Łukasz Jaworski 155 NCI

Lab 03

WebScrapping

Webscrapping to operacja ‘wydobycia’ informacji z innego źródła, w naszym przypadku witryny internetowej. W swoim labolatorium skorzystałem z biblioteki **Jsoup** dla Javy do pobrania listy linków oraz tytułów z rankingu najlepszych pozycji na Netflixie w serwisie **Filmweb.pl**.

```
import java.io.IOException;
public class WebScraper {
    public static void main(String[] args) throws IOException {
        String url = "http://www.filmweb.pl/ranking/netflix";
        System.out.println(("Fetching from "+ url));

        Document doc = Jsoup.connect(url).get();
        Elements filmLinks = doc.select( cssQuery: "a.film__link");

            for (Element link : filmLinks) {
                System.out.println(" Film link:" + link.attr( attributeKey: "abs:href") + ", film title: " + link.text());
            }
```

Program nie jest długi, jednak jest to właśnie zasługa biblioteki Jsoup.

Po określeniu adresu url z którego program ma ‘wyciągać’ dane, trzeba określić jakie elementy nas interesują. Po zbadaniu elementów na stronie dotarłem do interesujących mnie wpisów oraz informacji jak są zapisane na stronie.

```
<a class="film__link" href="/serial/Narcos-2015-680486">Narcos</a
```

Dzięki dokumentacji biblioteki (https://jsoup.org/apidocs/org/jsoup/select/Selector.html?fbclid=IwAR3VcE59mgYcJQBjgRJXOyjeW4O4Z-YrQbLSCtAbRMtalkmgE2rOZqPpdW4) określiłem formułę która pozwoli programowi wybrać interesujące mnie elementy.

Następnie w pętli wyświetlam wszystkie pobrane informacje z danej strony. Jako, że pobieram całą linię kodu programu, muszę określić też którą jej część mam aktualnie wyświetlić. Poniżej prezentuje wyniki pracy programu.

```
C:\Java\jdk1.8.0_191\bin\java.exe ...
Scrapping from http://www.filmweb.pl/ranking/netflix
 Review link:https://www.filmweb.pl/serial/Narcos-2015-680486, production title: Narcos
 Review link:https://www.filmweb.pl/serial/BoJack+Horseman-2014-718443, production title: BoJack Horseman
 Review link:https://www.filmweb.pl/serial/House+of+Cards-2013-620036, production title: House of Cards
 Review link:https://www.filmweb.pl/serial/Stranger+Things-2016-750359, production title: Stranger Things
 Review link:https://www.filmweb.pl/serial/Chef%27s+Table-2015-744119, production title: Chef's Table
 Review link:https://www.filmweb.pl/serial/Dark-2017-771383, production title: Dark
 Review link:https://www.filmweb.pl/serial/The+Crown-2016-747284, production title: The Crown
 Review link:https://www.filmweb.pl/serial/Abstrakt%3A+Sztuka+designu-2017-785613, production title: Abstrakt: Sztuka designu
 Review link:https://www.filmweb.pl/serial/Making+a+Murderer-2015-752805, production title: Making a Murderer
 Review link:https://www.filmweb.pl/serial/The+Get+Down-2016-737390, production title: The Get Down
```

Druga klasa **ThreadScrapper** działa na podobnej zasadzie – tym razem wykorzystałem jednak dane ze strony **http://scraping.pro/web-scraper-test-drive/**. Posiada ona listę narzędzi wykorzystywanych do WebScrapping'u.

W tej klasie wykorzystałem 3 wątki, aby kolejno pobrały różne informacje – nazwy testowanych programów, kryteria pod względem których były oceniane oraz komentarze użytkowników. Wyniki pracy programu widać poniżej.