Rikko Keuppens

R0790832

DevOps Case study 5 Webscraper Selenium

2de fase IT Factory

Academiejaar 2021-2022

Campus Geel, Kleinhoefstraat 4, BE-2440 Geel

INHOUDSTAFEL

INTRODUCTIE

Ik heb voor het vak DevOps case 5 de Webscraper met Selenium gekozen. Ik ben begonnen met 1 video te bekijken waar ik vervolgens op kon voort werken bij mijn scraper. Het programmeren ging zeer vlot en was vooral ook een zeer leuke opdracht. Voor mijn GitHub actions heb ik gebruik gemaakt van de tips met de links die op canvas stonden en meer eerdere kennis van actions.

Ik geef in dit document de links naar mijn GitHub repo en mijn YouTube video. Ik leg ook uit hoe ik tewerk ben gegaan. Ten slotte zie je mijn volledige code van de webscraper onderaan.

1 LINK NAAR YOUTUBE EN GITHUB REPO

Link naar YouTube video: https://youtu.be/Q6OC1NrEdJQ

Link naar GitHub repo: https://github.com/RikkoKeuppens/SeleniumScraper

2 UITLEG

Voor meer gedetailleerde uitleg staan er commentaar lijnen in mijn code.

Ik vraag bij mijn applicatie voor een keuze te maken tussen 3 websites. Als de gebruiker een bruikbare keuze geeft gaat de applicatie een while loop doen. Het programma blijft runnen als er een antwoord wordt opgegeven. Na de optie van een website moet je een term geven die je gaat opzoek. Dit kan een video zijn voor YouTube, een job voor indeed en een artikel voor de newyork times.

De applicatie gebruikt Xpaths voor de informatie te zoeken. Een chromedriver start op en elke klik wordt manueel met de xpaths geprogrammeerd. De term die gezocht wordt wordt ingegeven en we gaan zo de juiste dingen opzoeken.

Wanneer er niets wordt gevonden laat het systeem dit weten en kan je opnieuw iets opzoeken. Wanneer er wel iets wordt gevonden gaan er een bepaald aantal video's, advertenties of artikels worden gescraped. Hierbij worden telkens de links, titels en andere dingen zoals een auteur gescraped. De gescrapte informatie wordt weergegeven op de console en gaat naar een csv bestand gestuurd worden.

Vervolgens is er op GitHub een zip bestand dat kan gedownload worden via GitHub actions. Dit zorgt ervoor dat GitHub de laatste versie geeft van het exe bestand dat gedownload wordt. Hierbij kan een andere gebruiker dit downloaden en het exe bestand gebruiken. Het artifect wordt gemaakt door een yaml file in de directory.

3 BRONNENLIJST

Scrapax. *YouTube (2019). How to scrape websites using Selenium using c#*
https://www.youtube.com/watch?v=CpugqTr2j60&ab_channel=Scrapax

Thomas Ardul. Elmah.io(2021). Building and testing on multiple .net versions with github actions.
https://blog.elmah.io/building-and-testing-on-multiple-net-versions-with-github-actions/

Wade. .netCoreTutorials(2021). Publishing a single exe file in net core 3-0
https://dotnetcoretutorials.com/2019/06/20/publishing-a-single-exe-file-in-net-core-3-0/

Eric L Larson. Elanderson.net(2020). Use github to publish artifacts
https://elanderson.net/2020/06/github-use-actions-to-publish-artifacts/

