Webscraper vakantieparken

Realisatiedocument

Bachelor in de Toegepaste informatica
keuzerichting Application development

Wouter Vandueren

Academiejaar 2022-2023

Campus Geel, Kleinhoefstraat 4, BE-2440 Geel

LID VAN ASSOCIATIE KU LEUVEN

THOMAS MORE

Table of Contents

1

Realisatiedocument 1

Inleiding 3

Toelichting gebruikte technologieën 3

Webscrapers 3

Datapunten 5

Webapplicatieframework 6

Realisatie 7

Meehelpen en kijken met deployment websites 7

Webscraper voor 1 specifiek vakantiepark van Center Parcs 7

Webscraper voor alle vakantieparken van Center Parcs 8

Webscraper voor alle vakanties/aanbiedingen van Center Parcs 9

Webscraper voor alle bungalows van Center Parcs 10

Webscraper voor alle bungalows van Belvilla 11

Wegschrijven data van webscrapers 11

Flask applicatie voor webscrapers 12

Ondervonden problemen en oplossingen 14

Verbindingssnelheid internet 14

Communicatie met klant (Molenheide) 14

Conclusie 14

Inleiding

In dit document ga ik mijn stageopdracht wat meer toelichten en beschrijven wat ik gerealiseerd heb. Eerst ga ik uitleggen welke technologieën ik gebruikt heb om mijn opdracht te realiseren en waarom ik hiervoor gekozen heb gevolgd door het uitgebreid beschrijven van wat ik gerealiseerd heb op mijn stage. Verder ga ik ook nog beschrijven welke problemen ik zo al ondervonden heb en hoe ik deze opgelost heb. Dit document eindig ik met de conclusie waarin ik beschrijf wat ik zo al geleerd en ervaren heb gedurende mijn stage.

Voor mijn stageopdracht moest ik een webscraper programmeren die de prijzen van verschillende bungalowparken in België en Nederland kon gaan scrapen en deze vergelijkt met elkaar. De klant (Molenheide) gebruikt dan deze prijzen om te vergelijken met hun prijzen voor de bungalows zodat ze een competitieve prijs hebben. In het geval van kortingen zoals last-minute aanbiedingen willen ze kijken hoeveel korting de concurrentie plaatst op hun bungalows en hoe deze is opgesteld.

Toelichting gebruikte technologieën

Voor ik gestart ben met het realiseren van de webscraper heb ik eerst wat opzoek - en vergelijkingswerk gedaan tussen de verschillende tools om te scrapen. Ook keek ik welke datapunten (in dit geval vakantieparken) het beste zijn om te gaan scrapen.

Webscrapers

Eerst heb ik wat opzoekwerk gedaan over welke verschillende tools je allemaal kan gebruiken om websites te scrapen via de frontend. Je kan ook websites scrapen aan de hand van API-calls maar op de meeste websites staat hier security op waardoor je geblokkeerd kunt worden.

Ik had al ervaring met selenium omdat ik hier al eens een webscraper mee gemaakt had geprogrammeerd in c#. Ik heb ook wat meer info opgezocht over scrapy en beautifoulsoup. Deze 3 heb ik dan vergeleken.

Scrapy

De prestaties zijn belachelijk snel en het is een van de krachtigste bibliotheken die er zijn. Een van de belangrijkste voordelen van scrapy is dat het bovenop Twisted is gebouwd, een asynchroon netwerkframework, wat betekent dat scrapy het niet-blokkerende mechanisme gebruikt terwijl de verzoeken naar de gebruikers worden verzonden.

De belangrijkste kenmerken van Scrapy zijn -

1. Scrapy heeft ingebouwde ondersteuning voor het extraheren van gegevens uit HTML-bronnen met behulp van XPath-expressie en CSS-expressie.
2. Het is een draagbare bibliotheek, d.w.z. (geschreven in Python en draait op Linux, Windows, Mac en BSD)

3. Het kan eenvoudig worden uitgebreid.
4. Het is sneller dan andere bestaande scraping-bibliotheken. Het kan de websites 20 keer sneller extraheren dan andere tools.
5. Het verbruikt veel minder geheugen en CPU-gebruik.
6. Het kan ons helpen een robuuste en flexibele applicatie te bouwen met een heleboel functies.

Het heeft goede community-ondersteuning voor de ontwikkelaars, maar de documentatie is niet zo geweldig voor beginners omdat het geen beginnersvriendelijke documentatie heeft. Als je te maken hebt met een complexe scraping-operatie die een enorme snelheid en een laag stroomverbruik vereist, dan is Scrapy een goede keuze.

Beautifoul soup

Extraheert snel de gegevens van een bepaalde webpagina. Deze library helpt ons om de gegevens uit HTML- en XML-bestanden te halen, deze library vereist specifieke modules.

