

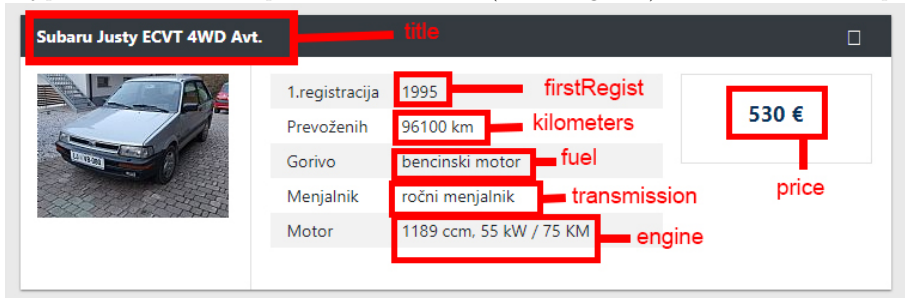
# IEPS Poročilo 2 - Ekipa "Vipavska Burja"

Matjaž Rupnik, Blaž Ličen

April 2021

## 1 Izbrana stran

Za svoj primer smo si izbrali spletno stran: avto.net (seznam oglasov) Primer iz seznama in prikaz podatkovnih polj:



## 2 Implementacije

### 2.1 Seznam regularnih izrazov, za ekstrakcijo podatkov iz spletnih strani

#### 2.1.1 rtvslo.si

```
<div class="author-name">([<]*)</div>
<div class="publish-meta">([/]*<*)</div>
<h1>([<]*)</h1>
<div class="subtitle">([<]*)</div>
<p class="lead">([<]*)</p>
<article^[^>]+?> (?=.)|(?=<p^[^>]*>\s*(.*)</p>)|(?=.) </article>
```

#### 2.1.2 overstock.com

```
<a href="http://www.overstock.com/cgi-bin/d2.cgi?PAGE=PROFRAME&PROD_ID=[0-9]+"><b>([<]*)</b></a>
<td align="left" nowrap="nowrap"><s>([<]*)
<span class="bigred"><b>([<]*)
<span class="littleorange">([<]*)</span>
<span class="normal">([<]*)<br>
```

#### 2.1.3 avto.net

```
<div class="GO-Results-Naziv bg-dark px-3 py-2 font-weight-bold text-truncate text-whitetext-decoration-none">[\n\r\s]+
<span>([<]*)</span>
<td class="w-75 pl-3">([<]*)</td>
<td class="d-none d-md-block pl-3">Prevoženih</td>[\n\r\s]+<td class="pl-3">([<]*)</td>
<td class="d-none d-md-block pl-3">Gorivo</td>[\n\r\s]+<td class="pl-3">([<]*)</td>
<td class="d-none d-md-block pl-3">Menjalnik</td>[\n\r\s]+<td class="pl-3 text-truncate">([<]*)
<td class="d-none d-md-block pl-3">Motor</td>[\n\r\s]+<td class="pl-3 text-truncate">([<]*)
<div class="d-none d-sm-block col-auto px-sm-0 pb-sm-3 GO-Results-PriceLogo">[\n\r\s]+<!--(?:.*?)->[\n\r\s]+
<!--(?:.*?)->[\n\r\s]+<div class="GO-Results-Price mt-0 mt-sm-3">[\n\r\s]+
<!--(?:.*?)->[\n\r\s]+<!--(?:.*?)->[\n\r\s]+<!--(?:.*?)->[\n\r\s]+
<div class="GO-Results-Price-Mid">[\n\r\s]+<div class="GO-Results-Price-TXT-Regular">([<]*)</div>
```

## 2.2 Seznam XPath izrazov, za ekstrakcijo podatkov iz spletnih strani

### 2.2.1 rtvslo.si

```
string(/html/body/div[@id="main-container"]/div[3]/div/div[1]/div[1]/div)

string(/html/body/div[@id="main-container"]/div[3]/div/div[1]/div[2])

string(/html/body/div[@id="main-container"]/div[3]/div/header/h1)

string(/html/body/div[@id="main-container"]/div[3]/div/header/div[2])

string(/html/body/div[@id="main-container"]/div[3]/div/header/p)

/html/body/div[@id="main-container"]/div[3]/div/div[2]/descendant::p/text()[not(ancestor::script)]
```

### 2.2.2 overstock.com

```
/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/a/b

/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/table/tbody/tr/td[1]/
table/tbody/tr[1]/td[2]/s

/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/table/tbody/tr/td[1]/
table/tbody/tr[1]/td[2]/s

/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/table/tbody/tr/td[1]/
table/tbody/tr[3]/td[2]/span

/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td/table/tbody/tr/td[2]/span
```

### 2.2.3 avto.net

```
/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[1]/span

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[4]/div/table/tbody/tr[1]/td[2]

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[4]/div/table/tbody/tr[2]/td[2]

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[4]/div/table/tbody/tr[3]/td[2]

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[4]/div/table/tbody/tr[4]/td[2]

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[4]/div/table/tbody/tr[5]/td[2]

/html/body/strong/div[1]/div[3]/div/div[5]/form/div/div[contains(@class, "GO-Results-PriceLogo")]/div[1]/div[1]/div
```

## 3 Implementacija avtomatske ekstrakcije

Za avtomatsko ekstrakcijo smo se odločili izdelati RoadRunner (like) algoritem. Z delom smo začeli, vendar nam naše izvedbe ni uspelo dokončati.