4 CODE

```
using System;
using OpenQA.Selenium;
using OpenQA.Selenium.Chrome;
using System.Diagnostics;

namespace WebScraper
{
   class Program
   {
      static void Main(string[] args)
      {
         //Intro
         Console.WriteLine("Welkom op mijn webscraper");
         Console.WriteLine("Kies een website die je wil scrapen...");
         Console.WriteLine("Opties: youtube(y), indeed(i), newyorktimes(n) of druk 'ctrl + c' om te
stoppen");
         string siteChoice = Console.ReadLine();

         //Webscraper opties
         ChromeOptions chOptions = new ChromeOptions();
         chOptions.AddArgument("--log-level=3");
         var chromeDriverService = ChromeDriverService.CreateDefaultService("./");
         chromeDriverService.HideCommandPromptWindow = true;

         //Programma blijft keuzes geven zolang het programma niet wordt gestopt
         //SiteChoice variable wordt gebruikt voor de keuze van website die gescraped wordt
         while (siteChoice != "")
         {
            if (siteChoice == "y" || siteChoice == "youtube")
            {
               //Video variable
               Console.WriteLine("Welke video wil je zoeken?");
               string title = Console.ReadLine();

               //Open youtube en vergroot scherm, klik op zoekbalk en geef de input in
               IWebDriver driver = new ChromeDriver(chromeDriverService, chOptions);
               driver.Navigate().GoToUrl("https://www.youtube.com");
               driver.Manage().Window.Maximize();
               var confirm = driver.FindElement(By.XPath("//*[@id=\"content\"]/div[2]/div[5]/div[2]/ytd-
button-renderer[2]/a"));
               confirm.Click();
               var element = driver.FindElement(By.XPath("/html/body/ytd-app/div/div/ytd-
masthead/div[3]/div[2]/ytd-searchbox/form/div[1]/div[1]/input"));
               element.Click();
               element.SendKeys(title);
               element.Submit();

               //Zoek de meest recente video
               var recent = driver.FindElement(By.XPath("/html/body/ytd-app/div/ytd-page-manager/ytd-
search/div[1]/ytd-two-column-search-results-renderer/div/ytd-section-list-renderer/div[1]/div[2]/ytd-search-
sub-menu-renderer/div[1]/div/ytd-toggle-button-renderer/a/tp-yt-paper-button"));
               recent.Click();
               var uploadDate = driver.FindElement(By.XPath("/html/body/ytd-app/div/ytd-page-
manager/ytd-search/div[1]/ytd-two-column-search-results-renderer/div/ytd-section-list-
renderer/div[1]/div[2]/ytd-search-sub-menu-renderer/div[1]/iron-collapse/div/ytd-search-filter-group-
renderer[5]/ytd-search-filter-renderer[2]/a/div/yt-formatted-string"));
               uploadDate.Click();
               System.Threading.Thread.Sleep(500); //wait half a second, otherwise it will scrape the most
popular videos instead of the most recent

               //Variable voor while loop the stoppen
               int counter = 1;
               int stopCounter = 0;

               Console.WriteLine();

               //Max 5 videos, wanneer er niets meer gevonden wordt stopt de loop
               while (counter < 6 && stopCounter < 1)
               {
```

```
                try
                {
                    //Variable gebruikt voor volgende video te slecter
                    string videoNumber = counter.ToString();

                    //Informatie van elke video wordt opgehaald met xpath, en in een variable gezet
                    string link = "/html/body/ytd-app/div/ytd-page-manager/ytd-search/div[1]/ytd-two-
column-search-results-renderer/div/ytd-section-list-renderer/div[2]/ytd-item-section-renderer/div[3]/ytd-
video-renderer[" + videoNumber + "]/div[1]/div/div[1]/div/h3/a";
                    var videoLink = driver.FindElement(By.XPath(link)).GetDomProperty("href");
                    string Title = "(//*[@id=\"video-title\"]/yt-formatted-string)[" + videoNumber + "]";
                    var videoTitle = driver.FindElement(By.XPath(Title)).Text;
                    string Author = "/html/body/ytd-app/div/ytd-page-manager/ytd-search/div[1]/ytd-two-
column-search-results-renderer/div/ytd-section-list-renderer/div[2]/ytd-item-section-renderer/div[3]/ytd-
video-renderer[" + videoNumber + "]/div[1]/div/div[2]/ytd-channel-name/div/div/yt-formatted-string/a";
                    var videoAuthor = driver.FindElement(By.XPath(Author)).Text;
                    string Views = "(//*[@id=\"metadata-line\"]//span[1])[" + videoNumber + "]";
                    string videoViews = driver.FindElement(By.XPath(Views)).Text;

                    //Info naar cosole geschreven