De afhankelijkheden van de Beautiful-soep zijn -

1. Er is een library nodig om een verzoek aan de website te doen, omdat deze geen verzoek aan een bepaalde server kan doen. Om dit probleem op te lossen, is de hulp nodig van de meest populaire bibliotheek genaamd Requests of urlib2. Deze libraries zullen ons helpen om ons verzoek aan de server te doen.
2. Na het downloaden van de HTML, XML-gegevens naar onze lokale machine, heeft Beautiful Soup een externe parser nodig om de gedownloade gegevens te ontleden. De bekendste parsers zijn: de XML-parser van lxml, de HTML-parser van lxml, HTML5lib, html.parser.

De voordelen van de Beautifoul-soup zijn

1. Het is gemakkelijk te leren en te beheersen. Bijvoorbeeld als we alle links van de webpagina willen extraheren. Het kan eenvoudig als volgt worden gedaan:

   ```
   from bs4 import BeautifulSoup
   soup = BeautifulSoup(html_doc, 'html.parser')

   for link in soup.find_all('a'): # It helps to find all anchor tag's
       print(link.get('href'))
   ```

1. In de bovenstaande code gebruiken we de html.parser om de inhoud van de html_doc te ontleden. Dit is een van de sterkste redenen voor ontwikkelaars om Beautiful soup te gebruiken als webscraping-tool.
2. Het heeft goede uitgebreide documentatie die ons helpt om de dingen snel te leren.
3. Het heeft goede ondersteuning van de gemeenschap om de problemen op te lossen die zich voordoen terwijl we met deze bibliotheek werken.

Selenium
Selenium is ontworpen om testen voor webapplicaties te automatiseren. Het biedt de ontwikkelaar een manier om tests te schrijven in een aantal populaire programmeertalen zoals C#, Java, Python, Ruby, enz. Dit framework is ontwikkeld om browserautomatisering uit te voeren.

API is erg beginnersvriendelijk, je kunt eenvoudig code schrijven met Selenium.

Het belangrijkste kenmerk van Selenium is -

1. Het kan gemakkelijk werken met Javascript-kernconcepten (DOM)
2. Het kan gemakkelijk AJAX- en PJAX-verzoeken verwerken.

Selenium: Als je te maken hebt met een Core Javascript-website, dan is Selenium de beste keuze. Maar de gegevensgrootte moet beperkt zijn.

Besluit
Uiteindelijk heb ik gekozen om selenium te gaan gebruiken voor het scrapen omdat ik al ervaring heb met deze tool en het ook deze voordelen heeft.

1. Dynamische inhoud: Selenium is ontworpen om te werken met dynamische webpagina's die JavaScript gebruiken om inhoud te genereren. Het kan de interacties van een gebruiker met de webpagina nabootsen, inclusief het uitvoeren van JavaScript-code, het klikken op elementen en het invullen van formulieren. Dit stelt Selenium in staat om de volledige inhoud van een JavaScript-website te laden en te extraheren.
2. Browserautomatisering: Selenium stelt je in staat om een webbrowser automatisch te besturen, zoals Google Chrome, Firefox, Safari, enz. Dit betekent dat je webpagina's kunt openen, door de inhoud kunt scrollen, elementen kunt selecteren en acties kunt uitvoeren die normaal gesproken door een menselijke gebruiker worden uitgevoerd. Hierdoor kun je toegang krijgen tot de volledige functionaliteit van een website, inclusief JavaScript-gestuurde interacties.
3. Brede ondersteuning: Selenium heeft brede ondersteuning voor verschillende programmeertalen, waaronder Python, Java, C#, Ruby en JavaScript. Hierdoor kun je Selenium integreren in je favoriete programmeertaal en het gebruiken om JavaScript-websites te scrapen op de manier die het beste bij jouw behoeften past.
4. Rijke functionaliteit: Selenium biedt een scala aan functies die nuttig zijn bij het scrapen van websites, zoals het zoeken naar elementen op basis van CSS-selectors of XPath, het wachten op bepaalde voorwaarden voordat een actie wordt uitgevoerd, het omgaan met pop-ups en frames, het manipuleren van cookies en nog veel meer. Deze functionaliteiten stellen je in staat om complexe scraping-taken uit te voeren en specifieke gegevens van JavaScript-websites te extraheren.

Datapunten
Vervolgens heb ik aan de klant Molenheide gevraagd welke vakantieparken ze de meeste interesse hadden. Zij gaven een lijst van een tiental vakantieparken en

deze heb ik vergeleken op een paar specificaties (bungalows in België/Nederland, up-to-date site, last-minute aanbiedingen…) en in een spreadsheet gezet. Daaruit bleek dat Center Parcs toch de meest geschikte kandidaat was. Deze is ook een van de grotere concurrenten van mijn opdrachtgever.

naam vakantiepark	bied bungalows/chalets aan	chalets in België/Nederland	Up-to-date site	Last minutes aanbieding	Aantal last-minutes aanbieding	Alle bungalows beschikbaar op 1 pagina	Is een specifiek vakantiepark
Oostappen vakantiepark							
EuroParcs							
Belvilla							
Center Parcs							
Landal Greenparks							
Slagharen							
Anzelhoef							
RCN							
Topparken							
Ardoer							
Roompot							
Bungalowspecials							
Molecaten							
Topparken							

De meeste vakantieparken hadden wel bungalows/chalets in België en Nederland, een up-to-date site en vakanties/aanbiedingen zoals last-minute aanbiedingen. Maar het grote struikelblok was dat niet alle bungalows beschikbaar waren op 1 pagina. Als alle bungalows beschikbaar zijn op 1 pagina vergemakkelijkt dit het proces van de webscraper, anders moet de scraper te veel wisselen van pagina's en dit kan voor problemen zorgen.

