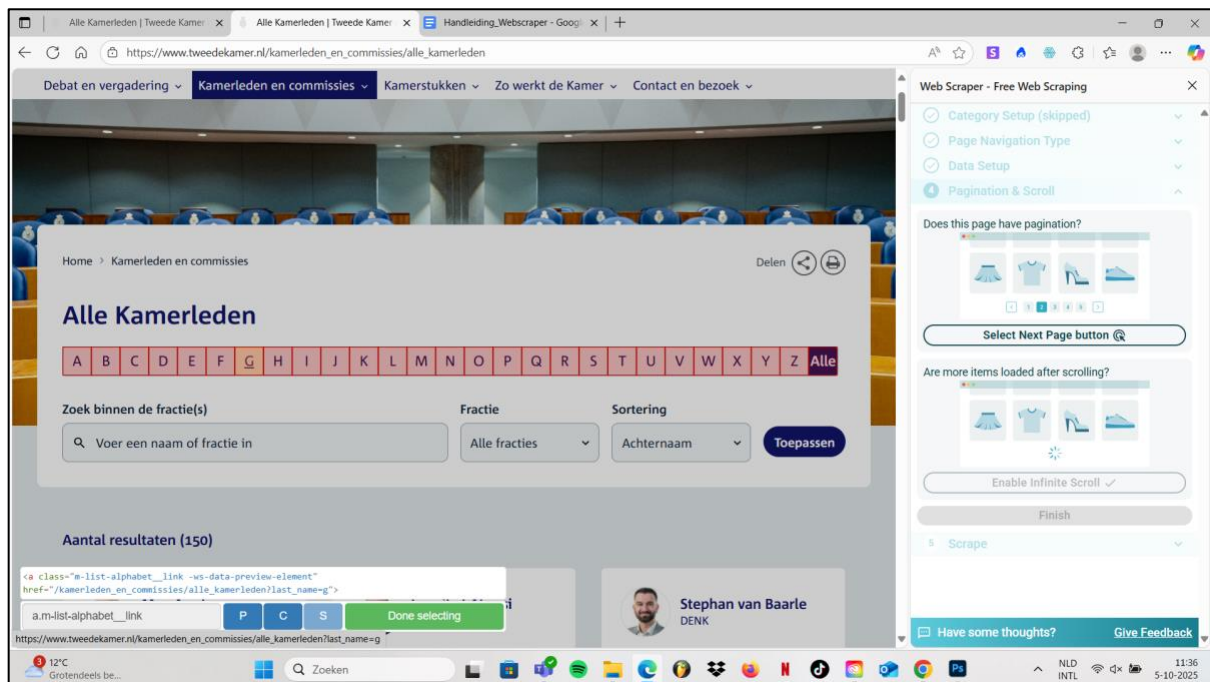
De Webscraper extensie weet nu welke elementen er gescrapet moeten worden. Nu volgt de laatste stap.

1. Open *Pagination & Scroll* door er op te klikken.

Websites met producten, items of in dit geval Kamerleden staan soms verdeeld op meerdere pagina's of op één lange pagina waar alle elementen op staan. Om van alle 150 Kamerleden data te verzamelen moet je de extensie eerst leren of alle Kamerleden verdeeld zijn over meerdere pagina's (Select Next Page button) of dat alle Kamerleden op één pagina staan (Enable Infinite Scroll). In dit voorbeeld kunnen beide opties.