                    Console.WriteLine("----------------------------------------------------------------------------
----");
                    Console.WriteLine("Titel van de video: " + videoTitle);
                    Console.WriteLine("Auteur: " + videoAuthor);
                    Console.WriteLine("Aantal weergaven: " + videoViews);
                    Console.WriteLine("Link van de video: " + videoLink);

                    //info naar csv file geschreven
                    string filepath = "c:/DevOpsScraperOutput/YouTube.csv";
                    using (System.IO.StreamWriter file = new System.IO.StreamWriter(filepath, true))
                    {
                        file.WriteLine(videoTitle + "," + videoAuthor + "," + videoViews + "," + videoLink);
                    }

                    //Counter gaat omhoog zodat max 5 video's worden gescraped
                    counter++;
                }
                catch
                {
                    //Wanneer er geen video is gevonden gaat stopCounter omhoog en stopt het programma
                    stopCounter++;
                }
                //wanneer er geen video's zijn gevonden verteld de console dit aan de gebruiker
                if (counter == 1)
                {
                    Console.WriteLine("Er zijn geen resultaten voor " + title + " op YouTube");
                    Console.WriteLine("Probeer het opnieuw:");
                }
            }
            if (counter > 1)
            {
                Console.WriteLine("----------------------------------------------------------------------------
");
            }
            //Opnieuw vragen voor website te scrapen
            Console.WriteLine();
            Console.WriteLine("Opties: youtube(y), indeed(i), newyorktimes(n) of druk 'ctrl + c' om te
stoppen");
            siteChoice = Console.ReadLine();
        }
        else if (siteChoice == "i" || siteChoice == "indeed")
        {
            //Advertentie variable
            Console.WriteLine("Kies de job die je wil opzoeken: ");
            string term = Console.ReadLine();

            //Open youtube en vergroot scherm, klik op zoekbalk en geef de input in
            IWebDriver driver = new ChromeDriver(chromeDriverService, chOptions);
            driver.Navigate().GoToUrl("https://be.indeed.com/");
            driver.Manage().Window.Maximize();var element =
driver.FindElement(By.XPath("//*[@id=\"text-input-what\"]"));
            element.Click();
```

```
            element.SendKeys(term);
            element.Submit();

            //Variable voor while loop the stoppen
            int counter = 1;
            int stopCounter = 0;

            //Wanneer je op indeed iets zoekt wat er niet is, kan je niet meer filteren op de laatste 3 dagen
            //Daarom moet de stop counter al gebruikt worden voordat de datum wordt geselecteerd
            while (stopCounter < 1)
            {
                try
                {
                    //Zoek op laatste 3 dagen
                    var button =
driver.FindElement(By.XPath("/html/body/table[1]/tbody/tr/td/div/div[2]/div/div[1]/button/div[1]"));
                    button.Click();
                    var lastDays =
driver.FindElement(By.XPath("/html/body/table[1]/tbody/tr/td/div/div[2]/div/div[1]/ul/li[2]/a"));
                    lastDays.Click();
                    System.Threading.Thread.Sleep(500); //Wacht voor popup
                    var closeButton = driver.FindElement(By.XPath("/html/body/div[5]/div[1]/button"));
                    closeButton.Click();
                    System.Threading.Thread.Sleep(500); // wacht voor nieuwste advertentie

                    //Max 16 resultaten (1 pagina) aan resultaten
                    while (counter < 17)
                    {
                        Console.WriteLine("-----------------------------------------------------------");
                        //Variable gebruikt voor volgende advertentie te slecteren
                        string addNumber = counter.ToString();

                        //Informatie van elke advertentie wordt opgehaald met xpath, en in een variable gezet
                        string Title = "//*[@id=\"mosaic-provider-jobcards\"]/a[" + addNumber +
"]/div[1]/div/div[1]/div/table[1]/tbody/tr/td/div[1]/h2/span";
                        var addTitle = driver.FindElement(By.XPath(Title)).Text.Trim();
                        string CompanyName = "//*[@id=\"mosaic-provider-jobcards\"]/a[" + addNumber +
"]/div[1]/div/div[1]/div/table[1]/tbody/tr/td/div[2]/pre/span";
                        var addCompanyName = driver.FindElement(By.XPath(CompanyName)).Text.Trim();