Webapplicatieframework

Flask

Flask is een lichtgewicht websysteem dat wordt gebruikt voor het bouwen van webapplicaties in Python. Het werd voor het eerst uitgebracht in 2010 en heeft kritische bekendheid gekregen vanwege zijn moeiteloosheid, flexibiliteit en capabele bruikbaarheid. Flask is ontworpen om ontwerpers een directe en gematigde benadering te bieden voor de ontwikkeling van webapplicaties.

Een van de belangrijkste kenmerken van Flask is dat het een microframework is, wat inhoudt dat het een verwaarloosbare set tools en bibliotheken biedt die essentieel zijn voor het bouwen van een webapplicatie. Dit geeft users de mogelijkheid om de functionaliteit van hun applicatie aan te passen en de gespecificeerde bibliotheken naar behoefte toe te voegen. Flask volgt een uitgemeten plan, waardoor users verschillende uitbreidingen kunnen consolideren om extra bruikbaarheid te bieden, zoals database-integratie, verificatie, vormen en meer.

Flask maakt gebruik van de WSGI (Web Server Gateway Interface) om te communiceren met webservers zoals Apache of Nginx. Het geeft sturende bruikbaarheid, waardoor users URL-regels kunnen karakteriseren die vergelijkbaar zijn met bepaalde capaciteiten in hun toepassing. Dit stelt users in staat om de geschikte code uit te voeren op basis van de door de klant gevraagde URL.

Flask ondersteunt ook de weergave van formaten, waardoor users energieke HTML-pagina's kunnen maken. Het maakt gebruik van het Jinja2-sjabloondialect, dat users in staat stelt lay-outs te maken met factoren, voorwaardelijke

grondgedachten en cirkelontwikkelingen. Dit maakt het voor ontwikkelaars eenvoudig om informatie van hun Python-code door te geven aan de HTML-pagina's die aan de client worden weergegeven.

Met betrekking tot database-integratie ondersteunt Flask verschillende bekende databases zoals SQLite, PostgreSQL en MySQL. Het biedt een eenvoudige manier om door te verbinden met een database en SQL-vragen uit Python-code uit te voeren. Flask heeft ook uitbreidingen zoals SQLAlchemy die een Object-Relational Mapper (ORM) bieden, waarmee ontwerpers databasetoewijzingen kunnen karakteriseren met behulp van Python-klassen en op een objectgeoriënteerde manier met de database kunnen worden verbonden.

Over het algemeen zou Flask een aanpasbaar en effectief websysteem kunnen zijn dat bekend staat om zijn rechtlijnigheid en uitbreidbaarheid. Het is perfect voor het bouwen van kleine tot middelgrote webapplicaties, Restful API's en prototypes. Het geeft users een ongelooflijke kans en controle over hun applicaties, terwijl het eenvoudig te onthouden en te gebruiken is.

Realisatie

Meehelpen en kijken met deployment websites

Gedurende mijn stage heb ik naast mijn stageopdracht ook veel kunnen meekijken met het hele proces van hoe een website ontwikkeld wordt. Van het bespreken met de klant wat ze willen als website tot het ontwikkelen van de website en het online zetten van de website. In het begin van mijn stage waren ze druk bezig met het online zetten van de webshop website van “Balo”, hier hadden ze mijn hulp bij gevraagd omdat de producten nog niet gecategoriseerd waren in de juiste categorieën. Dit heeft me wel een dag werk gekost maar zo leerde ik ook een beetje werken met wagtail cms, dit is een django content management system. Voor de rest heb ik gedurende mijn stage wat meegekeken met de mensen die de frontends maken voor deze websites in vue.js en de backends geprogrammeerd in python met django, wagtail…

Webscraper voor 1 specifiek vakantiepark van Center Parcs

Mijn eerste webscraper spitste zich toe op 1 vakantiepark, namelijk Erperheide van Center Parcs. De data die ik bijhield waren:

- Titel
- Nieuwe prijs
- Aantal personen
- Oppervlakte
- Datum
- Aantal slaapkamers

Ik heb ervoor gekozen om meer data dan enkel de prijs te scrapen zodat je Molenheide hun bungalows niet enkel op vlak van prijs kan vergelijken met de bungalows van Molenheide.

De data wordt op dit moment weggeschreven naar een csv bestand genaamd “cottages.csv”, elke keer dat er gescrapet wordt het csv bestand geleegd en opnieuw ingevuld met bungalows.

