Albert Heijn Web Scraper

Description

Voor onze applicatie hebben we een web scraper nodig voor onze eigen applicatie. De applicatie heeft namelijk een front end met daarop een visuele interface voor het aanpassen van boodschappenlijstjes. Hiervoor moeten er producten te kiezen zijn uit een ruime catalogus. De applicatie valt te bedienen via een server met api calls.

De web scraper haalt ongeveer 22000 producten op van de catalogus van de Albert Heijn. De informatie over de producten bevat onder andere: naam, ID, prijs, quantiteit, categorie en een url voor het plaatje.

De ID kan gebruikt worden als ID in het database, wat er voor zorgt dat er een directe link is tussen de ID in het database en die van AH.
De rest van de informatie gaat gebruikt worden in onze eigen applicatie om het product te kunnen gebruiken.

- Categorie kan gebruikt worden om te sorteren.
- Quantiteit is handig voor het inschatten van de hoeveelheid product per euro.
- De url voor het plaatje kan gebruikt worden voor het visuele deel van de applicatie.
- Prijs en naam spreken voor zich.

Prerequisites

- Python 3

Installation

1. Clone repository
2. Command “pip install -r requirements.txt”
3. Command “python app.py”
    - Runs on port 5002

End-points

De scaper moet op een commando kunnen werken vanuit een andere webserver. Dus de scraper heeft een aantal endpoints om de producten te controleren.

“/api/scraper/start” End point om het process te starten, Scraper geeft feedback over het process wanneer het bezig is en niet het process niet opnieuw kan starten.

“/api/scraper/status” End point voor het ophalen van de status (stopped, running, stopping, storingInDatabase)

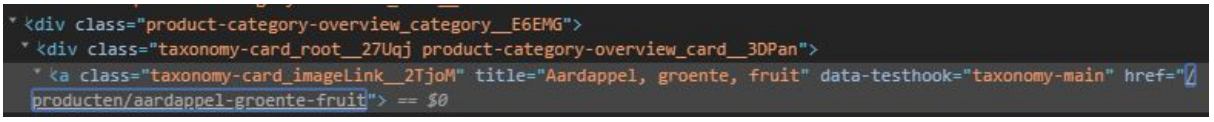
“/api/scraper/stop” End point om het process te stoppen

Functionality

Collecting from main page

De web scaper begint bij de website: “https://www.ah.nl/producten”.
De pagina bestaat uit meerdere categorieën die opgehaald worden door naar de class te zoeken in de return value van de get request. De class name is `class="product-category-overview_category__E6EMG"`

Daarna wordt er altijd gezocht naar de eerst volgende url with href.



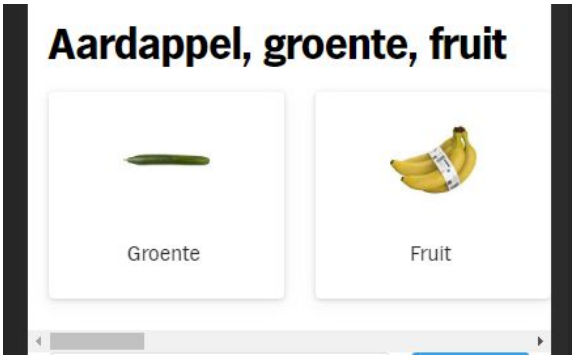
De link wordt opgeslagen in een lijst die wordt gebruikt om subcategorieën op te halen.

Collecting from subpages

Voor het ophalen van de subpages wordt de volgende class name gebruikt: `class="taxonomy-sub-selector_child__1yS69`

De request moet van een mobiel komen, want de pagina laad in een mobile version alle categorieën. Anders moet er eerst geklikt worden om de categorieën te krijgen.

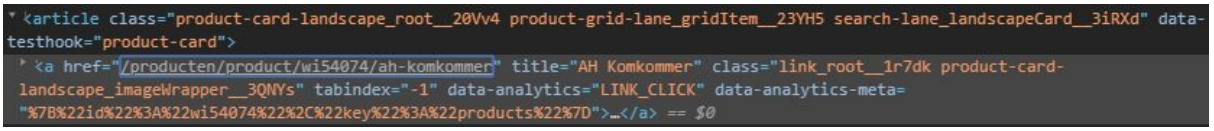
Desktop version



Mobile version

Collecting from products page

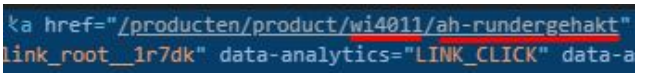
Voor de producten wordt de volgende class gebruikt:
`class="product-card-portrait_root__2J9q_ product-grid-lane_gridItem__23YH5"`

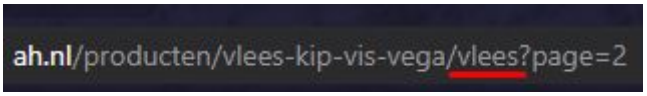


De request heeft als extra info “?page=100”, wat er voor zorgt dat alle paginas worden geladen. Dit gaat helaas niet verder dan 1000 items op een pagina. Daarom kon er niet direct gezocht worden op de categorie. Dat is waarvoor de subcategorie zijn opgehaald.

