Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Luka Dragar, Matej Vatovec, Aleksander Kovač

**EXTRACTION**

2. domača naloga za predmet Iskanje in ekstrakcija podatkov s spleta

MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

# Uvod

Poročilo opisuje implementacijo treh načinov ekstrakcije podatkov s treh tipov spletnih strani rtvslo.si, overstock.com in (custom).

# Opis izbranih spletnih strani

Za dodatni dve strani smo si izbrali dva produkta s strani mimovrste.si (povejta), kjer smo izbrali 6 parametrov in jih izluščili. Parametri so označeni na slikah 1 in 2. Se zmeniti.

# Implementacija

Pred prvim zagonom je potrebno zagnati ukaz pip install -r .\requirements.txt s katerim namestimo potrebne knjižnice.

## Tabela regularnih izrazov in XPath za posamezne strani

|  |
| --- |
| **REGEX** |
| **(RTV Slo)** Title: r˝<h1>(.\*).</h1>˝ |
| Subtitle: r˝<div class=\˝subtitle\˝>(.\*)</div>˝ |
| Lead: r˝<p class=\˝lead\˝>(.\*)</p>˝ |
| Author: r˝<div class=\˝author-name\˝>(.\*)</div>˝ |
| Meta: r˝<div class=\˝publish-meta\˝>\n\t\t(.\*)<br>˝ |
| Content: r˝<div class=\˝article-body\˝>(.\*?)<div class=\˝gallery\˝>˝ |
| **(Overstock)** Titles: r˝<td><a href=\˝http://www\.overstock\.com/cgi-bin/d2\.cgi\?PAG˝E=PROFRAME[\w\W]\*?\><b>(.\*?)</b>˝ |
| List prices: r˝<s>(.\*?)</s>˝ |
| Prices: r˝<span class=\˝bigred\˝><b>(.\*?)</b>˝ |
| Savings: r˝<b>You Save:[\w\W]\*?class=\˝littleorange\˝>(.\*?)</span>˝ |
| Contents: r˝<span class=\˝normal\˝>(.\*?)</span>˝ |
| **(Custom)** ID1 |
|  |
| XPATH |
| **RTV Slo** Title: //\*[@id=˝main-container˝]/div[3]/div/header/h1 (**[@id=˝main-container˝]** |
| Subtitle: //\*[@id=˝main-container˝]/div[3]/div/header/div[2] |
| Lead: //\*[@id=˝main-container˝]/div[3]/div/header/p |
| Author: //\*[@id=˝main-container]/div[3]/div/div[1]/div |
| Meta: //\*[@id=˝main-container]/div[3]/div/div[1]/div[2]/text()[1] |
| Content: //\*[@id=˝main-container]/div[3]/div/div[2]/article |
| **Overstock (glavni XPATH, se upošteva pri vse nadaljnjih – razen tam kjer je eksplicitno navedeno drugače)** : /html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr{i}/td[2] |
| Titles: /a/b |
| List prices: /table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s |
| Prices: /table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b |
| Savings: /table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span |
| Contents (XPATH se razlikuje pri delu, kjer je v glavni poti tr{i}):  /html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr[3]/td[2]/table/tbody/tr/td[2]/span/text() |

## Road runner

Algoritem prejme dve strani kot vhod. Najprej se pregleda dolžina teh strani. V primeru, da sta enako dolgi, se poskuša generirati wrapper. V nasprotnem primeru pa se stran razdeli na več blokov (html oznake na isti globini v drevesu) in se rekurzivno poskuša generirati wrapper za vsak blok.

Ideja algoritma je bila povzeta po članku <http://vldb.org/conf/2001/P109.pdf>, kjer opisuje RoadRunner algoritem, nekaj idej pa je bilo vzetih iz <https://www.unixuser.org/~euske/python/webstemmer/howitworks.html>.

Generirani wrapperji so priloženi v ločenih datotekah v repozitoriju znotraj mape wrappers.