                        string Location = "//*[@id=\"mosaic-provider-jobcards\"]/a[" + addNumber +
"]/div[1]/div/div[1]/div/table[1]/tbody/tr/td/div[2]/pre/div";
                        var addLocation = driver.FindElement(By.XPath(Location)).GetAttribute("innerHTML");
                        string Link = "//*[@id=\"mosaic-provider-jobcards\"]/a[" + addNumber + "]";
                        var addLink = driver.FindElement(By.XPath(Link)).GetAttribute("href").Trim();

                        //Info naar cosole geschreven
                        Console.WriteLine("Title of the add: " + addTitle);
                        Console.WriteLine("Company name: " + addCompanyName);
                        Console.WriteLine("Company location: " + addLocation);
                        Console.WriteLine("link for the add: " + addLink);

                        //Info naar csv file geschreven
                        string filepath = "c:/DevOpsScraperOutput/Indeed.csv";
                        using (System.IO.StreamWriter file = new System.IO.StreamWriter(filepath, true))
                        {
                            file.WriteLine(addTitle + "," + addCompanyName + "," + addLocation + "," +
addLink);
                        }

                        //Counter gaat omhoog zodat maximum 16 resultaten (1 pagina) wordt getoond
                        counter++;
                    }
                }
                catch
                {
                    //Als er geen resultaten worden gevonden stopt de loop
                    stopCounter++;
                }
            }
            //Als er geen resultaat gevonden is wordt dit getoond aan de gebruiker
            if (counter == 1)
            {
```

```
                Console.WriteLine();
                Console.WriteLine("Er zijn geen resultaten voor " + term + "op Indeed");
                Console.WriteLine("Probeer het opnieuw:");
            }
            //Opnieuw vragen voor website te scrapen
            Console.WriteLine();
            Console.WriteLine("Opties: youtube(y), indeed(i), newyorktimes(n) of druk 'ctrl + c' om te
stoppen");
            siteChoice = Console.ReadLine();
        }
        else if (siteChoice == "n" || siteChoice == "newyorktimes")
        {
            //Artikel variable wordt gevraagd
            Console.WriteLine("Kies een atrikel dat je wil opzoek, gebruik Engels: ");
            string term = Console.ReadLine();

            //NewyorkTimes wordt geopend, artikel variable wordt opgezocht, artikels van gisteren worden
weergegeven
            IWebDriver driver = new ChromeDriver(chromeDriverService, chOptions);
            driver.Navigate().GoToUrl("https://www.nytimes.com/");
            driver.Manage().Window.Maximize();
            var coockie = driver.FindElement(By.XPath("//*[@id=\"site-
content\"]/div[2]/div[1]/div/div[2]/button"));
            coockie.Click();
            var search =
driver.FindElement(By.XPath("//*[@id=\"app\"]/div[2]/div/header/section[1]/div[1]/div[2]/button"));
            search.Click();
            var searchTerm =
driver.FindElement(By.XPath("/html/body/div[1]/div[2]/div/header/section[1]/div[1]/div[2]/div/form/div/inp
ut"));
            searchTerm.SendKeys(term);
            var go =
driver.FindElement(By.XPath("/html/body/div[1]/div[2]/div/header/section[1]/div[1]/div[2]/div/form/button"
));
            go.Click();
            var date =
driver.FindElement(By.XPath("/html/body/div[1]/div[2]/main/div[1]/div[1]/div[2]/div/div/div[1]/div/div/butt
on"));
            date.Click();
            var yesterday =
driver.FindElement(By.XPath("/html/body/div[1]/div[2]/main/div[1]/div[1]/div[2]/div/div/div[1]/div/div/div/
ul/li[2]/button"));
            yesterday.Click();
            System.Threading.Thread.Sleep(500); // wacht voor nieuwste artikels

            //Variable voor while loop the stoppen
            int counter = 1;
            int stopCount = 0;

            Console.WriteLine();

            //Wanneer een term wordt ingegeven die heel breed is zoals "Politics", komt er een andere
volgorde van Xpaths
            //Daarom dat er 2 try's worden gedaan, als beide mislukken zijn er geen artikels
            while (counter < 10)
            {
                try
                {
                    //Variable gebruikt voor volgend artikel te slecteren
                    string addNumber = counter.ToString();

                    //Informatie van elk artikel wordt opgehaald met xpath, en in een variable gezet
                    string Title = "//*[@id=\"site-content\"]/div[1]/div[2]/div[1]/ol/li[" + addNumber +
"]/div/div/div/a/h4";