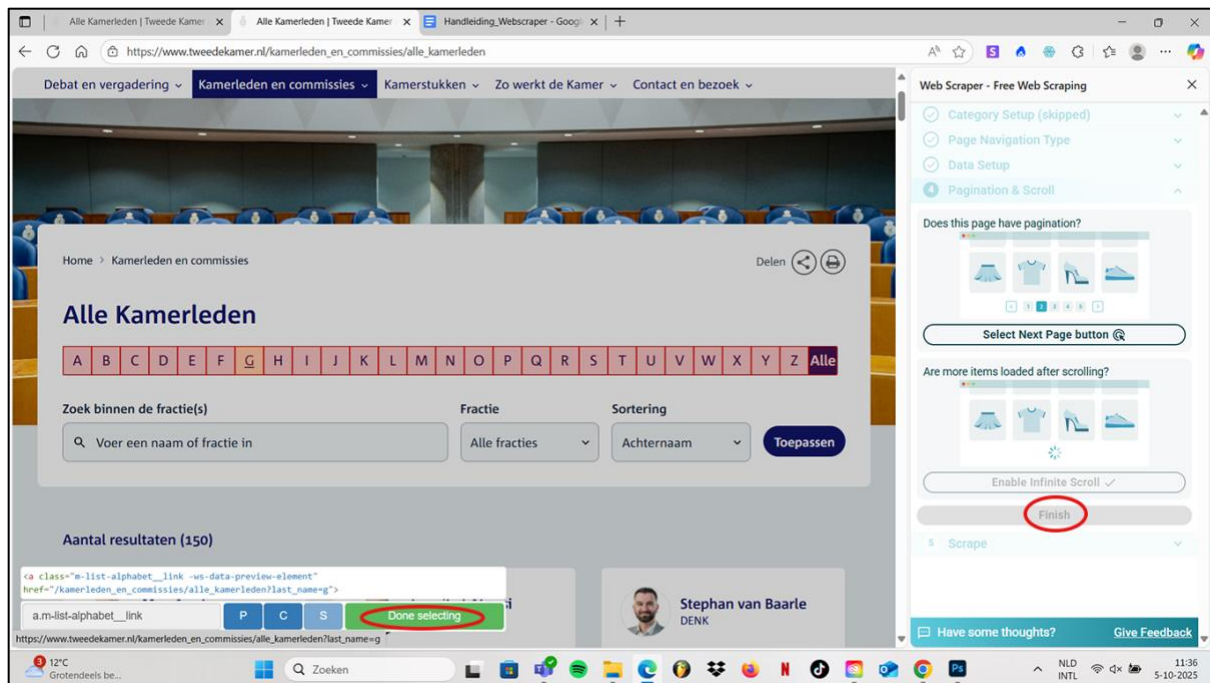
De Kamerleden staan op de website gesorteerd op de beginletters van hun naam. Elke letter is een pagina. Stel, je wilt dat de Web Scraper elke pagina af gaat om de data te scrapen, moet je de webscraper eerst leren hoe hij langs elke pagina gaat van elke beginletter.

2. Klik op *Select Next Page button*.
3. Selecteer een van de blokjes met daar in een letter totdat hij van geel naar rood verandert.
4. Klik vervolgens nog een blokje aan om de extensie duidelijk te maken dat al deze blokjes moeten worden meegenomen.



5. Klik op *Done selecting*.

6. Klik op *Finish*

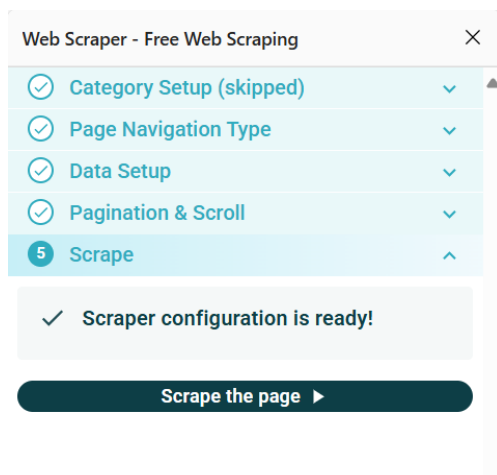


LET OP: Aangezien alle Kamerleden ook op één pagina staan genaamd 'Alle', kun je ook kiezen voor de eenvoudige optie *Enable Infinite Scroll*. Vervolgens klik je op Finish.

Stap 6: Scrapen en Excel

De extensie weet nu precies welke data waar vandaan moet worden gehaald.

1. Klik op de knop *scrape the page* om het scrape proces te starten.



Er opent een nieuw tabblad waar de extensie aan het scrapen is. Kom hier NIET aan en laat de extensie het proces afmaken. Dit kan even duren afhankelijk van de hoeveelheid data die er gescrept moet worden.

LET OP: Als de Web Scraper klaar is met scrapen, klik je NIET op de knop *Done*. Je bent dan alle verzamelde data kwijt.

2. Als het scrapen klaar is, klik je op de knop *.XLSX* om de verzamelde data geordend in een Excel bestand te downloaden.

3. Wanneer dat gelukt is, kun je op *Done* klikken.

Vervolgens kan je aan de gang met je verzamelde data in Excel.

Nadelen Web Scraper extensie

Hoewel het scrapen van websites door middel van de extensie makkelijk is voor gebruikers zonder kennis over HTML-codes en programmeren, is de AI-detectie niet perfect.

Dynamische sites (heavy JavaScript) kunnen lastiger zijn om te scrapen. De AI is niet altijd krachtig genoeg om de dynamische velden te herkennen. Dit zou nog met de hand moeten worden gescrapet. In het volgende hoofdstuk wordt dat uitgelegd aan de hand van het F12 menu.

Daarnaast zit er een ontwerpfout in het menu Category Setup. De meeste websites hebben namelijk een 'dropdownmenu' of 'hover-menu'. Zodra je de muis beweegt om op *Done selecting* te klikken, sluit het categorie menu. Daardoor kan Web Scraper die links niet zien en opslaan. Kortom verwacht de extensie vaste links, maar de meeste sites maken gebruik van dynamische links.

