Afbeelding met tekst, illustratie

Automatisch gegenereerde beschrijving **Campus Geel**

**DevOps & Security**

Case study

**Student: Ruben Heurckmans**

**R-number: R0831546**

**Academic year 2022-2023**

[Inleiding 3](#_Toc122030521)

[Link naar github 3](#_Toc122030522)

[Demo 3](#_Toc122030523)

[Video 3](#_Toc122030524)

[Uitleg 3](#_Toc122030525)

[Oplossing 6](#_Toc122030526)

[Aanmaken project 6](#_Toc122030527)

[Selenium toevoegen aan je project 7](#_Toc122030528)

[Code 8](#_Toc122030529)

[Full code 8](#_Toc122030530)

[Code in detail 13](#_Toc122030531)

[Main function 13](#_Toc122030532)

[GetUrl fuction 14](#_Toc122030533)

[StartScraper function 14](#_Toc122030534)

[YtScraper function 15](#_Toc122030535)

[IctJobScraper function 17](#_Toc122030536)

[TwitchScraper function 19](#_Toc122030537)

[WriteFiles function 20](#_Toc122030538)

[Bronnen 21](#_Toc122030539)

[De basis 21](#_Toc122030540)

[Extra 21](#_Toc122030541)

# Inleiding

In deze case study geef ik weer hoe ik een webscraper heb gemaakt voor YouTube, Ictjob en Twitch.

De YouTube scraper geeft de 5 meest recent geüploade video’s op basis van een ingegeven zoekterm. De waarde die je krijgt zijn: de titel, het kanaal van wie de video is, het aantal weergaven en een link naar de video.

De Ictjob scraper geeft de 5 meest recent aangemaakte jobs op basis van keywords die je ingeeft.  
De waarde die je krijgt zijn: de titel, het bedrijf, de locatie, alle keywords van de job en een link naar de detailpagina.

De Twitch scraper geeft de 5 streams met de meeste kijkers van een ingegeven categorie.  
De waarde die je krijgt zijn: de titel, de taal, het kanaal en het aantal live kijkers.

# Link naar github

<https://github.com/Ruben-Heurckmans/DevopsCaseStudy>

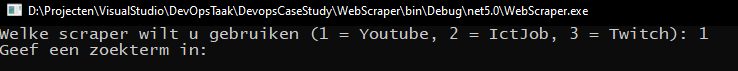
# Demo

## Video

<https://www.youtube.com/watch?v=UutNLhN8moM>

## Uitleg

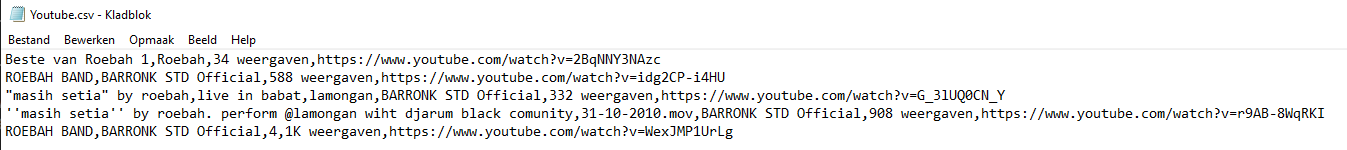
Je start het programma. Deze gaat je de keuze geven tussen drie verschillende webscrapers, deze kan je kiezen door een 1, 2 of een 3 te typen. Daarna vraagt hij om een zoekterm in te geven.



Hierin vul je je zoekterm in en dan gaat de scraper naar de gepaste link aan de hand van de gekozen scraper en de ingegeven zoekterm.

Onderstaande voorbeelden zijn resultaten van de YouTube scraper.



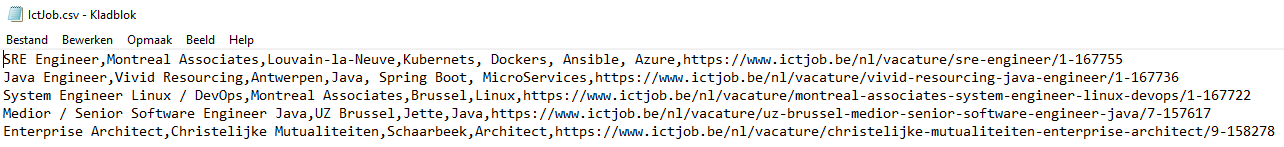


Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving

Hier enkele voorbeelden van de IctJob scraper  

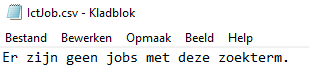
Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