	A	B	C	D	E	F
1	title	price	amount of persons	bedroom	duration	surface
2	Premium cottage Vernieuwd	€ 289	4 pers.	2 Slaapkamers	Van ma. 26 tot wo. 28 jun.	64m²
3	Premium cottage Rolstoelvriendelijk Vernieuwd	€ 619	4 pers.	2 Slaapkamers	Van di. 11 tot do. 13 jul.	73m²
4	Comfort cottage Vernieuwd	€ 249	4 pers.	2 Slaapkamers	Van wo. 28 tot vr. 30 jun.	62m²
5	Premium Appartement Vernieuwd	€ 189	2 pers.	1 Slaapkamer	Van di. 27 tot do. 29 jun.	32m²
6	Premium Appartement Vernieuwd	€ 239	3 pers.	2 Slaapkamers	Van wo. 28 tot vr. 30 jun.	39m²
7	VIP cottage Vernieuwd	€ 289	2 pers.	1 Slaapkamer	Van wo. 28 tot vr. 30 jun.	64m²
8	Dieren in het bos cottage	€ 369	4 pers.	2 Slaapkamers	Van wo. 28 tot vr. 30 jun.	64m²
9	VIP cottage Vernieuwd	€ 379	4 pers.	2 Slaapkamers	Van ma. 12 tot wo. 14 jun.	66m²
10	Premium cottage Vernieuwd	€ 339	5 pers.	3 Slaapkamers	Van di. 27 tot do. 29 jun.	64m²
11	Comfort cottage Vernieuwd	€ 279	5 pers.	3 Slaapkamers	Van ma. 26 tot wo. 28 jun.	64m²
12	Premium cottage Vernieuwd	€ 459	8 pers.	4 Slaapkamers	Van ma. 26 tot wo. 28 jun.	103m²
13	VIP cottage Vernieuwd	€ 499	6 pers.	3 Slaapkamers	Van di. 13 tot do. 15 jun.	103m²
14	Pony cottage Vernieuwd	€ 629	5 pers.	3 Slaapkamers	Van vr. 9 tot ma. 12 jun.	64m²

Webscraper voor alle vakantieparken van Center Parcs

Een webscraper voor 1 parc was een goed begin, maar zeker geen eindproduct. Ook werkte de vorige scraper aan de hand van een vaste URL en ik wil dat de user kan kiezen wel vakantie parc hij gaat scrapen.

In de tweede iteratie heb ik er dan voor gezorgd dat je kan kiezen van welke vakantiepark je de bungalows wil scrapen. Dit doe ik door 3 user inputs te vragen: country, vacation parc en vacation parc code. Deze 3 variabelen vult her programma dan in de url en zo weet de webscraper welke pagina deze moet scrapen. Deze parknamen en parkcodes heb ik weggeschreven naar een excel bestand zodat het de gebruiker deze niet elke keer moet gaan opzoeken.

	A	B	C	D	E	F	G	H	I	J	K
1	**belgië**			**nederland**			**frankrijk**			**duitsland**	
2	**Park Naam**	**Park Code**		**Park Naam**	**Park Code**		**Park Naam**	**Park Code**		**Park Naam**	**Park Code**
3	terhills	TH		de-kempervennen	KV		le-lac-d-ailette	LA		park-allgau	AG
4	erperheide	EP		het-heijderbos	HB		les-trois-forets	TF		bispinger-heide	BS
5	park-de-haan	HA		port-zelande	PZ		villages-nature-paris	VN		park-bostalsee	BT
6	les-ardennes	AR		de-eemhof	EH		les-landes-de-gascogne	LG		park-hochsauerland	SL
7	de-vossemeren	VM		park-zandvoort	ZV		les-bois-francs	BF		park-nordseekuste	BK
8				het-meerdal	MD		les-hauts-de-bruyeres	CH		park-eifel	HE
9				limburgse-peel	LH		le-bois-aux-daims	BD			
10				parc-sandur	SR						
11				de-huttenheugte	HH						

De webscraper vraagt ook in het begin of de user de bungalows wilt filteren op een bepaald aantal personen (volwassenen, huisdieren, ouderen, kinderen) en vertrek en aankomstdatum. Deze waarde worden ook opgeslagen in een variabele en geïmplementeerd in de url waarop de webscraper gaat scrapen. De specificaties die van de bungalows worden gescrapet zijn:

- Titel
- Nieuwe prijs
- Aantal personen
- Oppervlakte
- Datum
- Aantal slaapkamers
- Land

- Vakantie parc

De data wordt nu ook weggeschreven naar een PostGreSQl database, ik heb ook toegevoegd dat de webscraper een melding geeft wanneer alles gescrapet