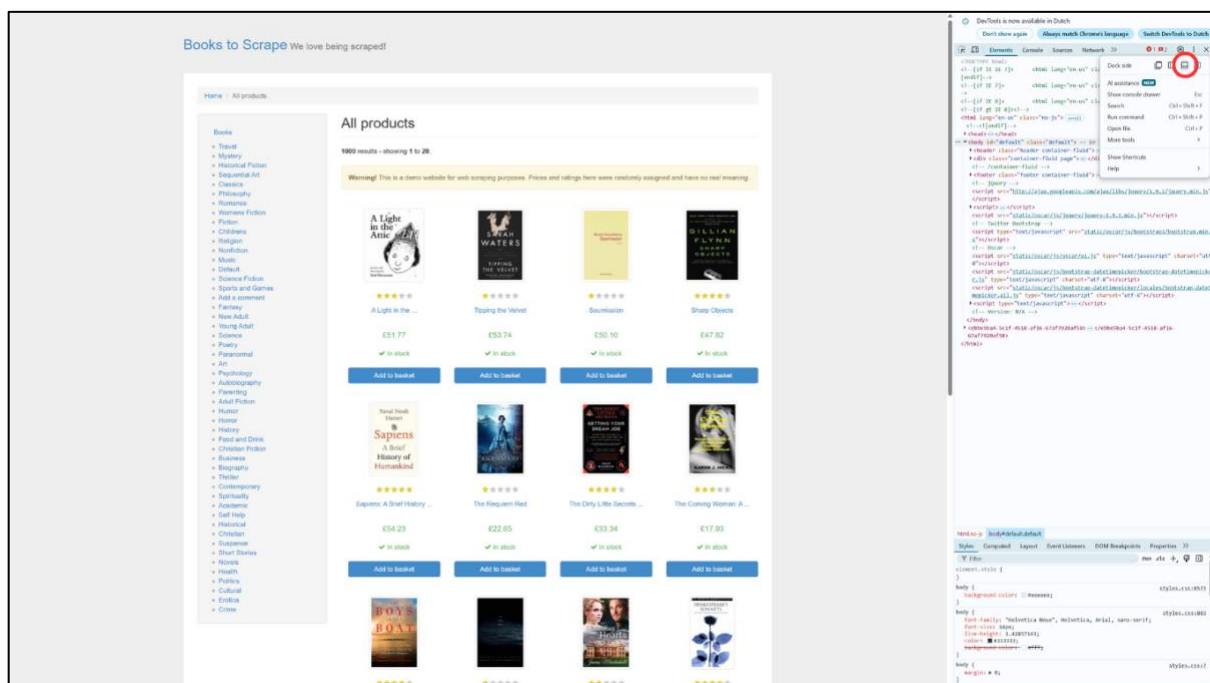
Stap-voor-stap voorbeeldscrape (F12 versie)

Geschreven door Joas

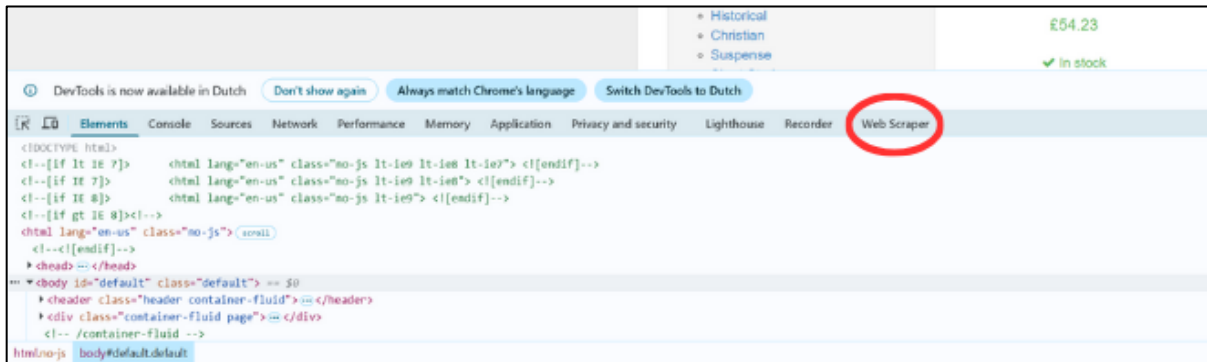
Met de F12-versie van Webscraper.io kan er een duidelijk beeld verkregen worden van hoe een webpagina is opgebouwd. Dit maakt het makkelijk om precies aan te wijzen welke informatie er verzameld moet worden. De selecties kunnen direct getest en bijgesteld worden, wat ervoor zorgt dat sitemaps goed blijven werken. Deze aanpak geeft volledige controle over hoe je door een website navigeert. Denk bijvoorbeeld aan het klikken op knoppen voor de volgende pagina, het omgaan met pop-ups, of het laden van nieuwe informatie zonder de pagina te verversen. Dit is vooral handig bij ingewikkelde websites. Je bent dus niet afhankelijk van automatische keuzes. Dit vermindert de kans op fouten en maakt het oplossen van problemen met sitemaps veel eenvoudiger, omdat alles in één omgeving gebeurt. In het volgende deel wordt de basis van Webscraper.io uitgelegd door via F12 Developer Tools te werken.

