Webscraper

Project plan

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

Project plan: Wouter Vandueren - Wisemen – Molenheide

Table of Contents

Who

Wisemen is een digitaal marketingbureau dat bedrijven helpt groeien en slagen in de digitale wereld. Het bedrijf is gevestigd in Diepenbeek, België, en biedt een scala aan digitale marketingdiensten, waaronder SEO (Search Engine Optimization), PPC-advertenties (Pay-Per-Click), marketing via sociale media, webdesign en -ontwikkeling, branding en contentmarketing.

Tot voor kort heette het bedrijf altijd Appwise, omdat er vroeger meer gefocust werd op het enkel ontwikkelen van apps. Sinds maart is de naam veranderd naar Wisemen, dit omdat het bedrijf vandaag geëvolueerd is naar een digitale totaalpartner

Wat het bedrijf Wisemen uniek maakt is dat ze een volledig team van ervaren digitale marketingexperts hebben die met klanten samenwerken om op maat gemaakte strategieën te creëren die aansluiten bij hun zakelijke doelen en doelstellingen. Dit team zorgt ook voor interessante informatiefilmpjes op sociale media die op deze manier potentiële klanten lokken. Hun focus ligt op het leveren van meetbare resultaten en het helpen van bedrijven om de gewenste resultaten te bereiken, of het nu gaat om het genereren van leads, het vergroten van het websiteverkeer of het verbeteren van de naamsbekendheid.

Net zoals de meeste it-bedrijven heeft Wisemen ook een development team, hierin worden websites, apps etc. ontwikkeld. In Wisemen heb je verschillende 'tribes' binen deze service, elke tribe bestaat uit front- end en back-end developers en in de meeste tribes is er ook een ux/ui designer aanwezig. Elke tribe is toegewijd aan het ontwikkelen van website/webapplicaties of het ontwikkelen van apps voor op de telefoon.

Als laatste service heeft wisemen ook een strategie team, hierin wordt de beste strategie bepaald om een bedrijf te laten evolueren naar een meer digitaal bedrijf. Wisemen hanteert een gegevensgestuurde benadering van digitale marketing en gebruikt analyses en inzichten om hun strategieën en campagnes te onderbouwen. Ze leggen ook sterk de nadruk op creativiteit en innovatie en werken nauw samen met klanten om unieke oplossingen te ontwikkelen waarmee ze zich onderscheiden van de concurrentie.

Over het algemeen is Wisemen een full-service digitaal marketingbureau dat bedrijven helpt een sterke online aanwezigheid op te bouwen en hun

marketingdoelen te bereiken door middel van een reeks digitale marketingdiensten.

What

Molenheide / Saraland is een vakantiepark gevestigd in Houthalen Helchteren en Hotton, een uniek totaalconcept in de Benelux. De voorbije 20 jaar heeft Molenheide geïnvesteerd in eigen boekingssoftware. Dit resulteert in een ongeziene applicatie waar zelfs gespecialiseerde aanbieders van bungalowsoftware niet aan kunnen tippen.De applicatie is echter niet future proof. Enerzijds door de sterk verouderde technologie, anderzijds omdat alle (technische) kennis bij één persoon gebundeld is, die de pensioengerechtigde leeftijd binnen nadert. Een nieuwe applicatie drong zich op en daar heeft Wisemen het voorbije zijn schouders onder gezet.

Voor mijn stageopdracht moet ik ervoor zorgen dat Molenheide de prijzen kan raadplegen van bungalows van hun concurrenten. Ik moet beginnen van één datapunt (Center Parcs) maar dit kan nog evolueren naar meerdere datapunten. Om deze prijzen van de websites te kunnen afhalen ga ik een webscraper moeten schrijven. Ik mag zelf kiezen in welke programmeertaal ik de webscraper schrijf, alleen voor de database moet ik gebruik maken van postgresql. Eerst ga ik een webscraper schrijven die vanuit de frontend scrapet, later kan ik kijken of dit niet beter gaat via API calls.