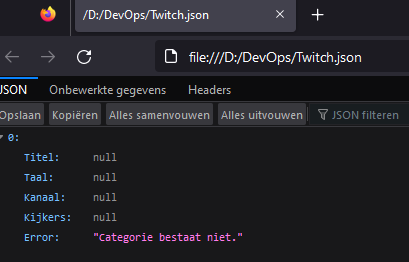
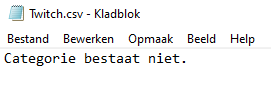
Automatisch gegenereerde beschrijving


Afbeelding met tekst, schermafbeelding, monitor, zwart

Automatisch gegenereerde beschrijving

En hier enkelen van twitch

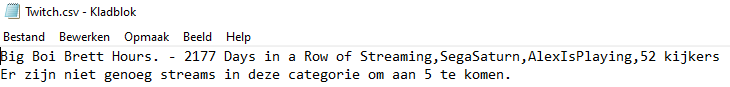
Afbeelding met tekst

Automatisch gegenereerde beschrijving Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving.

Afbeelding met tekst

Automatisch gegenereerde beschrijving Afbeelding met tekst, schermafbeelding, monitor, zwart

Automatisch gegenereerde beschrijving

# Oplossing

## Aanmaken project

Om te beginnen open je visual studio en maak je een Console App aan.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Daarna geef je je project een naam. Mijn project noemt “Scraper”.

Afbeelding met tekst, schermafbeelding, monitor, zwart

Automatisch gegenereerde beschrijvingVoor framework heb ik “.NET 5.0” genomen.

Daarna druk je op create.

## Selenium toevoegen aan je project

Als het project is aangemaakt moeten we onze scraping tool nog toevoegen aan ons project. Ik gebruik hier selenium voor. Om dit toe te voegen klik je rechts op “Solution (naam van je project)” en kies je voor “Manage NuGet Packages for Solution”.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

In de package manager zoek je naar Selenium.WebDriver. De package manager scrolt zelf een beetje naar beneden dus om de juiste te zien moet je wat naar boven scrollen. Selecteer de package en check de checkboxes zoals in de afbeelding. De versie die ik gebruikte is 4.6.0. Druk daarna op install.