De ID in de link kan gebruikt worden in het database, want de ID kan direct gebruikt worden om informatie op te halen van de website. Voor de link hieronder is het nummer wi54074
https://www.ah.nl/producten/product/wi54074/ah-komkommer

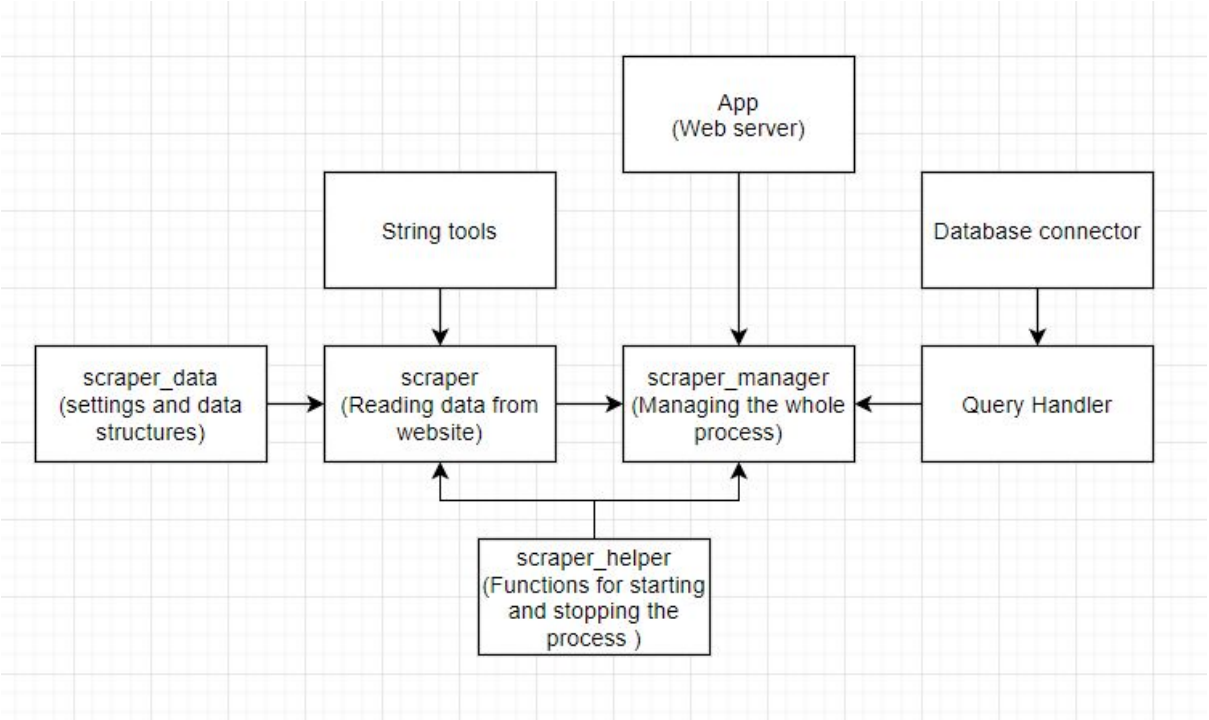
Op deze product page kan de informatie die nodig is worden opgehaald zoals de plaatjes, de categorie, prijs en de quantiteit.





Wanneer dit voor alle producten is gedaan wordt het in de database gezet.

Overview



Resultaten

De web scraper vindt ongeveer 22000 producten van de website en haalt alle gewenste informatie op van de producten (naam, ID, prijs, plaatje, quantiteit, category).
De scraper doet hier ongeveer 20 minuten over. De resultaten worden in een database opgeslagen waar ze gebruikt kunnen worden voor onze eigen interface.

```
------------------------------------------------------------
total products collected: 21882
total time: 1070.5244567394257
------------------------------------------------------------
```

NOTE: Dit process kan het best maar een paar keer per dag worden uitgevoerd worden per IP, want als het te vaak wordt uitgevoerd, wordt tijdelijk de snelheid van de connectie vermindert door AH.