Web Scraping W PHP

Katarzyna Mierzwa 269578

Mateusz Sęk 250487

Mateusz Kwiatkowski 274140

SPIS TREŚCI

Co to jest web scraping?

Web Scraping to metoda, która wyciąga dane ze stron internetowych, co zastępuje odręcznie i powtarzające się kopiowanie i wklejanie danych. Najczęściej wykorzystywany jest:

- do porównywania cen przeróżnych produktów/ usług (ceneo.pl/skyscanner.pl)
- do pobierania danych o potencjalnych klientach pożądanych przez firmę
- do zbierania informacji o produktach/ usługach konkurencji
- do stron, które zbierają informacje z różnych źródeł na dany temat (np. oferty pracy)

Jak działa Web Scraping?

Nasz kod wysyła żądanie do serwera, na którym znajduje się strona, którą chcemy scrapować. Pobiera on kod źródłowy strony, w takim sam sposób jak robi to przeglądarka. Następnie filtruje stronę i szuka elementów HTML, który wyodrębniliśmy w naszym kodzie, by następnie wyświetlić zawartość, którą chcemy wyodrębnić.

Index

index.php > ...

```php
<?php
    require_once('simple_html_dom.php');//dołączamy biblioeke php do tego projektu...
    //biblioteka sluzy do przeszukiwania kodu html pod katem wyszukiwanych wartosci

    require_once('ScrapperItem.php');//dołaczamy klasa na podstwie ktorej tworzymy obiekty poszczegolnych produktow

    require_once('euroScrapper.php');//pliki dołączamy za pomoca ktorego pobieramy dane z euro agd
    require_once('mediaExpertScrapper.php');//...z media expert

    if(count($_GET) > 0){ //sprawdzamy czy ktos uzyl wyszukiwarki
        $mediaExpertURL = generateURLForMediaExpertScrapping($_GET); // jezeli tak generujemy adres url media expert do pobrania produktow

        if($mediaExpertURL !== false){
            $mediaExpertProducts = getMediaExpertHTMLContent($mediaExpertURL); //jesli adrs media epert jest rozny od false
            //bedzie rowny false tylko wtedy gdy uzytkownik nie wybierze sklepu z wyszukiwarki
        }
        else{
            $mediaExpertProducts = [];
            //nie bedzie zadnych produktow nie bedzie niczego
        }

        $euroURL = generateURLForEuroScrapping($_GET); //ta sama historia

        if($euroURL !== false){
            $euroProducts = getEuroHTMLContent($euroURL);
        }
        else{
            $euroProducts = [];
        }

```

index.php > ...

```
30
31          }
32          else{
33              $euroProducts = getEuroHTMLContent(); //pobieramy domyslne produkty z euro
34              $mediaExpertProducts = getMediaExpertHTMLContent(); //pobieramy z media
35              // tylko wtedy gdy nic nie wybral nie wybral zadnego sklepu
36          }
37
38
39
40          $products = array_merge($euroProducts, $mediaExpertProducts);   //laczymy w jedna tablice wszystkie produkty z tablic z euro i media
41          shuffle($products); // miwszzamy tablice zeby wyswietamy sie w lososowej kolejnosci
42      ?>
43
44      <!DOCTYPE html>
45      <html lang="en">
46      <head>
47          <meta charset="UTF-8">
48          <meta http-equiv="X-UA-Compatible" content="IE=edge">
49          <meta name="viewport" content="width=device-width, initial-scale=1.0">
50          <title>Laptopy</title>
51          <link rel="stylesheet" href="https://cdn.jsdelivr.net/npm/bootstrap@5.1.3/dist/css/bootstrap.min.css" integrity="sha384-1BmE4kWBq78iYhF
52      </head>
53      <body>
54      <nav class="navbar navbar-expand-lg navbar-light bg-light">
55        <div class="container-fluid">
56          <a class="navbar-brand" href="#">Wyszukiwanie laptopów</a>
57          </div>
58        </div>
59      </nav>
```