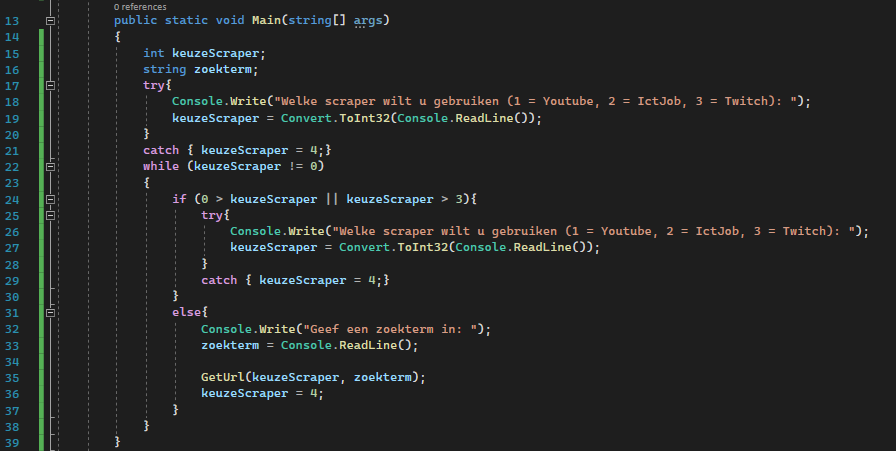
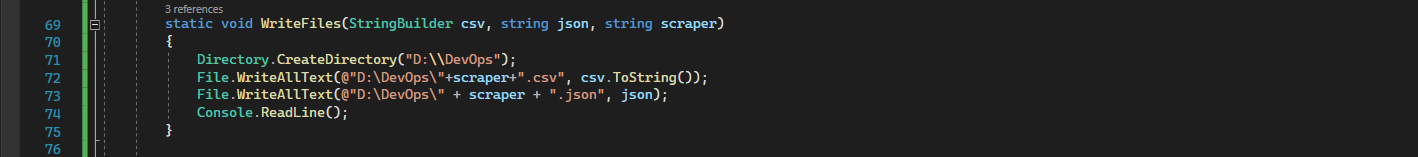
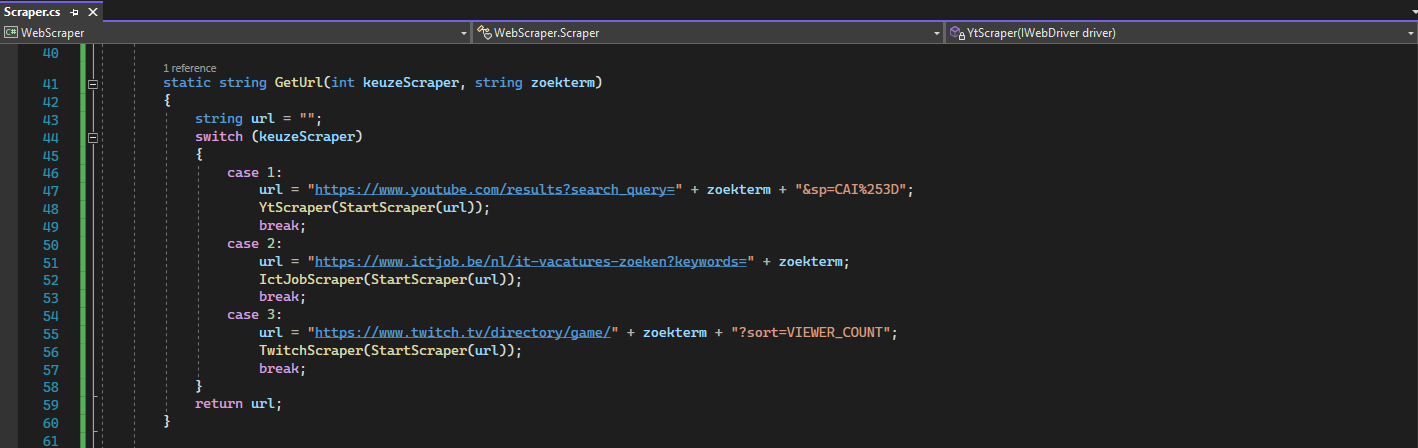