title character varying (255)	new_price character varying (255)	amount_of_persons character varying (255)	bedroom character varying (255)	duration character varying (255)	surface character varying (255)	country character varying (255)	vacation_parc character varying (255)
VIP cottage Rolstoelvriendelijk	259	4 pers.	2 Slaapkamers	Van ma. 15 tot wo. 17 mei.	84 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage	259	4 pers.	2 Slaapkamers	Van ma. 15 tot wo. 17 mei.	77 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage	339	2 pers.	1 Slaapkamer	Van ma. 15 tot wo. 17 mei.	58 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive Lakeside cottage	409	4 pers.	2 Slaapkamers	Van ma. 15 tot wo. 17 mei.	77 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive cottage	599	4 pers.	2 Slaapkamers	Van vr. 14 tot ma. 17 apr.	77 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive cottage	789	2 pers.	1 Slaapkamer	Van di. 11 tot vr. 14 apr.	58 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP Lakeside cottage	344	4 pers.	2 Slaapkamers	Van di. 12 tot do. 14 dec.	77 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage Rolstoelvriendelijk	289	6 pers.	3 Slaapkamers	Van ma. 15 tot wo. 17 mei.	105 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage	289	6 pers.	3 Slaapkamers	Van ma. 15 tot wo. 17 mei.	98 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP Lakeside cottage	369	6 pers.	3 Slaapkamers	Van ma. 15 tot wo. 17 mei.	98 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage	409	8 pers.	4 Slaapkamers	Van ma. 15 tot wo. 17 mei.	125 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive Lakeside cottage	429	6 pers.	3 Slaapkamers	Van ma. 15 tot wo. 17 mei.	98 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP Lakeside cottage	509	8 pers.	4 Slaapkamers	Van ma. 15 tot wo. 17 mei.	125 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive Lakeside cottage	589	8 pers.	4 Slaapkamers	Van ma. 15 tot wo. 17 mei.	125 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
VIP cottage	719	12 pers.	6 Slaapkamers	Van ma. 15 tot wo. 17 mei.	182 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive cottage	719	6 pers.	3 Slaapkamers	Van vr. 12 tot zo. 14 mei.	98 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive Lakeside cottage	769	12 pers.	6 Slaapkamers	Van ma. 15 tot wo. 17 mei.	182 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Exclusive cottage	1219	12 pers.	6 Slaapkamers	Van di. 18 tot vr. 21 apr.	182 m²	België, Limburg, Dilsen-Stokkem	Terhills Resort by Center Parcs
Comfort cottage	€159	4 pers.	2 Slaapkamers	Van ma. 15 tot wo. 17 mei.	59 m²	België, Ardennen, Vielsalm	Les Ardennes
Comfort cottage	€129	4 pers.	2 Slaapkamers	Van ma. 15 tot wo. 17 mei.	50 m²	België, Ardennen, Vielsalm	Les Ardennes
Comfort cottage	€159	6 pers.	3 Slaapkamers	Van ma. 15 tot wo. 17 mei.	75 m²	België, Ardennen, Vielsalm	Les Ardennes

Webscraper voor alle vakanties/aanbiedingen van Center Parcs

Center Parcs biedt ook veel kortingen op bungalows aan de hand van vakanties en aanbiedingen. Hier was Molenheide ook zeer geïnteresseerd in dus heb ik dit ook toegevoegd in de webscraper. Center Parcs biedt veel verschillende aanbiedingen aan zoals last-minute voordeel, vroegboekvoordeel, ecocheques… en veel vakantie aanbiedingen zoals paasvakantie, zomervakantie, herfstvakantie…

Dit heb ik gedaan door aan het begin van het programma te vragen aan de user wat hij wilt gaan scrapen (een specifiek vakantiepark, een specifiek land of vakanties/aanbiedingen). Als de user doorgeeft dat hij vakanties/aanbiedingen wilt gaan scrapen vraagt het programma welke vakantie of aanbiedingen en dit wordt opgeslagen in een variabele die vervolgens wordt ingevuld in een url. De webscraper gaat vervolgens alle mogelijke bungalows scrapen en deze data wegschrijven naar een csv bestand met de naam van desbetreffende vakantie/aanbieding en naar een PostGreSQL database met de naam offer/vacation.

De specificaties die van de bungalows worden gescrapet bij vakanties en offers zijn:

- Titel
- Nieuwe prijs
- Oude prijs
- Aantal personen
- Oppervlakte
- Datum
- Aantal slaapkamers
- Land
- Vakantie parc

- Timestamp
- Korting in euro’s
- Korting in procent

Ik heb nu ook een timestamp toegevoegd, zo kunnen ze makkelijk nakijken wanneer een bungalow gescrapet is.

Om het Molenheide ook makkelijker te maken heb ik nu ook de oude prijs van de bungalows gescrapet, zo kon ik ook in het programma zelf de korting van de bungalows in euro’s en procent berekenen aan de hand van de oude prijs en de nieuwe prijs van de bungalows.

Webscraper voor alle bungalows van Center Parcs

Dit programma is geschreven in Python en maakt gebruik van de Selenium-webdriver en andere pakketten om gegevens van de Center Parcs-website te extraheren en op te slaan in een PostgreSQL-database en CSV-bestanden. Het programma heeft verschillende functies:

1. **Get_data_all(url, program_cat):** Deze functie ontvangt een URL en een programmabeschrijving als invoer. Het opent een Chrome-webdriver en navigeert naar de opgegeven URL. Vervolgens klikt het op een knop om alle beschikbare bungalowss weer te geven en wacht tot de knop niet meer zichtbaar is. Daarna worden de gegevens van elke bungalow op de pagina verzameld en opgeslagen in een lijst van dictionaries. Deze gegevens omvatten de titel, prijzen, aantal personen, aantal slaapkamers, duur, oppervlakte, land, vakantiepark, tijdstempel en kortingen. Vervolgens maakt de functie verbinding met een PostgreSQL-database en voegt de bungalowgegevens toe aan de database. Ten slotte retourneert de functie de verzamelde gegevens. Deze functie wordt opgeroepen als de user kiest om vakanties/aanbiedingen te scrapen