Instructies

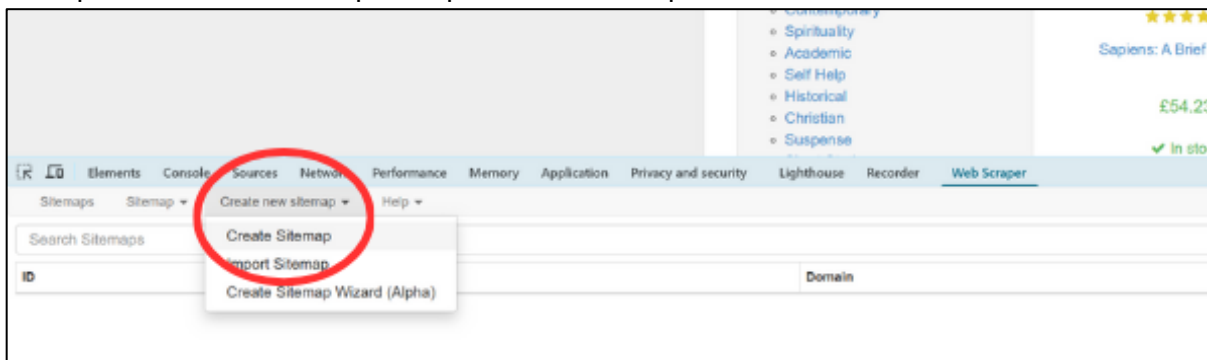
Ga naar books.toscrape.com en druk op F12. Aan de rechterkant opent het inspecteren menu. Klik vervolgens op de drie puntjes rechtsboven en dan op het met rood omcirkelde teken ("Dock to bottom"). Webscraper werkt namelijk niet volledig als deze aan de rechterkant van het scherm staat.



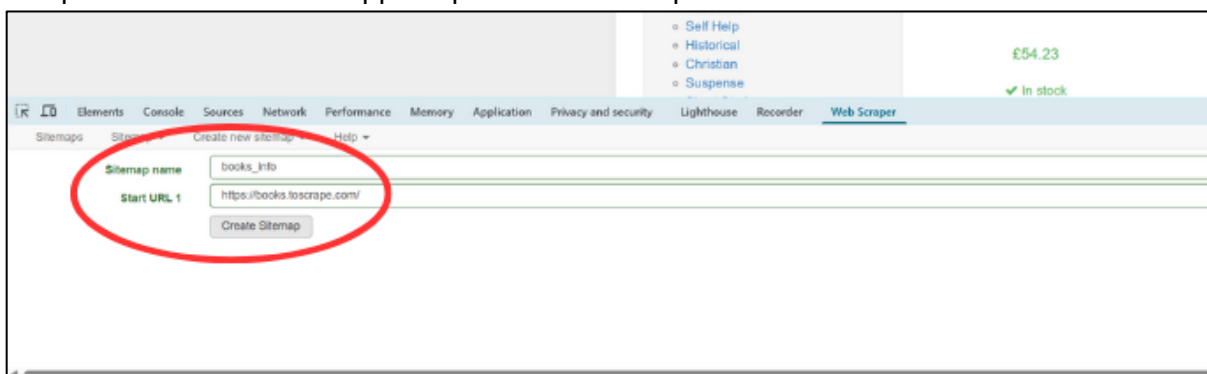
Klik vervolgens op “Webscraper”



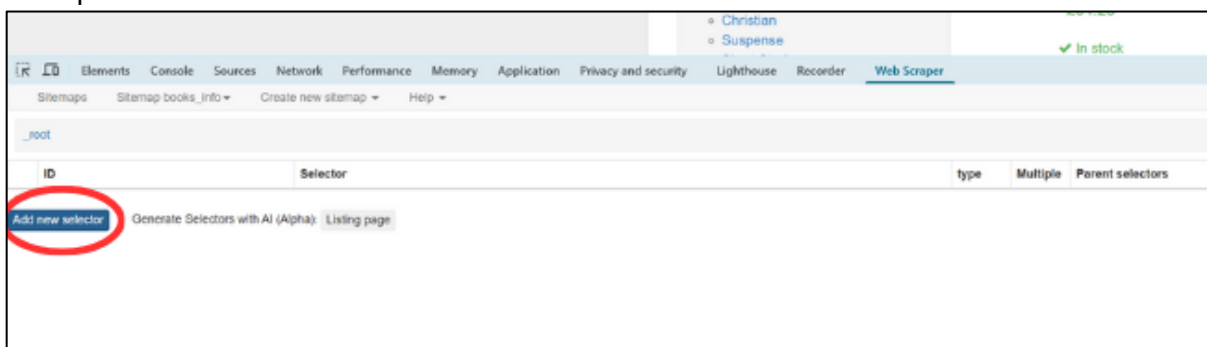
Klik op “Create new sitemap” en op “Create Sitemap”