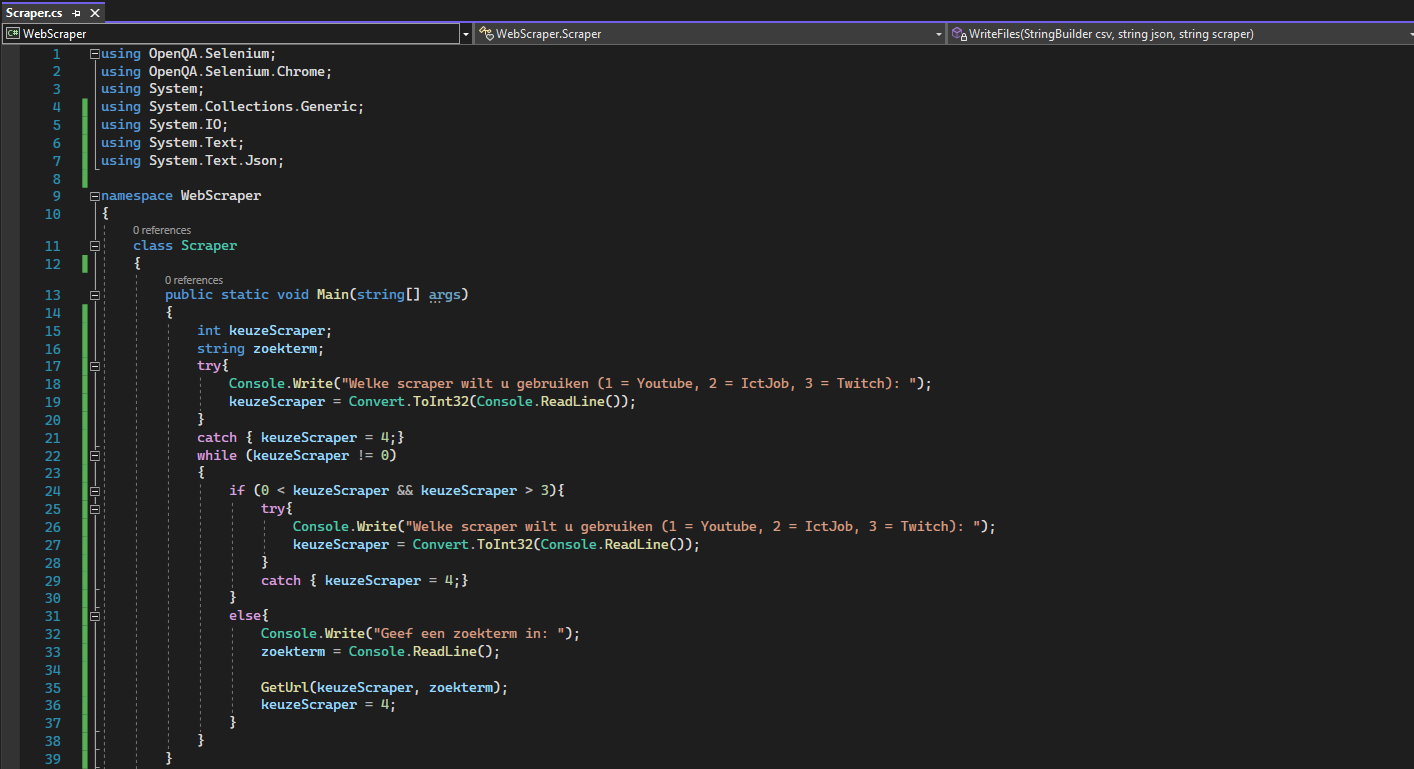
Afbeelding met tekst, monitor, schermafbeelding, binnen

