

GitHub: <https://github.com/TheBestSoftInTheWorld/web-scraping>

After run project we get data from website defined **searchUrl** field. The data are save to XML file.

1)command line:

mvn -q exec:java -e -Dexec.mainClass=com.mycompany.app.App

```
C:\moje_aplikacje\web-scraping>mvn -q exec:java -e -Dexec.mainClass=com.mycompany.app.App
title: a123456789z
<?xml version="1.0" encoding="UTF-8"?><URL><inner id="1"><url>#services</url><text>Poznaj Nas</text></inner><inner id="2"><url>#welcome</url><text>a123456789z</text></inner><inner id="3"><url>#services</url><text>Co robimy ?</text></inner><inner id="4"><url>#resume</url><text>STACK TECHNOLOGICZNY</text></inner><inner id="5"><url>#portfolio</url><text>PORTFOLIO</text></inner><inner id="6"><url>#testimonials</url><text>OFERTA</text></inner><inner id="7"><url>#contact</url><text>kontakt</text></inner><inner id="8"><url>#contact</url><text>Zatrudnij nas</text></inner><inner id="9"><url>#</url><text>WSZYSTKIE REALIZACJE</text></inner><inner id="10"><url>#</url><text>PORTALE INTERNETOWE</text></inner><inner id="11"><url>#</url><text>INNE PRACE</text></inner><innerURL id="1"><url>http://a123456789z.com/wp-content/themes/01_parallax-background/assets/img/portfolio/www1.jpg</url><text></innerURL><innerURL id="2"><url>http://a123456789z.com/wp-content/themes/01_parallax-background/assets/img/portfolio/pdf-2127829_1920.png</url><text></innerURL><inner id="12"><url>#contact</url><text>Zatrudnij nas</text></inner><inner id="13"><url>#</url><text>+48 883 664 616</text></inner><inner id="14"><url>#</url><text>hi@a123456789z.com</text></inner><outerURL id="1"><url>https://web.facebook.com/a123456789zcom-2001135716793805</url><text></outerURL><outerURL id="2"><url>https://www.linkedin.com/company/27165975/</url><text></outerURL></URL>
```

2) Eclipse

The screenshot shows the Eclipse IDE with a Java project named 'App'. The main class is 'App.java', which contains the following code:

```
134
135     }
136
137     }
138
139     // write the content into xml file
140     TransformerFactory transformerFactory = TransformerFactory.newInstance();
141     Transformer transformer = transformerFactory.newTransformer();
142     DOMSource source = new DOMSource(doc);
143
144     // StreamResult result = new StreamResult(new File("C:/XML/file.xml"));
145     StreamResult result = new StreamResult(System.out);
146     transformer.transform(source, result);
147
148     } catch (ParserConfigurationException pce) {
149         pce.printStackTrace();
150     } catch (TransformerException tfe) {
151         tfe.printStackTrace();
152     }
153 }
154
155 private static void generateList(String text, String href) {
156     if (href.startsWith("http") || href.startsWith("https")) {
157
```

The right sidebar shows the 'Outline' view with the following structure:

- com.mycompany.app
 - App
 - searchUrl : String
 - url : List<URLContent>
 - main(String[]) : void
 - generateXML() : void
 - generateList(String, String) : void

The bottom console shows the output of the application:

```
<terminated> App (4) [Java Application] C:\Program Files\Java\jre1.8.0_161\bin\javaw.exe (1 kwi 2019, 17:05:24)
title: a123456789z
<?xml version="1.0" encoding="UTF-8"?><URL><inner id="1"><url>#services</url><text>Poznaj Nas</text></inner><inner id="2"><url>#welcome</url><text>a1234
```