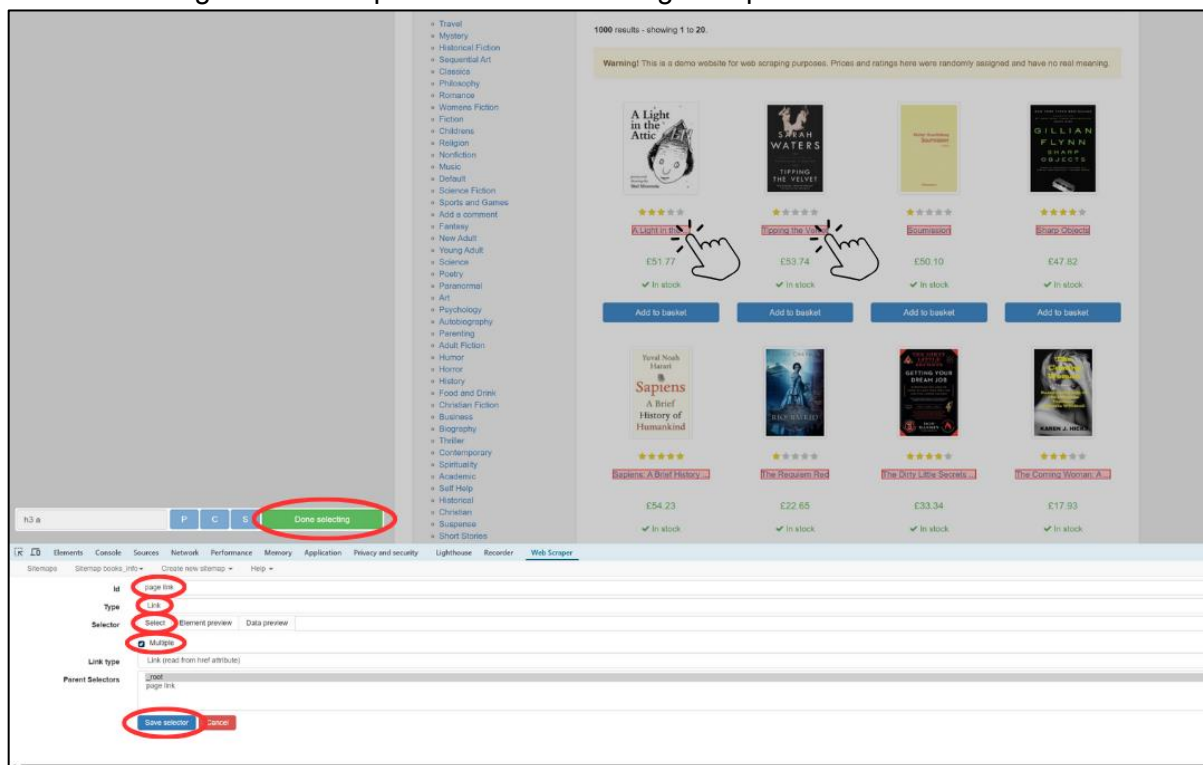
Vul bij “Sitemap name” een titel in en bij “Start URL 1” de URL van de pagina waarop je start. Let op bij het invoeren van de titel dat Webscraper moeilijk kan doen met hoofdletters en spaties. Klik na deze stappen op “Create Sitemap”.



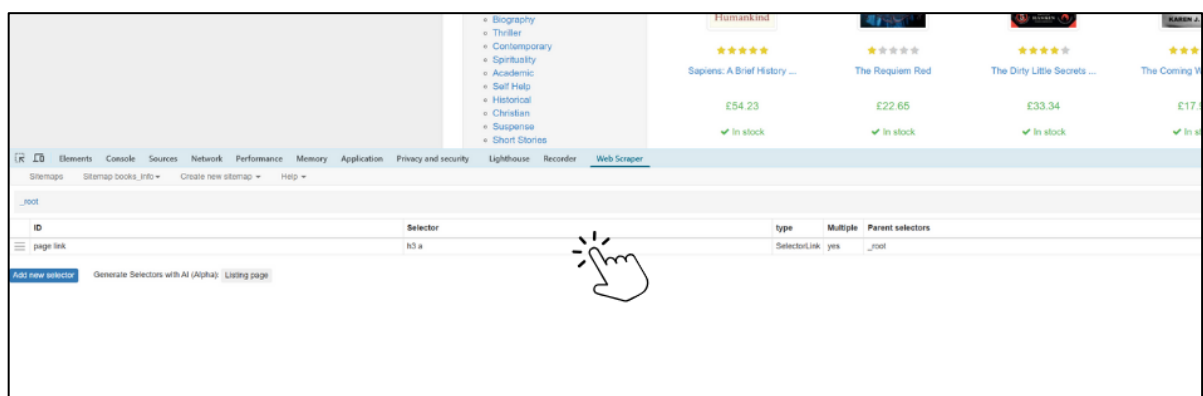
Klik op “Add new selector”.