2. **Get_data(url):** Deze functie ontvangt een URL als invoer. Het opent een Chrome-webdriver en navigeert naar de opgegeven URL. Daarna worden de gegevens van elke bungalow op de pagina verzameld en opgeslagen in een lijst van dictionaries. Deze gegevens omvatten de titel, prijzen, aantal personen, aantal slaapkamers, duur, oppervlakte, land, vakantiepark, tijdstempel en kortingen. Vervolgens maakt de functie verbinding met een PostgreSQL-database en voegt de bungalowgegevens toe aan de database. Ten slotte retourneert de functie de verzamelde gegevens. Deze functie wordt opgeroepen als de user kiest om een specifiek vakantiepark of land te scrapen.

3. **main():** Dit is de hoofdfunctie van het programma. Het vraagt de gebruiker om invoer voor het aantal keren dat het programma moet worden uitgevoerd en het type programma (resort, country, offer/vacation). Afhankelijk van het gekozen programma, worden er verdere vragen gesteld om de benodigde informatie te verzamelen.

Vervolgens wordt de juiste URL gegenereerd en wordt de corresponderende functie (get_data_all() of get_data()) aangeroepen om de gegevens van de Center Parcs-website te extraheren. De verzamelde gegevens worden opgeslagen in een CSV-bestand en een bericht wordt afgedrukt met het aantal geëxtraheerde bungalows en de locatie van het CSV-bestand.

Het programma maakt ook gebruik van de psycopg2-bibliotheek om verbinding te maken met een PostgreSQL-database en de gegevens daarin op te slaan. De databaseverbinding wordt tot stand gebracht met behulp van de gegevens voor host, poort, database, gebruikersnaam en wachtwoord die in het programma zijn opgegeven.

Let op: de programmabeschrijvingen en de bijbehorende URL's zijn in dit programma niet volledig ingevuld en moeten door de gebruiker worden aangevuld om het programma correct te laten werken.

Webscraper voor alle bungalows van Belvilla

Bij het vergelijken van al de verschillende concurrenten van Molenheide leek Belvilla ook een geschikte kandidaat voor het scrapen van bungalows. Het grote verschil tussen Molenheide en Belvilla is dat Belvilla geen vakakantiepark is dat bungalows verhuurd. De bungalows die Belvilla aanbiedt zijn meestal huizen van particulieren die hun huis enige tijd niet nodig hebben.

Ik kan niet dezelfde webscraper gebruiken voor Belvilla als voor Center Parcs omdat de webscraper voor Center Parcs specifiek gebouwd is voor hun website, maar ik kon hem wel op dezelfde manier opbouwen. Belvilla heeft wel veel minder aanbiedingen/vakanties op hun website in vergelijking met Center Parcs maar veel meer bungalows ter beschikking voor België en Nederland. Belvilla heeft ook veel minder specificaties bij hun bungalows staan, bijvoorbeeld ze hebben geen vertrek en aankomstdatum wat Center Parcs en Molenheide wel hebben.

Wegschrijven data van webscrapers

Ik heb eerst opgezocht hoe ik de data die ik terugkrijg van de webscraper wegschrijf naar een csv bestand, dit waren maar enkele lijnen code die ik moest toevoegen. Ik had ook aangepast dat het csv bestand een naam krijgt aan de hand van welk vakantiepark hij scrapet om ervoor te zorgen dat hij niet alles in 1 csv bestand wegschrijft.

Ik heb de webscraper de data ook naar een PostgreSQL database laten wegschrijven, ik had nagevraagd aan de mensen in mijn tribe op stage welke database er het meest gebruikt wordt en dat was PostgreSQL. In PostGreSQL heb ik ook verschillende tabellen gemaakt (vakantieparken, aanbiedingen, vakanties) zodat de data meer gestructureerd wordt weggeschreven.

Flask applicatie voor webscrapers

Het laatste deel van mijn stageopdracht was het creëren van een aantrekkelijke module voor de webscrapers. Hier heb ik wat opzoekwerk naar gedaan om te kijken welk framework hier het meest geschikt voor is en uiteindelijk heb ik gekozen om flask te gebruiken.

Dit programma is een HTML-formulier dat dynamisch wordt aangepast op basis van de geselecteerde waarde in het dropdown-menu voor het "Programma". Het programma maakt gebruik van JavaScript om de formulierelementen te tonen of te verbergen op basis van de geselecteerde optie.

Wanneer de pagina wordt geladen, wordt de JavaScript-functie handleProgramChange() gedefinieerd. Deze functie wordt aangeroepen wanneer de waarde van het "Programma" dropdown-menu verandert.

De functie haalt de geselecteerde waarde op van het "Programma" dropdown-menu en wijst deze toe aan het variabele program. Vervolgens worden de formulierelementen geselecteerd door hun id's en toegewezen aan de overeenkomstige variabelen, zoals programDate, arr_date, ret_date, country, vacationParkCode, vacationPark, children, en childrenAges.