Ik moet er ook voor zorgen dat de webscraper makkelijk gebruikt kan worden en dat er een timer opgezet kan worden zodat deze webscraper bijvoorbeeld om de week bungalows gaat scrapen. De klant kan dan met queries de gegevens uit de database halen om hun prijzen ermee te vergelijken. De webscraper moet ook gepresenteerd worden in een aantrekkelijke module. Uiteindelijk kan ik ook nog een front end implementeren zodat ze via daar de data uit de database kunnen ophalen.

Why

Huidig business model klant

Park Molenheide is een vakantiepark in Midden Limburg met 350 woningen, een groot overdekt zwemparadijs (inclusief 4 baden en 4 glijbanen), een overdekte binnenspeeltuin (grootste van de Benelux), verschillende horeca aangelegenheden en een award winning indoor mini-golfbaan (grootste van Europa). Jaarlijks trekken ze ongeveer 200.000 bezoekers aan. Het unieke aan Park Molenheide is dat elke bungalow vrijstaand is, midden in de natuur ligt en grenst aan een wild- en wandelpark.

Hun kernactiviteit omvat het korte verblijfstoerisme in vakantiecentra (bungalows, chalets en campingplaatsen), voornamelijk rechtstreeks voor particuliere bezoekers. Het eendagstoerisme omvat daarnaast ongeveer 30% van hun omzet.

Hun doelgroep bestaat voornamelijk uit jonge ouders met kinderen. Een persoonlijke opvang en familiale, kindvriendelijke service staat centraal in onze werking.

Binnen onze kernprocessen valt het plannen van onze diensten, om ervoor te zorgen dat elke bungalow tijdig instapklaar is wanneer er bezoekers verwacht worden. Dit proces is een lange ketting van verschillende stappen en actoren. Een groot deel van deze actoren staat op de loonlijst van Park Molenheide, ter ondersteuning van deze bemanning werken we met interimcontracten op momenten dat er tijdelijk een verhoogde werkdruk verwacht wordt.

Park Molenheide is onderdeel van de Molenheide Groep, waaronder ook verblijfspark Moulin d'Hotton is opgenomen sinds 2017. Dit park omvat 30 woningen, op het moment van schrijven worden er 20 extra bijgebouwd. Een derde verblijfspark van 19 woningen in Rochefort, is momenteel in de maak. In tegenstelling tot Park Molenheide, bieden deze twee verblijfsparken geen extra faciliteiten zoals horeca of zwembaden en bestaat het personeel er voornamelijk uit een lokale poetsdienst. Het verdere beheer van deze verblijfsparken wordt uitgevoerd vanuit de centrale locatie van Park Molenheide.

Evolutie business model klant

Momenteel wordt een digitaal systeem uitgewerkt dat centraal zal staan in de werking van Park Molenheide. Dit omvat een performant en gebruiksvriendelijk boekingssysteem dat **in de toekomst** gekoppeld kan worden met externe websites, ticketprinters, toegangscontrole,

mailservices, …, alles raadpleegbaar in de cloud door stakeholders met verschillende toegangsniveaus.
De transformatie die we beogen in deze kanteling gaat veel verder. We willen de grote hoeveelheid datapunten binnen en buiten ons digitale ecosysteem verbinden, analyseren en aggregeren om onze operationele processen efficiënter en schaalbaar te maken en onze economische processen dynamisch en slimmer te maken. Een voorbeeld is de user journey van het aanmaken van een boeking tot en met het reinigen van de woning voor en na gebruik. Dit is een keten van verschillende processen en actoren, waarin elke kleine fout een kettingreactie kan initiëren.
Om een sterke datagedreven onderneming te worden, ambiëren zij een zelflerend model op basis van artificiële intelligentie, dat met verschillende interne en externe parameters (bv. historische data, buitenlandse economische ontwikkelingen, vakantiedata, gebruik van de faciliteiten, weermodellen, online prijzen van concurrenten, …) voorspellingen zal maken en inzichten zal bieden op vlak van:

1. prijszetting: Deze prognose stelt Molenheide in staat om de prijs voor een verblijf in een vakantiepark en de bijhorende activiteiten dynamisch en automatisch in te stellen, enerzijds competitief naar concurrenten toe en anderzijds geoptimaliseerd naar eigen marges toe (de juiste prijs op het juiste moment gericht op de juiste doelgroep).
2. bezettingsgraad bungalows: Een prognose op basis van historische data, actuele en toekomstige data over verlofperiodes in verschillende lokale en internationale regio's, weersvoorspellingen, … die ons in staat stelt om de bezettingsgraad te voorspellen op korte en langere termijn.
3. doelgerichte targeting: Data en prognoses over de bezettingsgraad laten ons toe om automatische doelgerichte digitale marketing via verschillende kanalen op de juiste momenten in de juiste regio's uit te kunnen voeren om laag voorspelde bezettingsgraden te verhogen.
4. werkplanning korte en lange termijn: Op basis van punten 1, 2 en 3 streven we een verhoogde bezettingsgraad na met een verhoogde continuïteit. Een lager aantal pieken en dalen in de bezettingsgraad, samen met de mogelijkheid om deze langer op voorhand te voorspellen (en bij te stellen op basis van nieuwe data) biedt Park Molenheide de mogelijkheid om werktijden van de interne diensten en processen (bv. poetsdienst, horeca, uitbaten bowling, …) slimmer in te schatten en in te plannen
5. energieverbruik: Door stijgende energieprijzen worden onze kosten drastisch verhoogd. Met een intelligent gebouw- beheer willen we energieprijzen en verbruikdata opnemen als parameters binnen het datagedreven ecosysteem. In periodes met hoge energieprijzen kan het netto voordeliger zijn om tijdelijk de bezettingsgraad te verlagen. Daarnaast voorzien we een slimme, dynamische sturing om het energieverbruik te optimaliseren en minimaliseren (bv. slimme sturing

van de verwarming van bungalows op basis van geolocatie van de bezoekers, patronen in historische gebruikersdata, …).

We voorzien een gefaseerde aanpak: (1) future-proof data- verwerking en analyse als basis voor de use cases (2) POC voor datagedreven prijszetting en marketing en POC voor werkplanning optimalisatie voor 1 specifieke dienst. Op basis van deze onderzoeken zal het aantal databronnen en de configuratie graad van de artificiële intelligentie algoritmes verder uitgebreid en geautomatiseerd worden, zullen de andere interne processen slimmer gemaakt worden en zal het onderzoek starten om energieverbruik te optimaliseren.

When

In het begin van mijn stage heb ik een introductie gehad over de opdracht van molenheide. Dit is een project waar ze al een jaar aan bezig zijn en dit project gaat ook langer duren als mijn stage is. Ik moet maar een klein deel maken van dit grote project, zoals in het vorig puntje beschreven moet ik enkel ervoor zorgen dat de klant Molenheide de prijzen van hun concurrentie kan raadplegen om zo hun prijzen iets competitiever te maken. Ik sta ook als enige van de tribe waar ik inzit op dit deel van de opdracht dus ik ga veel zelfstandig moeten werken.

INFO OPZOEKEN (27/2 – 3/3)

Mijn eerste opdracht gaat zijn het maken van een webscraper die prijzen en andere specificaties van bungalows van verschillende vakantieparken haalt. Ik ga beginnen bij center parcs omdat deze een van de grote concurrenties van molenheide en deze site heeft veel type bungalows. Ik ga beginnen met het onderzoeken wat de gelijkenissen zijn tussen hoe Molenheide de bungalows aanbied en hoe Center Parcs bungalows aanbied. Want ik wil niet alleen de prijzen gaan scrapen maar ook bijvoorbeeld hoeveel personen er kunnen verblijven en de periode zodat je makkelijker de bungalows tussen Molenheide en Center Parcs kunt vergelijken. Ik ga ook opzoeken met welke tools ik het beste deze webscraper kan maken, ik zou het liefst de webscraper in python programmeren maar dan heb ik nog de keuze of ik selenium, scrapy of beautifoulsoup ga gebruiken.

STAP 1 (6/3 – 10/3)

Ik ga beginnen met het maken van een scraper die specifiek van een vakantiepark alle bungalows gaat scrapen, deze data wordt dan weggeschreven naar een hardcoded csv file. Wat de webscraper van bungalows zal scrapen zijn de titel, prijs, aantal personen dat er kan

logeren, aantal slaapkamers, datum van verblijf en aankomst, oppervlakte van de bungalow, land waarin de bungalow ligt en het vakantiepark waarin de bungalow ligt. Ik heb ook besloten dat ik de webscraper ga schrijven in python met behulp van selenium. Ik heb gekozen voor selenium omdat ik hier al eens een webscraper mee gemaakt heb en voor python omdat ik deze syntax makkelijker vind om mee te werken.