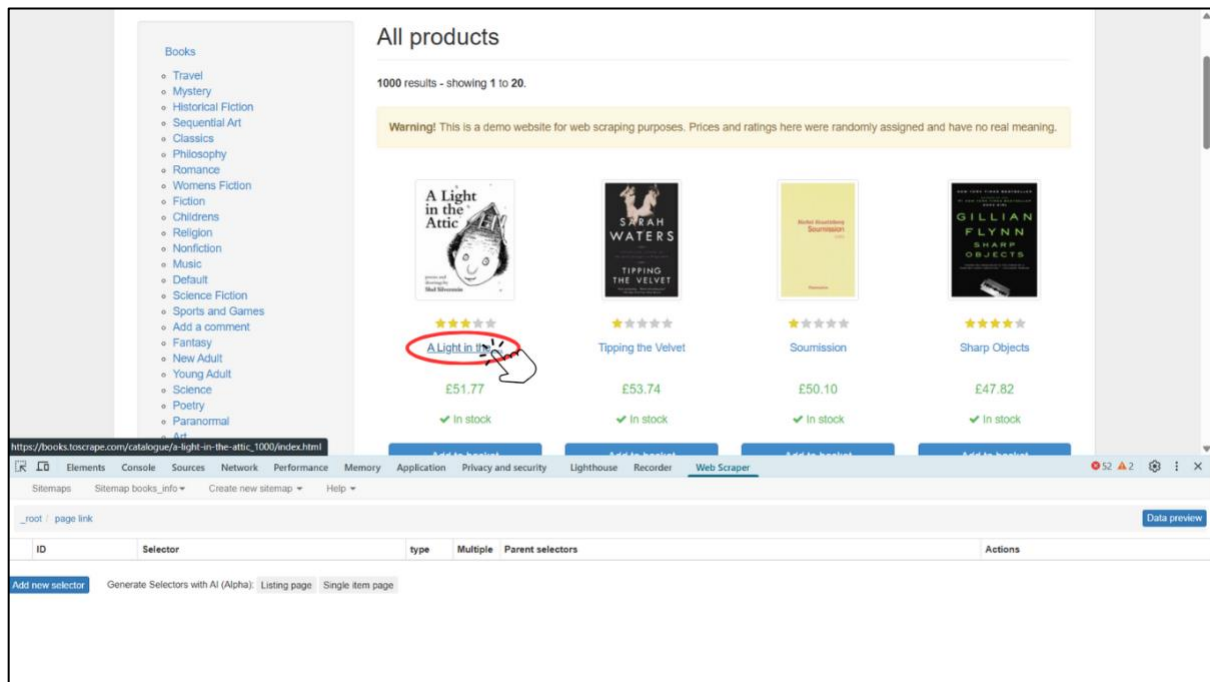
Vul bij “Id” de naam in van datgene wat je wilt selecteren. Bij “Type” kan je kiezen voor meerdere opties. In dit geval willen we de link selecteren. Er kan hier ook text of bijvoorbeeld een afbeelding geselecteerd worden. Klik bij “Selector” op “Select” en dan kan je elementen op de pagina selecteren. Bij het klikken van de eerste twee titels vult het programma automatisch in dat je de rest van die titels ook geselecteerd wilt hebben. Klik daarna op “Done selecting”. Vink “Multiple” aan en klik vervolgens op “Save selector”.