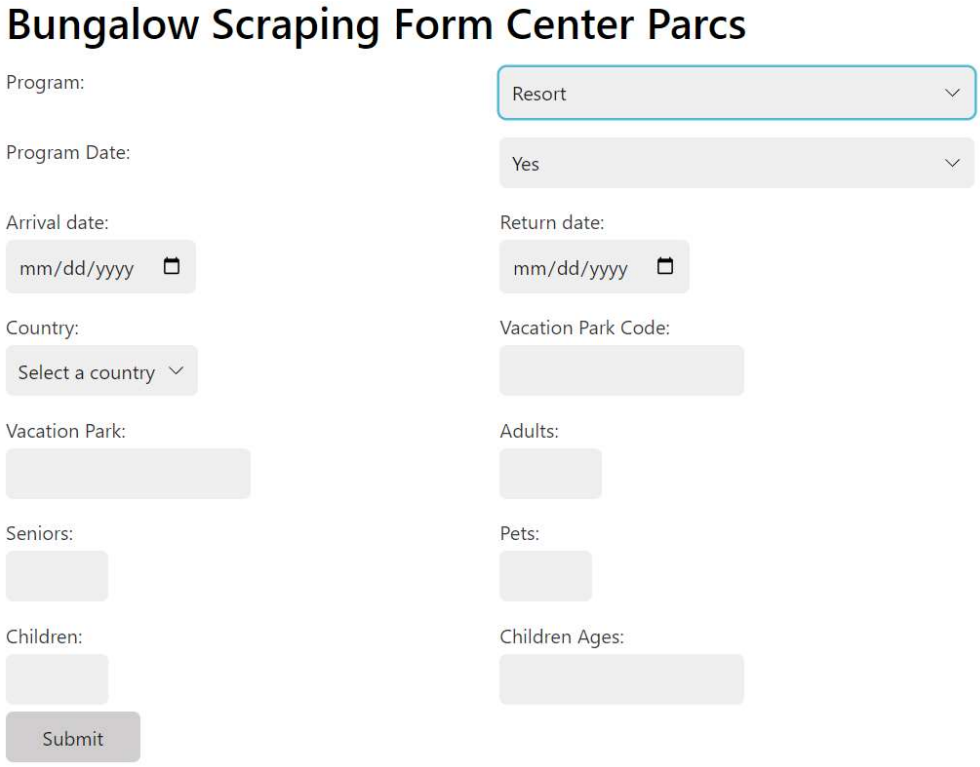
Als het geselecteerde programma gelijk is aan "resort", worden de formulierelementen voor de resorteigenschappen getoond door de style. display eigenschap op "block" in te stellen. Als het programma "resort" is en de geselecteerde waarde voor "Programma Datum" gelijk is aan "yes", worden ook de aankomst- en vertrekdatumvelden getoond. Als de geselecteerde waarde voor "Programma Datum" "no" is, worden deze velden verborgen door de style.display eigenschap op "none" in te stellen.

Als het geselecteerde programma niet gelijk is aan "resort", worden de formulierelementen voor de resorteigenschappen verborgen door de style.display eigenschap op "none" in te stellen.

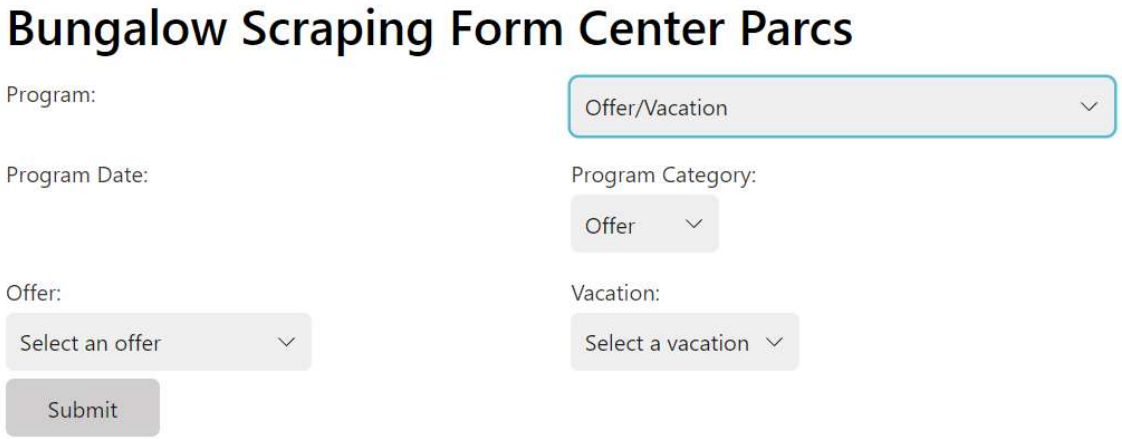
Het formulier zelf heeft de attributen onchange="handleProgramChange()" op het "Programma" dropdown-menu en onsubmit="handleProgramChange()" op het formulier. Dit zorgt ervoor dat de functie handleProgramChange() wordt aangeroepen wanneer de geselecteerde waarde verandert of wanneer het formulier wordt ingediend. Hierdoor wordt het formulier dynamisch aangepast op basis van de geselecteerde optie voordat het wordt ingediend naar de server.

