

Poročilo za drugo domačo nalogo(Data extraction from the Web)

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

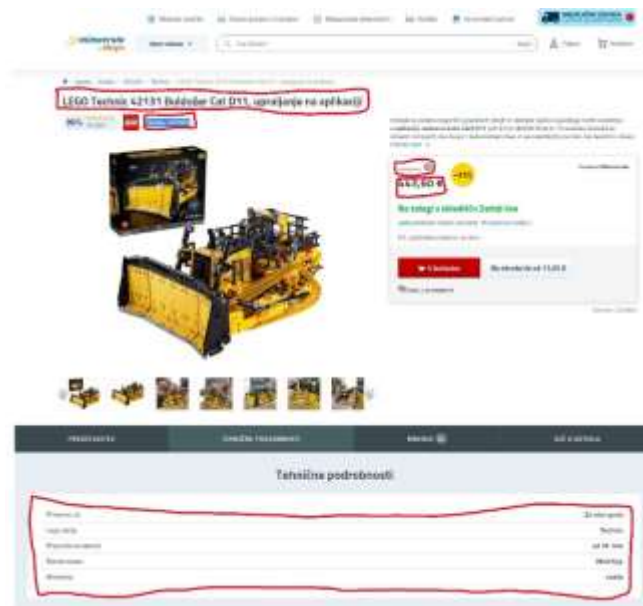
MENTORJA: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

1. Uvod

V tem poročilu bomo opisali izdelavo treh načinov ekstrakcije podatkov iz spleta: regularni izrazi, xpath in implementacija splošnega algoritma Webstemmer. Izbrane spletne strani so overstock.com, rtv.si in mimovrste.si.

2. Opis izbranih spletnih strani

Za dodatne dve strani smo izbrali produkte iz spletne strani mimovrste.si. Iz spletne strani smo ekstrahirali: naslov, številko izdelka, ceno, znižano ceno, dostopnost in tehnične lastnosti izdelka. Parametri so označeni na sliki 1.



Slika 1 ekstrahirani podatki

3. Implementacija regularnih izrazov in xpath-a

Regular Expressions (Regex) omogočajo relativno učinkovito iskanje želenih elementov v HTML kodi, čeprav orodje ni zasnovano za to uporabo. XPath je orodje, posebej zasnovano za iskanje elementov v XML in HTML strukturah. Omogoča enostavno definiranje lokacije elementa v drevesni strukturi in po potrebi iskanje njegovih pod-elementov. Iskanje običajno poteka tako, da definiramo tip in prepoznavno lastnost zelenega elementa.

[illegible]

```

<?xml
title = <L.*>{.*}</L>
subtitle = <div class="<title">{.*}</div>
lead = <p class="lead">{.*}</p>
author = <div class="author-name">{.*}</div>
time = <div class="publish-meta">{.*}</div>
content = <div class="article-body">{.*}</div class="gallery">{

```

```
#discrete
title = cld("class=trial_title", "1", "2", "3", "4")
owner = cgl("class=panel-order-title_inn", "0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
availability = cgl("class=panel-order-title_availability", "0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
old_price = cgl("class=price", "0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
final_price = cld("class=price_wap", "0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
table param = cld("class=product_parameters", "parameter=0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
table values = cld("class=product_parameters", "parameter=0", "1", "2", "3", "4", "5", "6", "7", "8", "9")
```

[illegible]

```

<?xml
  title = //header[@class="article-header"]/*[text()
  subtitle = //title[@class="subtitle"]/*[text()]]&
  lead = //p[@class="lead"]/*[text()
  author = //div[@class="author-name"]/*[text()
  date = //div[@class="publish-meta"]/*[text()
  content = //div[@class="article-body-media"]/*[Figure|FigureCaption|text()]] //div[@class="article-body"]/*[article|p|text()