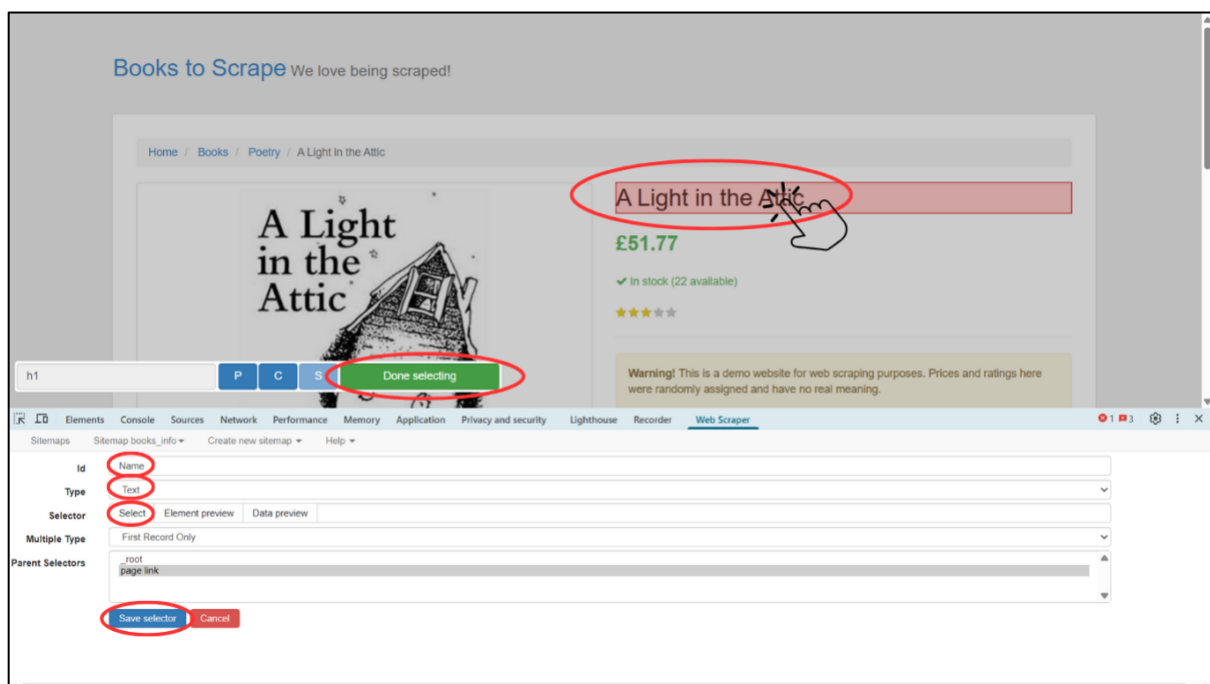
We zeggen nu eigenlijk: start met ‘root’ en ga daarna naar elk van die links kijken. Dubbelklik nu op het balkje van “page link”.



Open de onderstaande pagina om de informatie te selecteren die we willen scrapen.



Klik op "Add new selector". Vul de "Id" in. Bij "Type" gaan we nu voor "Text". Klik op "Select" en klik op de Titel van de pagina. Klik op "Done selecting" en op "Save selector". Herhaal deze stappen voor Titel, Product Type, Price (excl. tax), Price (incl. tax), Availability and Number of reviews.



Nu staat er onder "pagelink" het volgende rijtje aan informatie.

The screenshot shows the 'Product Information' section with the following data:

Field	Value
UPC	a897fe39b1053632
Product Type	Books
Price (excl. tax)	£51.77
Price (incl. tax)	£51.77
Tax	£0.00
Availability	In stock (22 available)
Number of reviews	0

Below this is a table of selectors for the 'page link' path:

ID	Selector	type	Multiple	Parent selectors	Actions
Name	h1	SelectorText	no	page link	Element preview, Data preview, Edit, Delete
Product Type	th.contains('Product Type') + td	SelectorText	no	page link	Element preview, Data preview, Edit, Delete
Price (excl. tax)	tr.nth-of-type(3) td	SelectorText	no	page link	Element preview, Data preview, Edit, Delete
Price (incl. tax)	tr.nth-of-type(4) td	SelectorText	no	page link	Element preview, Data preview, Edit, Delete
Availability	th.contains('Availability') + td	SelectorText	no	page link	Element preview, Data preview, Edit, Delete
Number of reviews	th.contains('Number of reviews') + td	SelectorText	no	page link	Element preview, Data preview, Edit, Delete

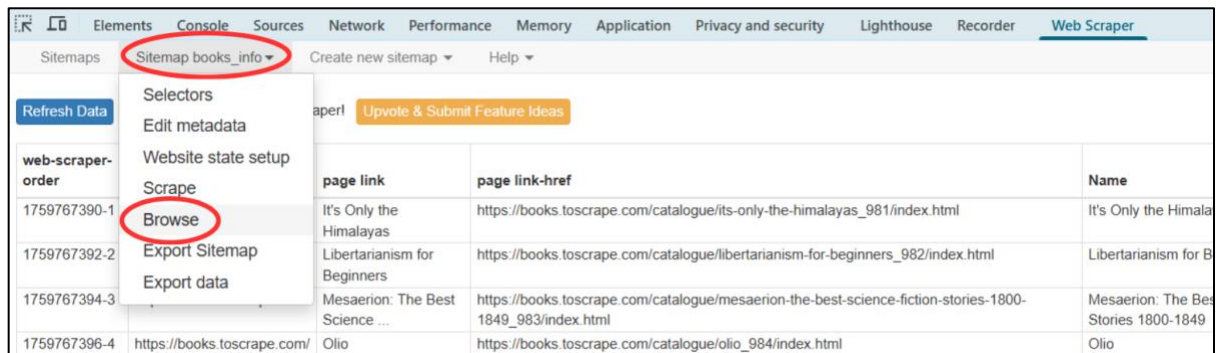
Klik nu op "Sitemap books_info" en vervolgens op "Scrape" om de informatie van de website te scrapen.