Dit programma maakt gebruik van HTML, CSS en JavaScript om een interactief formulier te creëren dat zich aanpast aan de geselecteerde waarde van het "Programma" dropdown-menu.

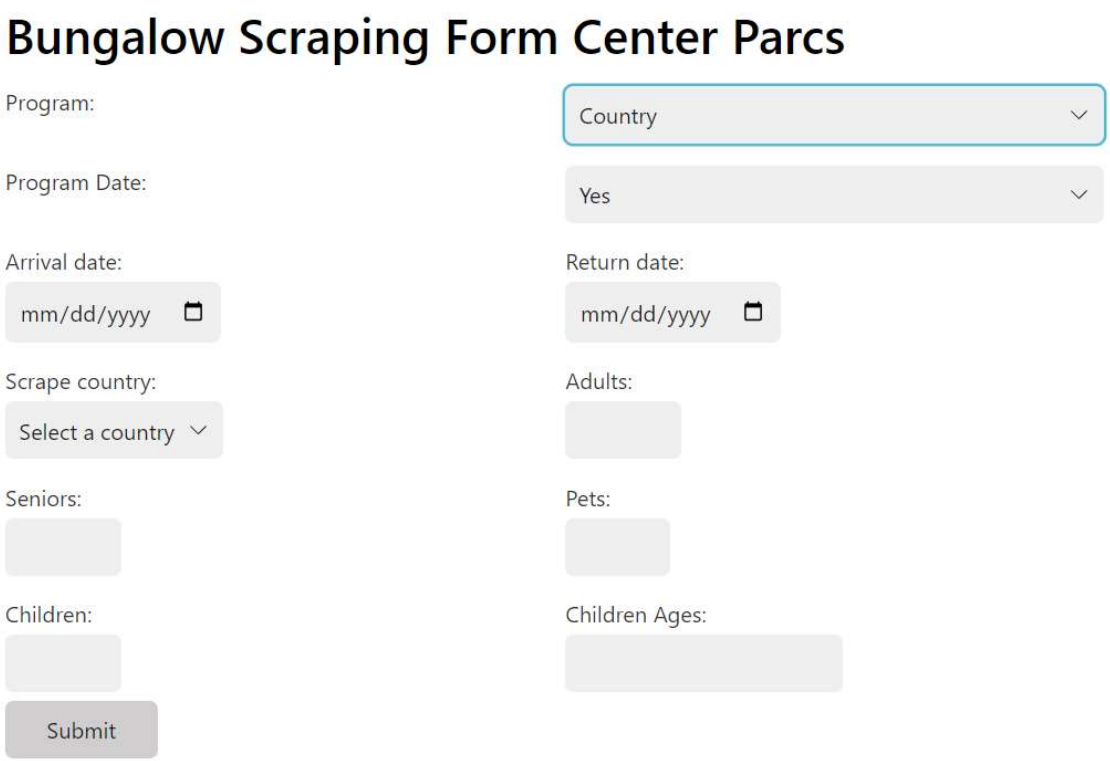
Frontend voor een speciefiek vakantie parc



Frontend voor vakanties/aanbiedingen



Frontend voor een specifiek land



Ondervonden problemen en oplossingen

Verbindingssnelheid internet

Bij het testen van de webscraper was het belangrijk dat ik een snelle internetverbinding had omdat de webscraper voordat hij alle bungalows gaat scrapen moet blijven klikken op de "toon meer knop", hier had ik een time.sleep van 4 seconden op ingesteld maar omdat het internet soms niet al te snel was op stage kon hij de "toon meer knop" niet op tijd inladen en dacht de webscraper dat alle bungalows al getoond waren op de pagina. Om dit probleem op te lossen heb ik de time.sleep iets groter gezet voor de zekerheid, dit vertraagt wel het hele scrape proces maar zo heb ik dat probleem wel nooit meer voorgehad.

Communicatie met klant (Molenheide)

In het begin van de stage had ik weinig tot geen contact met de klant maar dat was toen geen probleem omdat ik wist wat ik moest maken. Nadat ik de webscraper voor Center Parcs helemaal afhad heeft mijn stagementor de data van last-minutes en vroegboekvoordeel aanbiedingen doorgestuurd in csv bestanden naar Molenheide, hier waren ze erg gelukkig mee.

Conclusie

Tijdens mijn stage bij Wisemen heb ik mijn stageopdracht kunnen voltooien en heb ik veel bijgeleerd. Naast het merendeel zelfstandig programmeren van de

webscraper heb ik ook achter de schermen kunnen meekijken hoe websites ontwikkeld worden en de processen die er allemaal voor nodig zijn.

De klant (Molenheide) was ook zeer tevreden met het resultaat dat ik nu al heb kunnen tonen, de webscraper is zeker nog niet helemaal klaar maar nu hebben ze wel al een start met wat ik al gemaakt heb.