Automatisch gegenereerde beschrijving

## Code

### Full code

Scraper.cs (door kleine aanpassingen kloppen niet alle line numbers)



Afbeelding met tekst

Automatisch gegenereerde beschrijving

Afbeelding met tekst

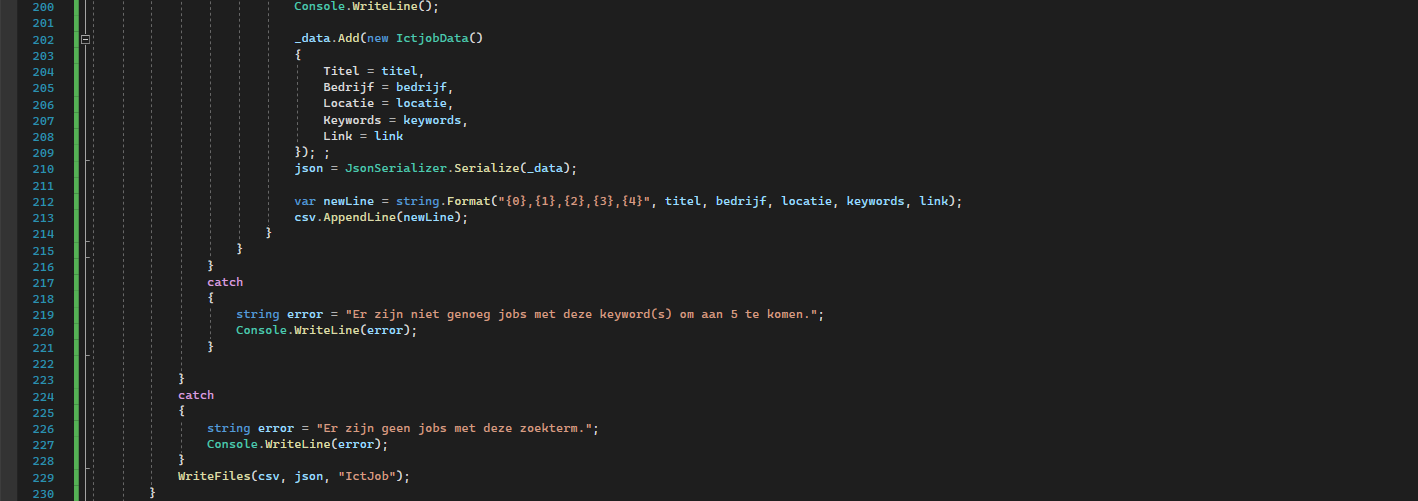
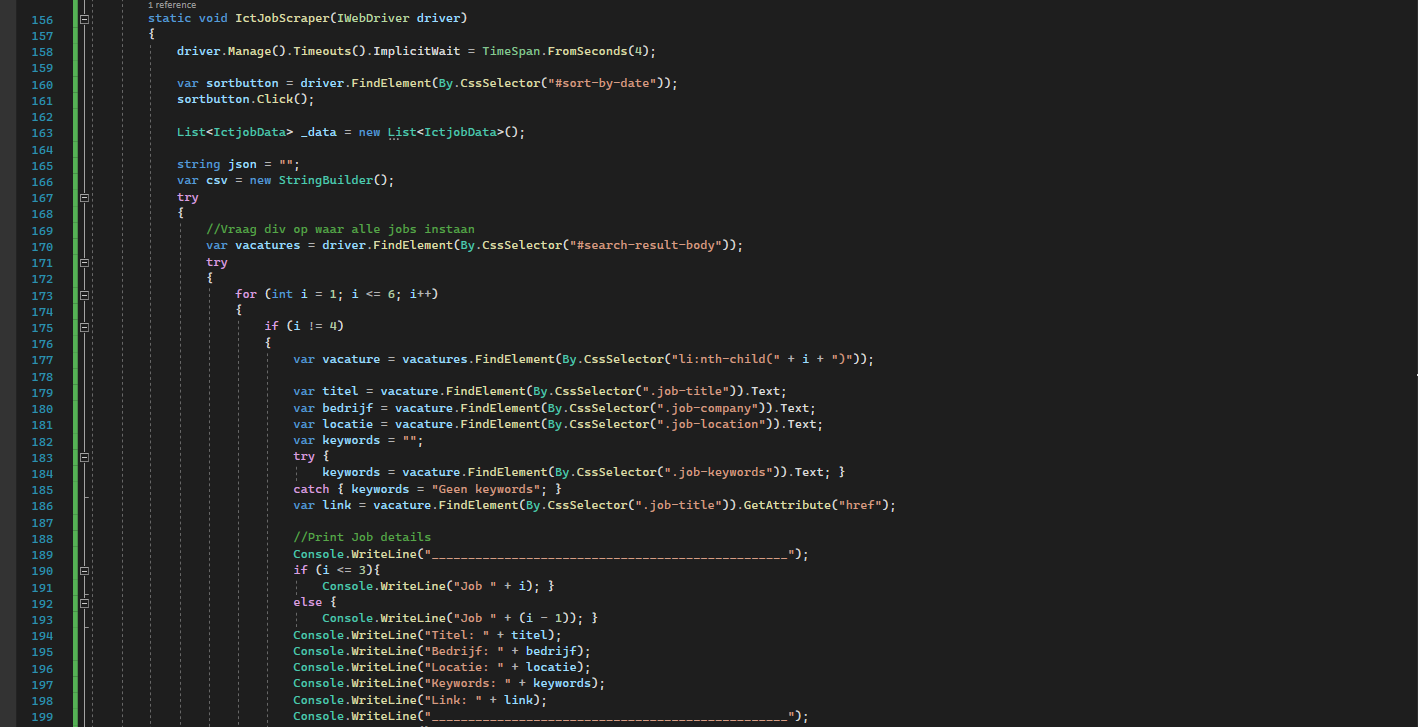
Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving

Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving



Afbeelding met tekst

Automatisch gegenereerde beschrijvingAfbeelding met tekst

Automatisch gegenereerde beschrijving

YoutubeData.cs

Afbeelding met tekst, monitor, scherm, zilver

Automatisch gegenereerde beschrijving

IctjobData.cs

Afbeelding met tekst

Automatisch gegenereerde beschrijving

TwitchData.cs

Afbeelding met tekst, schermafbeelding, elektronica, scherm

Automatisch gegenereerde beschrijving

### Code in detail

#### Main function

Als eerste declareren we de variabelen ”keuzeScraper” en ”zoekterm”.   
Daarna gaan we de gebruiker vragen welke scraper hij/zij wilt gaan gebruiken. De waarde die wordt ingelezen converteren we naar een Int32 en slaan we op in ” KeuzeScraper”. Dit zetten we in een try zodat de applicatie geen error geeft als de gebruiker iets anders dan een geheel getal invult. Als de gebruiker bijvoorbeeld een woord zou ingeven wordt dit opgevangen door de catch waarbij ”keuzeScraper” altijd 4 wordt.

Dan start de while loop. Zolang ”keuzeScraper” niet 0 is blijft deze loop gaande. Als je 0 invult dan stopt de applicatie. Zolang je niet 1, 2 of 3 invult zal de vraag zich blijven herhalen. Ook hier staat dezelfde try, catch rond zodat de applicatie geen error krijgt.