index.php > ...

```php
            <div class="form-group">
                <label class="mb-2">Sklep</label>
                <select multiple class="form-control" name="shop[]">
                    <option value="euro" <?php if(isset($_GET['shop']) && in_array('euro', $_GET['shop'])): ?> selected <?php endif; ?>>Euro RTV AGD</option>
                    <option value="media" <?php if(isset($_GET['shop']) && in_array('media', $_GET['shop'])): ?> selected <?php endif; ?>>Media Expert</option>
                </select>
            </div>
        </div>
        <div class="col-3">
            <div class="form-group">
                <label class="mb-2">Producent</label>
                <select class="form-control" name="producent">
                    <option disabled selected>Wybierz</option>
                    <option <?php if(isset($_GET['producent']) && $_GET['producent'] === 'hp'): ?> selected <?php endif; ?> value="hp">HP</option>
                    <option <?php if(isset($_GET['producent']) && $_GET['producent'] === 'lenovo'): ?> selected <?php endif; ?> value="lenovo">Lenovo</option>
                    <option <?php if(isset($_GET['producent']) && $_GET['producent'] === 'acer'): ?> selected <?php endif; ?> value="acer">Acer</option>
                    <option <?php if(isset($_GET['producent']) && $_GET['producent'] === 'dell'): ?> selected <?php endif; ?> value="dell">Dell</option>
                    <option <?php if(isset($_GET['producent']) && $_GET['producent'] === 'Apple'): ?> selected <?php endif; ?> value="Apple">Apple</option>
                </select>
            </div>
        </div>
        <div class="col-3">
            <div class="form-group">
                <label class="mb-2">Przekątna ekranu</label>
                <select class="form-control" name="screen">
                    <option disabled selected>Wybierz</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '10'): ?> selected <?php endif; ?> value="10">10 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '11'): ?> selected <?php endif; ?> value="11">11 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '12'): ?> selected <?php endif; ?> value="12">12 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '13'): ?> selected <?php endif; ?> value="13">13 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '14'): ?> selected <?php endif; ?> value="14">14 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '15'): ?> selected <?php endif; ?> value="15">15 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '16'): ?> selected <?php endif; ?> value="16">16 cali</option>
                    <option <?php if(isset($_GET['screen']) && $_GET['screen'] === '17'): ?> selected <?php endif; ?> value="17">17 cali</option>
                </select>
            </div>
        </div>
```