The screenshot shows the 'Sitemap books_info' dropdown menu open, with the 'Scrape' option highlighted. The menu options are: Selectors, Edit metadata, Website state setup, Scrape, Browse, Export Sitemap, and Export data.

Er opent een pagina waar een aantal waarden ingesteld kan worden over hoe lang er bijvoorbeeld gewacht wordt tussen pagina's tijdens het scrapen, zodat pagina's volledig geladen zijn. Klik vervolgens op "Start scraping".

The screenshot shows the 'Start scraping' button highlighted in the Web Scraper interface. The button is located at the bottom of the page, below the 'Request interval (ms)' and 'Page load delay (ms)' settings.

Nu opent er een ander venster en start het scrapen. Wanneer dit klaar is kan je door de data scrollen door te klikken op “Sitemap books_info” en dan op “Browse”.



Als de data compleet is, kan je klikken op “Export data” en kan je de data downloaden als .XLSX of als .CSV.

Problemen en Fouten

Geschreven door Floris

Net als elk ander programma is het mogelijk dat er dingen fout gaan. Zo kan je per ongeluk een column verwijderen die je toch nog wilt gebruiken, dit is makkelijk op te lossen door de set te her genereren. Het is ook mogelijk de verkeerde elementen te selecteren op een pagina. Het is in dat geval mogelijk om nogmaals op het element te klikken om het weer te deselecteren.

Als een pagina niet goed doorlaad, kan het komen door de hoeveelheid links die de Webscraper opent. In dit geval is het handig om je paginations te verminderen en je data pagina per pagina te verzamelen. Een ander gevolg van een pagina met veel links scrapen is dat de site je IP adres tijdelijk blokkeert. In dit geval is het handig om zo'n 10 minuten te wachten en het nog een keer te proberen. Zorg er voor dat je ook hier minder paginations en links gebruikt. Anders is het mogelijk dat je weer wordt geblokkeerd.

Conclusie

Geschreven door Kellen

Webscraping vormt een krachtig hulpmiddel in een tijd waarin data de basis is voor inzicht, onderzoek en verantwoording. In deze handleiding heb je geleerd hoe Webscraper.io werkt. Van het installeren van de extensie tot het exporteren en analyseren van data. De toegankelijkheid van de tool maakt het mogelijk om zonder programmeerkennis grote hoeveelheden web informatie te verzamelen. Daarmee biedt Web Scraper niet alleen een praktische vaardigheid, maar ook een nieuwe manier van denken over digitale bronnen: websites worden datasets en informatie verandert in analyseerbare kennis.

Toch reikt het gebruik van webscraping verder dan alleen techniek. Tijdens het werken met Webscraper.io leer je pas echt goed data scrapen als je ook begrijpt hoe een website technisch werkt, hoe de tekst, knoppen en tabellen in de code (HTML) zijn opgebouwd. Door dat te snappen kun je betere keuzes maken in wat je verzamelt. Het toevoegen van een HTML-verdieping in deze handleiding helpt gebruikers niet alleen om complexere websites te scrapen, maar ook om meer controle te krijgen over de kwaliteit en betrouwbaarheid van de data. Voor gebruikers die zich verder willen verdiepen in de mogelijkheden van Webscraper.io, met name in het gebruik van de F12-versie en het werken met HTML-codes, verwijzen we graag naar de officiële website pagina met tutorials:

<https://webscraper.io/tutorials>

In deze handleiding is bewust gekozen om bij de basis te blijven, zodat iedereen de benodigde stappen van webscraping onder de knie krijgt. Op de tutorial pagina vind je uitgebreide uitleg, voorbeelden en video's die helpen om de meer geavanceerde functies van Webscraper.io beter te begrijpen en toe te passen.

Tot slot vraagt webscraping om bewustzijn en verantwoordelijkheid. Data is nooit neutraal en het geautomatiseerd verzamelen ervan roept vragen op over privacy, eigendom, en ethiek. Door zorgvuldig om te gaan met bronvermelding, context en juridische kaders, kun je webscraping inzetten als instrument voor transparantie, kennisdeling en data gedreven inzichten. Daarmee vormt het een relevant instrument binnen hedendaags onderzoek en journalistiek.