Als je 1, 2 of 3 ingeeft gaat de console vragen om een zoekterm in te vullen. Deze wordt dan opgeslagen onder “zoekterm”.

Daarna wordt de functie “GetUrl()” opgeroepen met “keuzeScraper” en “zoekterm” als parameter.  
Daarna maken we “keuzeScraper” terug 4 zodat hij niet automatisch weer gaat vragen voor een zoekterm.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

#### GetUrl fuction

We declareren als eerste de variabele string. Deze gaat later als parameter gebruikt worden in de fuctie “StartScraper()”.

Om te bepalen welke url we nodig hebben gebruiken we een switch met 3 cases. Eentje voor als de waarde 1 is, eentje voor 2 en eentje voor 3. Welke case het wordt hangt af van de waarde van “keuzeScraper”.

In de cases worden de spaties in “zoekterm” verandert met de waarde dat de site gebruikt om spaties in de url te verwerken. Bij youtube en de ictjob site is dit een “+” en bij twitch “%20”.

Daarna roepen we de functie op van de juiste scraper met als parameter de waarde die we krijgen uit de functie StartScraper met als parameter de url die zonet gemaakt is.

Afbeelding met tekst

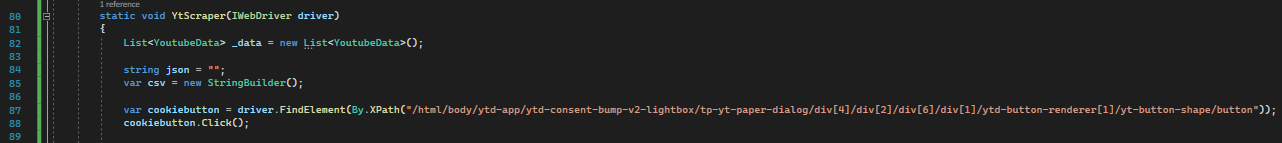
Automatisch gegenereerde beschrijving

#### StartScraper function

Deze functie start de driver, navigeert naar de url die doorgegeven is als parameter en geeft daarna de waarde driver terug. Deze gaat dan gebruikt worden in de scraper functies.  
Afbeelding met tekst

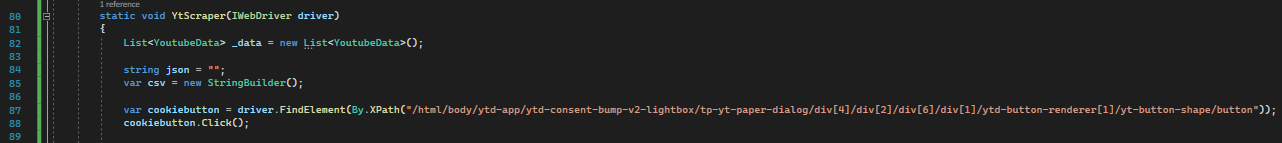
Automatisch gegenereerde beschrijving

#### YtScraper function

Om te beginnen maken we een List aan met als type YoutubeData. Dit is een class in de file YoutubeData.cs, die zichtbaar is in onderstaande screenshot. Deze list gaan we gebruiken om onze json file op te bouwen. Daarna maken we nog een variabele json en csv aan waar we de data naartoe gaan schrijven.

Afbeelding met tekst, monitor, schermafbeelding

Automatisch gegenereerde beschrijvingYoutubeData.cs

Youtube gaat vragen om akkoord te gaan met de cookies. Als we deze pop-up niet wegdoen kunnen we de data niet lezen dus we laten de driver klikken op de knop “akkoord”.

In onderstaande screenshot is de 2de try even ingevouwen zodat het duidelijker wordt.  
We laten de driver zoeken naar een video. Als deze er niet zou zijn krijgen we normaal een error en stopt onze applicatie. Dit vangen we op met een try, catch waarbij er door de catch een errormelding in de console komt te staan en deze melding wordt ook in de Json file en de Csv file gezet.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

De 2de try en catch dient voor als er niet genoeg video’s zijn voor een bepaalde zoekterm. De catch daarvan is precies dezelfde als die van de eerste “try” alleen heeft die een andere error message.

Afbeelding met tekst

Automatisch gegenereerde beschrijvingIn de “try” wordt een for loop 5 keer uitgevoerd. Deze zoekt 5 keer naar een video element en haalt daar de details zoals de titel, het aantal weergaven, … uit. Een video element ziet er uit zoals in de screenshot.