                    var articleTitle = driver.FindElement(By.XPath(Title)).Text;
                    string Date = "/html/body/div[1]/div[2]/main/div[1]/div[2]/div[1]/ol/li[" + addNumber +
"]/div/span";
                    var articleDate = driver.FindElement(By.XPath(Date)).Text;
                    string Author = "//*[@id=\"site-content\"]/div[1]/div[2]/div[1]/ol/li[" + addNumber +
"]/div/div/div/a/p[2]";
                    var articleAuthor = driver.FindElement(By.XPath(Author)).Text;
```

```
                    string Link = "//*[@id=\"site-content\"]/div[1]/div[2]/div[1]/ol/li[" + addNumber +
"]/div/div/div/a";
                    var articleLink = driver.FindElement(By.XPath(Link)).GetAttribute("href");

                    //Info naar cosole geschreven
                    Console.WriteLine("------------------------------------------------------------");
                    Console.WriteLine("Title of the article: " + articleTitle);
                    Console.WriteLine("Date of the article: " + articleDate);
                    Console.WriteLine(articleAuthor);
                    Console.WriteLine("Link to the article: " + articleLink);

                    //Info naar csv file geschreven
                    string filepath = "c:/DevOpsScraperOutput/NewYorkTimes.csv";
                    using (System.IO.StreamWriter file = new System.IO.StreamWriter(filepath, true))
                    {
                        file.WriteLine(articleTitle + "," + articleDate + "," + articleAuthor + "," + articleLink);
                    }

                    stopCount = 0;
                }
                catch
                {
                    counter++;
                    stopCount++;
                }
                try
                {
                    string addNumber = counter.ToString();

                    string Title = "//*[@id=\"site-content\"]/div[1]/div[2]/div[2]/ol/li[" + addNumber +
"]/div/div/div/a/h4";
                    var articleTitle = driver.FindElement(By.XPath(Title)).Text;
                    string Date = "//*[@id=\"site-content\"]/div[1]/div[2]/div[2]/ol/li[" + addNumber +
"]/div/span";
                    var articleDate = driver.FindElement(By.XPath(Date)).Text;
                    string Author = "//*[@id=\"site-content\"]/div[1]/div[2]/div[2]/ol/li[" + addNumber +
"]/div/div/div/a/p[2]";
                    var articleAuthor = driver.FindElement(By.XPath(Author)).Text;
                    string Link = "//*[@id=\"site-content\"]/div[1]/div[2]/div[2]/ol/li[" + addNumber +
"]/div/div/div/a";
                    var articleLink = driver.FindElement(By.XPath(Link)).GetAttribute("href");

                    Console.WriteLine("------------------------------------------------------------");
                    Console.WriteLine("Title of the article: " + articleTitle);
                    Console.WriteLine("Date of the article: " + articleDate);
                    Console.WriteLine(articleAuthor);
                    Console.WriteLine("Link to the article: " + articleLink);

                    string filepath = "c:/DevOpsScraperOutput/NewYorkTimes.csv";
                    using (System.IO.StreamWriter file = new System.IO.StreamWriter(filepath, true))
                    {
                        file.WriteLine(articleTitle + "," + articleDate + "," + articleAuthor + "," + articleLink);
                    }
                    stopCount = 0;
                }
                catch
                {
                    counter++;
                    stopCount++;
                }
                if (stopCount == 2)
                {
                    //Wanneer geen resultaten worden gevonden, toont de console dit aan de gebruiker
                    Console.WriteLine("Er zijn geen resultaten voor " + term + "op de NewYorkTimes");
                    Console.WriteLine("Probeer het opnieuw:");
                }
            }
            //Opnieuw vragen voor website te scrapen
            Console.WriteLine();
            Console.WriteLine("Opties: youtube(y), indeed(i), newyorktimes(n) of druk 'ctrl + c' om te
stoppen");
            siteChoice = Console.ReadLine();
```

```
            }
            //Wanneer een fout resultaat wordt gegeven bij SiteChoice, geeft de console opnieuw de keuze
aan de gebruiker
            else
            {
               Console.WriteLine("\"" + siteChoice + "\" is een verkeerde keuze, probeer opnieuw");
               Console.WriteLine("Opties: youtube(y), indeed(i), newyorktimes(n) of druk 'ctrl + c' om te
stoppen");
               siteChoice = Console.ReadLine();
            }
         }
      }
   }
}
```