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

V poročilu smo povzeli tri načine, kako se lahko izluščijo podatki iz treh različnih spletnih strani. Prva je bila rtvslo.si ter druga overstock.com.

# Opis izbranih spletnih strani

Za 2 izbirni strani smo si izbrali ponudnike študentske prehrane, kjer smo izbrali katere podatke želimo izluščiti s pomočjo XPATH-a, regexov in algoritma »Road Runner«. Sledeča slika prikazuje izbrane parametre.

Slika, ki vsebuje besede besedilo, posnetek zaslona, spletna stran, pisava

Opis je samodejno ustvarjen

.

# Implementacija

Ukaz »pip install -r requirements.txt namesti potrebne pakete potem pa zaženemo run-extraction.py skripto ter odvisno od zahtevanega algoritma dodamo arugment (A – regex, B – xpath, C – »Road Runner«).

## Tabela regularnih izrazov in XPath za posamezne strani

|  |
| --- |
| ***REGEX*** |
| **RTV Slo**˝ |
| regex\_list = [              (r"<h1>(.\*)</h1>", "Title"),              (r"<div class=\"subtitle\">(.\*)</div>", "SubTitle"),              (r"<p class=\"lead\">(.\*)</p>", "Lead"),              (r"<div class=\"author-name\">(.\*)</div>", "Author"),              (r"<div class=\"publish-meta\">\n\t\t(.\*)<br>", "PublishedTime"),              (r"<div class=\"article-body\">(.\*?)<div class=\"gallery\">", "Content")          ] |
| **Overstock** |
| regex\_list = [          (r"<td><a href=\"http://www\.overstock\.com/cgi-bin/d2\.cgi\?PAGE=PROFRAME[\w\W]\*?\"><b>(.\*?)</b>", "Title"),          (r"<s>(.\*?)</s>", "List price"),          (r"<span class=\"bigred\"><b>(.\*?)</b>", "Price"),          (r"<b>You Save:[\w\W]\*?class=\"littleorange\">(.\*?)</span>", "Saving"),          (r"<span class=\"normal\">(.\*?)</span>", "Content")      ] |
| **Študentska prehrana** |
| regex\_list = [          (r"<h3 class=\"no-margin bold\">(.\*?)</h3>", "Title"),          (r"<small>(.\*?)</small>[\w\W]\*", "Address"),          (r"<span class=\" color-light-grey\">(.\*?)</span>", "Price"),          (r"<span class=\" color-light-grey\">(.\*?)<\/span>\s\*<\/small>", "List price"),          (r"<div class=\"col-md-12 text-bold\">(.\*?)</div>", "Work time"),          (r"<strong class=\" color-blue\">(.\*?)</strong>", "Main dish"),          (r"<ul class=\"list-unstyled\">(?:\s\*<li>.\*?</li>){1}\s\*<li>.\*?<i>(.\*?)</i>", "Salad"),          ] |
| ***XPATH*** |
| **RTV Slo** |
| {  "Title": '//˝\*[@id="main-container"]/div[3]/div/header/h1',  "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[1]/div',  "PublishedTime":'//\*[@id="main-container"]/div[3]/div/div[1]/div[2]/text()[1]',  "Content": '//\*[@id="main-container"]/div[3]/div/div[2]/article',          } |
| **Overstock**  common\_path = '/html/body/'  common\_path = f'{common\_path}table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr[{i}]/td[2]'  {  "Title": f'{common\_path}/a/b',  "List price": f'{common\_path}/table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s',  "Price": f'/{common\_path}/table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b',  "Saving": f'{common\_path}/table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span',  "Content": '/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()'                  } |
| **Študentska prehrana**  common\_path = '/html/body/'  common\_path = f'{common\_path}div[3]/div[2]'   {                      "Locale name": f'{common\_path}/div[1]/div[1]/div[2]/div[1]/h3[1]',                      "Address": f'{common\_path}/div[1]/div[1]/div[2]/div[1]/small[1]',                      "Price": f'{common\_path}/div[1]/div[1]/div[2]/div[1]/small[2]/span[2]',                      "List price": f'{common\_path}/div[1]/div[1]/div[2]/div[1]/small[2]/span[4]',                      "Work time": f'{common\_path}/div[2]/div[1]/div[1]/div[1]/div[3]/div[1]/div[2]/div[1]',                      "Main dish": f'//strong[@class="color-blue"][{i}]',                      "Salad": f'//ul[@class=list-unstyled][{i}]/li[2]/i[1]'                  } |

## 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.