Verder wordt de data in de console getoond en toegevoegd aan de list om dan in de json variabele gestoken te worden en de waarden worden ook in de csv variabele gezet.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Afbeelding met tekst, schermafbeelding, monitor, scherm

Automatisch gegenereerde beschrijving

En op het einde van de functie wordt de “WriteFiles” functie gebruikt met als parameters csv, json en de string “youtube”. Deze functie wordt op het einde uitgelegd.



#### IctJobScraper function

Aangezien er niet veel verschillen zijn tussen de scraperfuncties ga ik vooral focus leggen op de verschillen.

Afbeelding met tekst

Automatisch gegenereerde beschrijvingAls eerste laten we de driver 5 seconden wachten. Dit doen we omdat ictjob een protectie heeft voor scrapers waarbij je even moet wachten voor de pagina laadt. In de screenshot is dit te zien.

Afbeelding met tekst, monitor, schermafbeelding, computer

Automatisch gegenereerde beschrijvingDaarna klikken we op de sort-by-date knop zodat we de juiste sorteermethode gebruiken.

En natuurlijk maken we weer de list aan alleen deze keer met het type IctjobData uit de file “Ictjobdata.cs”

Afbeelding met tekst

Automatisch gegenereerde beschrijving

De 2de try is weer collapsed. We zien hier dat de error message aangepast is zodat het past bij de site die we scrapen.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

De 2de try is ook grotendeels hetzelfde, natuurlijk met andere CssSelectors maar de basis blijft hetzelfde.

De grote verschillen zijn dat we nu de loop 6 keer gaan doen ipv 5 keer. Dit komt omdat er een call to action op de site staat om zelf een job listing aan te maken en dit zit in een div met dezelfde CssSelector als de joblistings zelf.



Hierdoor moeten we natuurlijk de gepaste if, else schrijven zodat hij de vierde loop geen data opvraagt en zodat in de 5de en 6de loop het juiste getal in de console typt.

Afbeelding met tekst

Automatisch gegenereerde beschrijvingOok zien we dat we de juiste data naar de json en csv variabele schrijven.

Daarna roepen we weer de WriteFiles functie aan met de csv, json en de string “IctJob” als parameters.



#### Afbeelding met tekst, monitor, scherm, binnen Automatisch gegenereerde beschrijvingTwitchScraper function

We maken weer een list maar nu met TwitchData.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

In de eerste try gaan we kijken of de categorie wel bestaat. Als deze niet bestaat is er de catch om de error op te vangen

Afbeelding met tekst, schermafbeelding

Automatisch gegenereerde beschrijving

In de 2de try gaan we weer door de loop, bij de taal hebben we een extra try, catch omdat sommige mensen geen taal hebben aanstaan. Ook is het te merken dat er op twitch veel aziatische tekens in de titels, talen, … staan. Om deze leesbaar te maken zetten we de OutputEncoding van de console op UTF8. Dit is vaak niet voldoende, in de demo video zie je hoe ik het font van de console aanpas zodat deze tekens zichtbaar worden.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

Afbeelding met tekst

Automatisch gegenereerde beschrijving

En dan heb je nog de data die naar de json en csv variabele worden geschreven.

De catch die een aangepaste error message heeft

En op het einde weer de Writefiles functie met de jusite parameters.



#### WriteFiles function

De WriteFiles functie maakt een directory aan in de D schijf als deze er nog niet is en plaatst daar de juiste json en csv file in aan de hand van de string die meegegeven is onder “scraper”.

Afbeelding met tekst

Automatisch gegenereerde beschrijving

# Bronnen

## De basis

Scrapax. (2019, 25 augustus). *How to scrape websites using Selenium in C#* [Video]. YouTube. <https://www.youtube.com/watch?v=CpugqTr2j60>

## Extra

Joshi, M. (2022, 29 september). *Wait Commands in Selenium C and C#*. BrowserStack. <https://www.browserstack.com/guide/wait-commands-in-selenium-c-and-c-sharp>

*How to write a JSON file in C#?* (2013, 4 juni). Stack Overflow. <https://stackoverflow.com/questions/16921652/how-to-write-a-json-file-in-c>

*Writing data into CSV file in C#*. (2013, 12 september). Stack Overflow. <https://stackoverflow.com/questions/18757097/writing-data-into-csv-file-in-c-sharp>