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

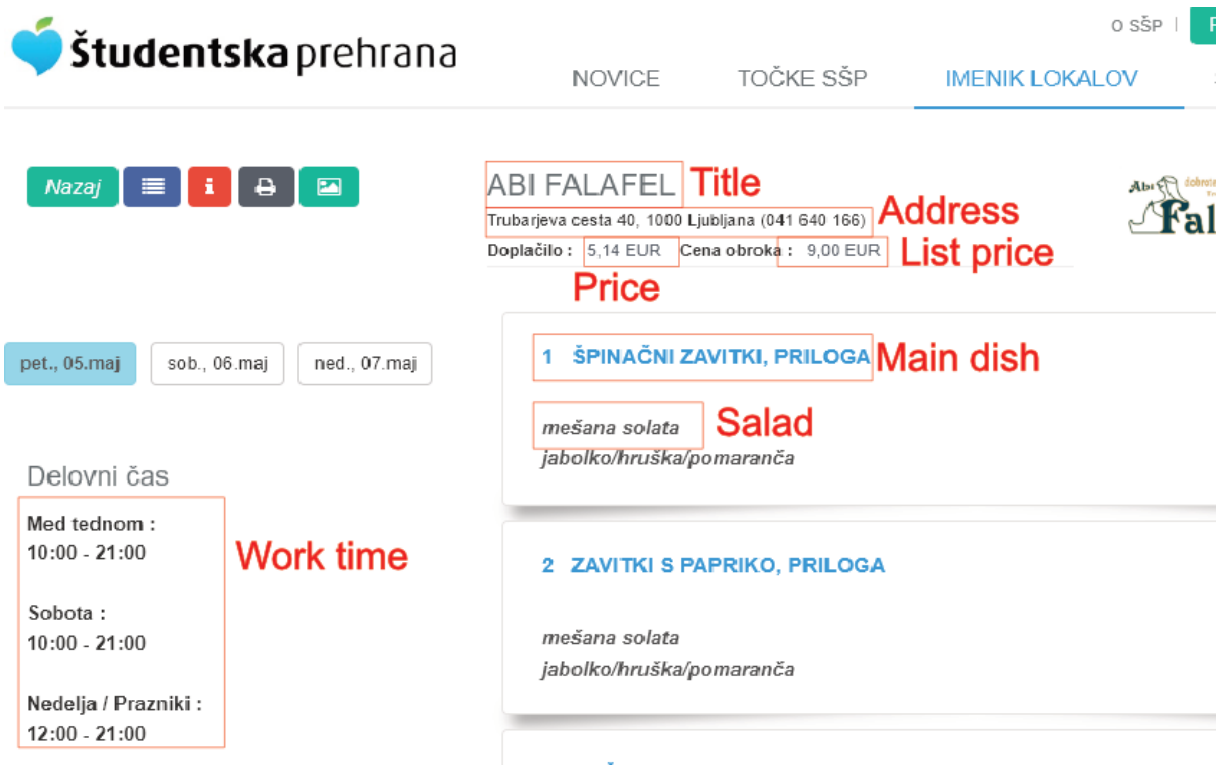
MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

1. 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.

2. 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.



.

3. 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«).

3.1 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"<div class=\"col-md-12 text-bold\">(.*?)</div>", "Work time"),
    (r"<strong class=\" color-blue\">(.*?)</strong>", "Main dish"),
    (r"<i class=\"text-bold color-dark\">(.*?)</i>", "Salad"),
]
```

XPATH

RTV Slo

```
for key, xpath in {
    "Title": '//header[@class="article-header"]/h1',
    "SubTitle": '//div[@class="subtitle"]',
                "Lead": '//p[@class="lead"]',
                "Author": '//div[@class="author"]/div',
                "PublishedTime": '//div[@class="publish-meta"]/text()[1]',
                "Content": '//article[@class="article"]', #regex would be a better option for this case.
}.items():
    element = site_string.xpath(xpath)
    if element:
```

Overstock

```
common_path = f'{common_path}table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr[1]/td[2]'
data = {}
for key, xpath in {
    "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'
```

Študentska prehrana

```
data['Locale name'] = single_value(
    site_string, f'{common_path}/div[1]/div[1]/div[2]/div[1]/h3[1]')
data['Address'] = single_value(
    site_string, f'{common_path}/div[1]/div[1]/div[2]/div[1]/small[1]')
data['Price'] = single_value(
    site_string, f'{common_path}/div[1]/div[1]/div[2]/div[1]/small[2]/span[2]')
data['List price'] = single_value(
    site_string, f'{common_path}/div[1]/div[1]/div[2]/div[1]/small[2]/span[4]')
data['Work time'] = single_value(
    site_string, f'{common_path}/div[2]/div[1]/div[1]/div[1]/div[1]/div[1]/div[3]/div[1]/div[2]/div[1]')  # need combining with regex to arhieve best result
data['Salad'] = single_value(
    site_string, '//ul[@class="list-unstyled"][1]/li[2]/i[1]')
data['Main dish'] = single_value(
    site_string, f'//strong[@class=" color-blue"][1]')
print(json.dumps(data, indent=4, ensure_ascii=False))
```

3.2 Road runner

Glaven del algoritma predstavlja rekurzivna funkcija match(). Metoda se premika po obeh dokumentih in primerja vrstice. Če se ujemata, vrstico doda v wrapper. Če se ne pa imamo dve možnosti ali je iterator ali optional. Podrobnosti metode so prikazane v psevdokodi:

```
function match(wrapper, sample, new_wrapper, start1, start2):
    if start1 >= len(wrapper) or start2 >= len(sample)
        return new_wrapper

    if wrapper[start1] == sample[start2] or (wrapper[start1] and sample[start2]
are text):
        update new_wrapper
        Return match(wrapper, sample, new_wrapper, start1+1, start2+1)
    else:
        found = False
        repeat for both wrapper and sample:
            end = find_end(wrapper, start1)
            if end is not None:
                satart = find_start(wrapper, start1-1)
                if start is not None:
                    section1 = wrapper[start:start1]
                    section2 = wrapper[start1:end+1]
                    iterator = match(section1, section2, [], 0, 0)
                    if iterator is not None:
                        found = True
                        return   match(wrapper,   sample,   combine(new_wrapper,
iterator))
        if not found:
            i1 = index where wrapper matches sample[start2]
            i2 = index where sample matches wrapper[start1]
            if i1 < i2:
                for i in range(start1, i1):
                    mark wrapper[i] as optional in new_wrapper
                return match(wrapper, sample, new_wrapper, i1, start2)
            else if i2<i1:
                for i in range(start2, i2):
                    mark wrapper[i] as optional in new_wrapper
                return match(wrapper, sample, new_wrapper, i1, start2)
            else:
                return None
```

```
(opt) (<div>)?
(opt) (<div>)?
(<figure>
<a>
<img>
</a>
<figcaption>
MMC RTV SLO
</figcaption>
</figure>
)+</div>
<div>
<div>
</div>
<div>
<figure>
<a>
<img>
</a>
<figcaption>
.*
</figcaption>
</figure>
<figure>
<a>
<img>
</a>
<figcaption>
.*
</figcaption>
</figure>
<figure>
<a>
<img>
</a>
<figcaption>
.*
</figcaption>
</figure>
<figure>
<a>
<img>
</a>
<figcaption>
```

Primer wrapperja na spletni strani rtv.si