STAP 2 (13/3 – 17/3)

Na dit te hebben geprogrammeerd ga ik de webscraper updaten door deze te laten werken voor elk vakantiepark in elk land. Dit ga ik doen door een paar user inputs te vragen als het programma gerund word, aan de hand hiervan kan ik de juiste url meegeven om de bungalows te scrapen. De eerste user input die ik vraag is het land waarin het vakantiepark ligt, de tweede user input is de naam van het vakantiepark en de laatste user input is de code van het vakantiepark (dit is een 2 letterige afkorting van de naam van het vakantiepark). Deze user inputs worden dan ingevuld in de url, de webscraper gaat dan op deze url de bungalows scrapen en gaat deze wegschrijven naar een csv bestand dat begint met de naam van het opgegeven vakantiepark gevolgd door "_cottages.csv", deze csv wordt opgeslagen in een folder met de naam van het opgegeven land. Ik heb voor de gemakkelijkheid ook alle landen en vakantieparken codes opgeslagen in een csv bestand.

STAP 3 (20/3 – 22/3)

Hierna ga ik de webscraper aanpassen zodat de data niet alleen naar een csv wordt weggeschreven, maar ook naar een database. Ik ga er eerst voor zorgen dat de data wordt weggeschreven naar een SQLite database omdat ik al ervaring heb met dit type database, veel zal ik niet moeten aanpassen in het programma om dit mogelijk te maken. Ik ga er gewoon voor moeten zorgen dat er een connectie is tussen mijn programma en de database en dat de data in de juiste kolommen wordt weggeschreven.

STAP 4 (23/3 – 24/3)

Na navraag aan de mensen binnen mijn tribe heb ik gehoord dat ze liever met PostgreSQL werken dan met SQLite dus hiervoor ga ik mijn code nog aanpassen. Ik ga ook een kleine security implementeren in mijn code dat wanneer er geen verbinding met de database gemaakt kan worden het programma een error meegeeft. Om mijn programma te laten werken met PostgreSQL zal ik wat lijnen moeten aanpassen en toevoegen, zoals de connectie met de database. Voor het moment zal ik de database nog lokaal draaien maar ik heb altijd de mogelijkheid om deze online te zetten.

STAP 5 (27/3 – 31/3)

Als al de vorige stappen zijn gelukt zal de webscraper al werken voor alle specifieke vakantieparken in elk land van Center Parcs. Het enige probleem is dat deze webscraper per keer maar van één vakantiepark de bungalows scrapet. Er zullen ook mogelijkheden moeten zijn dat je kan kiezen of je specifieke aanbiedingen/vakanties gaat scrapen. Hier ga ik waarschijnlijk op 2 grote problemen stuiten, **probleem1**: de webscraper die ik al geschreven heb zal ik hiervoor moeten aanpassen, dus ik ga eigenlijk een nieuwe webscraper moeten maken. Dit kan ik wel oplossen door wanneer het programma gerund worden te vragen aan de user of hij alle bungalows wilt van alle vakantieparken of specifiek van 1 vakantiepark en aan de hand hiervan kan ik zeggen tegen het programma welke functie hij moet uitvoeren. **Probleem 2**: het aantal bungalows die het programma in 1 keer zal moeten scrapen is aanzienlijk groter, bij het specifiek scrapen zat het aantal bungalows tussen de 5-30 bungalows en bij het alles scrapen zal het rond de 400 bungalows zitten. Dit zal het runproces volgens mij enorm vertragen

STAP 6 (3/4 – 7/4)

Nadat de webscraper ook werkt voor specifieke aanbiedingen/vakanties te scrapen ga ik extra gegevens toevoegen die de webscraper moet gaan scrapen. Voordien werd alleen de nieuwe prijs gescrapet maar nu ga ik ook de oude prijs scrapen, zo kan ik ook in de webscraper zelf de berekening gaan maken wat het het verschil is in euro's en procent tussen de nieuwe prijs en de oude prijs. Ik ga ook nog een timestamp meegeven zodat he precies kan zien wanneer deze gegevens zijn gescrapet. Ik ga de user ook een keuze laten wanneer deze specifiek een vakantiepark wilt scrapen dat hij ook kan instellen voor hoeveel personen en tussen welke 2 datums. Voor het aantal personen kan je kiezen uit adults, children en seniors. Je kan ook nog pets toevoegen.