EuroScrapper

euroScrapper.php > PHP Intelephense > getEuroHTMLContent

```php
<?php

function getEuroHTMLContent($url = 'https://www.euro.com.pl/laptopy-i-netbooki.bhtml'){ //funkcja majaca za zadanie pobranie danych ze strony euro i przetworzenie ich do
    $euroHTML = file_get_html($url); //pobranie danych ze strony euro

    $products = []; //deklarjemy pusta tablice produktow

    foreach($euroHTML->find('.product-box') as $product){ //przechodzimy petlo po wszystkich elemetach html ktora zwrocila funkcja
        $attribiutes = []; //pusta tablica w ktorej znajda sie poszczegone czesci opisu produktow

        if(!count($product->find('.product-name')) || !count($product->find('.product-photo .photo-hover')) || !count($product->find('.price-normal'))){
            continue; //sprawdzamy czy istnieje nazwa zdjecie i cena , jezeli nie istnieje ktorych z tych elementow to przechodzmy do kolejnej iteracji
        }

        $title = trim(preg_replace('/\s\s+/', ' ', $product->find('.product-name')[0]->plaintext)); //pobieramy tytul usuwamy z niego spacje przed i po
        $image = 'https://www.euro.com.pl' . $product->find('.product-photo .photo-hover')[0]->{'data-hover'}; //pobieramy link do zdjecia prduktu
        $price = trim(preg_replace('/\s\s+/', ' ', $product->find('.price-normal')[0]->plaintext));//pobieramy cene i usuwamy spacje z pocztaku i konca

        foreach($product->find('.product-attributes .attributes-row') as $attribute){//iterujemy po poszczegolnych fragmentach opisu produktow
            array_push($attribiutes, trim(preg_replace('/\s\s+/', ' ', $attribute->plaintext)));  //i wkladmy do tablicy i usuwamy biale znaki
        }

        $shopUrl = 'https://www.euro.com.pl' . $product->find('.product-name a')[0]->href; //pobieramy link do produkty
        //dodajemy do tablicy nowy obiekt
        array_push($products, new ScrapperItem($title, $image, $price, $attribiutes,  $shopUrl, 'https://f00.esfr.pl/img/desktop/euro/logo.png'));
    }


    return $products; //zwracamy cala tablce produktow
}

function getScreenSize($size){ //funkcja ktora przeksztalca wielkosc ekranu na taka ktora jest akceptowalna przez wyszukiwarke euro
    $sizes = [
        '10' => '11',
        '11' => '!11-2-13-1',
        '12' => '!11-2-13-1',
        '13' => '!13-2-14-1',
```

```php

function generateURLForEuroScrapping($params){ //funkcja ktora generuje adres url
    if(isset($_GET['shop']) && !in_array('euro', $_GET['shop'])){ //jesli w wyszukiwarce nie zostal wybrany sklep euro to funcja zwroci false
        return false;
    }

    $url = 'https://www.euro.com.pl/laptopy-i-netbooki{producent}{przekatna}{ram}{dysk}{matryca}.bhtml';//bazowy adres wyszukiwarki

    //sprawdzamy czy w wyszukiwarce zostal wybrany producent jezeli tak to dadajemy go do linku z pozostalymi wartosciami tak samo czyli ram itd
    if(isset($_GET['producent'])){
        $url = str_replace('{producent}', ",_" . $_GET['producent'], $url);
    }
    else{
        $url = str_replace('{producent}', "", $url);
    }

    if(isset($_GET['screen'])){
        $url = str_replace('{przekatna}', ",przekatna-ekranu-cale-" .  getScreenSize($_GET['screen']), $url);
    }
    else{
        $url = str_replace('{przekatna}', "", $url);
    }

    if(isset($_GET['ram'])){
        $url = str_replace('{ram}', ",pamiec-ram_2!" .  $_GET['ram'] . "-gb", $url);
    }
    else{
        $url = str_replace('{ram}', "", $url);
    }
    if(isset($_GET['disc'])){
        $url = str_replace('{dysk}', ",dysk-ssd!" .  getDiskCapacity($_GET['disc']), $url);
    }
    else{
        $url = str_replace('{dysk}', "", $url);
    }

    if(isset($_GET['matrix'])){
```

ScrapperItem

ScrapperItem.php > PHP Intelephense > ScrapperItem

```php
<?php

class ScrapperItem{
    public $title;
    public $image;
    public $price;
    public $description;
    public $shop_url;
    public $shop;

    public function __construct($title = '', $image = '', $price = '', $description = [], $shop_url = '', $shop = ''){
        $this->title = $title;
        $this->image = $image;
        $this->price = $price;
        $this->description = $description;
        $this->shop_url = $shop_url;
        $this->shop = $shop;
    }
}
```

Biblioteka Simple HTML DOM Parser

Jest to skrypt który potrafi pobrać zawartość innej strony oraz przefiltrować ją w poszukiwaniu interesującego nas elementu. PHP Simple HTML DOM potrafi dokładnie skopiować wybrany fragment strony. Działaniem przypomina bibliotekę jQuery bo właśnie w podobny sposób poszukujemy wybranych elementów, po klasie, id, po nazwie tagu HTML. Możemy używać skrypt w pętli co daje potężne możliwości analizy całych serwisów.

Technologia

Stos technologiczny:

- Visual Studio Code
- PHP
- Bootstrap
- Xampp Apache

Bibliografia

1. Link
2. Link
3. Link
4. Link