STAP 7(10/4 – 14/4)

De webscraper voor alle bungalows van Center Parcs te scrapen is zo goed als klaar, deze week ga ik de webscraper de hele tijd testen om te kijken of er nog ergens fouten in de code zitten. Ik ga de code ook zo goed mogelijk proberen te optimaliseren zodat de webscraper zo snel mogelijk werkt.

STAP 8(17/4 – 28/4)

Nadat de webscraper van Center Parcs goed getest is en deze volledig werkt kan ik beginnen aan een webscraper voor een andere concurrent van Molenheide. Er zal wel eerst nagevraagd moeten worden van welke

concurrenten Molenheide de prijzen het liefst gescrapet krijgt. Indien ze meerdere concurrenten voorstellen zal ik moeten kijken welke concurrent de meeste bungalows/chalets aanbied in de landen België/Nederland en welke website het meest up to date is. Ik ga deze websites vergelijken aan de hand van een excel bestand waar ik de verschillende eisen waar deze sites aan moeten voldoen met elkaar ga vergelijken.

STAP 9(1/5 – 10/5)

Na te hebben gekeken welke concurrent het meest geschikt is om een webscraper op te bouwen ga ik dezelfde stappen ondernemen zoals ik bij de vorige webscraper heb gedaan. Ik ga weer identiek een webscraper maken met python en selenium maar omdat ik via de frontend manier scrape ga ik een geheel nieuwe webscraper moeten maken. Het volgende datapunt da tik ga nemen is Belvilla, omdat deze website het meest geschikt is voor mijn type webscraper en omdat Belvilla veel bungalows/chalets aanbied in België en Nederland.

STAP 10(11/5 – 27/5)

Nu de webscrapers voor Center Parcs en Molenheide klaar zijn ga ik me bezig houden met het realiseren van een frontend webapplicatie voor de beide webscrapers. Zo kan de user de webscrapers makkelijker gebruiken en zo kan ik ook beter tonen bij de bachelorproef presentatie tonen wat ik gemaakt heb.

Hier ga ik eerst wat meer opzoekwerk over moeten doen want ik heb nog nooit een python programma omgezet in een frontend webapplicatie. Ik ga ook nog toevoegen dat de excel bestanden gedownload worden op de computer van de user die de webscraper gebruikt als deze klaar is.

STAP 11(27/5 – 2/6)

Testen van de webscrapers, kijken of er nog fouten in de code zitten en documenten schrijven.

Communication

Tijdens mijn stage moet ik elke week een status report maken waarin ik kort beschrijf wat ik elke dag gedaan heb. Dit status report stuur ik elke vrijdag door naar mijn placement mentor en naar mijn internship supervisor. Al de informatie over het project waar ik nu aan bezig ben (volledige project molenheide) staat beschreven op jira.

Ik werk alleen aan dit deel van het project dus ik kan niet echt met iemand communiceren over hoe ik de webscraper ga maken. Om bij te houden wat mijn realisaties zijn tot nu heb ik elke week een planning met

mijn placement mentor waar ik kan laten zien wat ik tot dan gemaakt heb en waar ik feedback kan vragen. Later in mijn stage heb ik ook de mogelijkheid om mijn code van de webscraper te laten nakijken door iemand van mijn tribe, maar die heeft het nu te druk met andere projecten.

Who is who

Ik werk in een tribe die gefocust is op het ontwikkelen van websites en webapplicaties. De tribe bestaat uit front-end developers, back-end developers, 1 tribe lead (mijn workspace mentor) en 1 ui/ux developer. Op het project molenheide zijn we nu met 3 mensen aan het werken, mezelf meegerekend. Ik werk wel alleen aan de webscraper, de 2 andere mensen werken aan de frontend/backend van de website van molenheide om deze te verbeteren.