```

```

<script>
title //id[@class="detail_title detail_title_desktop"]//text()
number //span[@class="detail-panel-under-title_text"]//text()
availability //div[@class="availability-box status availability-box status-available"]//text()
priceOld //span[@class="price_wrap_box_old_container_value"]//span/text()
priceFinal //div[@class="price_wrap_box_final"]//span/text()
table_row_params //table[@class="product-parameters_table"]//tbody/tr/td[@class="product-parameters_parameter"]//text()
table_row_value //table[@class="product-parameters_table"]//tbody/tr/td[@class="product-parameters_parameter product-parameters_value"]//text()

```

Regularni izrazi

Xpath-a

Slika 2 Implementacija regularnih izrazov in xpath

4. Implementacija Webstemmer

Funkcija webstemmer deluje tako, da sprejme dve HTML strani in pragove za merjenje podobnosti ter odkrivanje naslova in glavnega besedila. Razčleni strani, izračuna matriko podobnosti in združi podobne strani v grozde. Nato odstrani statične bloke in odkrije naslov ter glavno besedilo znotraj grozdov. Rezultate pretvori v JSON objekt in jih vrne kot končni rezultat.

Pseudocode rešitve

1. `webStemmer(html_page1, html_page2, sim_threshold, diff_threshold, title_threshold, main_text_threshold)`
 - `wrapper ← []`
2. Razčlenjevanje strani:
 - `parsed_pages ← parse_page(html)` za vsako html stran
3. Izračun podobnosti:
 - `sim_matrix ← compute_similarity_matrix(parsed_pages)`
4. Grupiranje in ustvarjanje vzorca postavitve:
 - `clusters ← cluster_pages(sim_matrix, sim_threshold)`
5. Obdelava vsakega grozda:
 - Za vsak grozd:
 - Izračunaj `diff_scores` za bloke v grozdu
 - Odstrani statične bloke (`diff_scores > diff_threshold`)

- Poišči naslov in glavno besedilo (upoštevaj `title_threshold`, `main_text_threshold`)
 - Dodaj rezultate v wrapper
- 6. Vrne wrapper kot JSON niz

```
[
  "a",
  "span"
],
"Click here to purchase."
],
[
  [
    [
      "tr",
      "td",
      "tr",
      "td",
      "table",
      "tbody",
      "tr",
      "td",
      "a",
      "tr",
      "td"
    ],
    "More Info..."
  ],
  [
    [
      "td",
      "a"
    ],
    "14-kt. Diamond 7.5-8 mm Pearl Pendant"
  ],
  [
    [
      "table",
      "tbody",
      "tr",
      "td",
      "table",
      "tbody",
      "tr",
      "td"
    ],
    "List Price:"
  ],
  [
    [
      "td"
    ],
    "$196.99"
  ],
  [
    [
      "tr",
      "td"
    ],
    "Price:"
  ],
  [

```

Overstock

```
"Temni na\u010din BETA"
],
[
  [],
  "BETA"
],
[
  [
    [
      "label",
      "div",
      "div",
      "div",
      "div"
    ],
    "Prijavljen"
  ],
  [
    [
      "button"
    ],
    "\u00d77"
  ],
  [
    [
      "div"
    ],
    "Odjava"
  ],
  [
    [
      [],
      "Uporabni\u0161ki ra\u010dun"
    ],
    [
      [
        "div",
        "div"
      ],
      "Temni na\u010din BETA"
    ],
    [
      [],
      "BETA"
    ],
    [
      [
        [
          "label",
          "div",
          "div",
          "div",
          "div"
        ],
        "\u010casovno obdobje po meri"
      ],
      [
        "button"
      ]
    ]
  ]

```

RTV

```
],
"Partner"
],
[
  [
    "g"
  ],
  "Pridru\u017eite se mimovrste=)Partner programu"
],
[
  [],
  "Item 1 of 5"
],
[
  [
    [
      "div",
      "div",
      "div",
      "div",
      "div",
      "ul"
    ],
    "Nakupovanje na mimovrste=)"
  ],
  [
    [
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "li",
      "a",
      "div",
      "div",
      "ul"
    ],
    "Kontakt in pomo\u0161"
  ],
  [
    [
      "li",
      "a",
      "li",
      "a"
    ]
  ]

```

Mimovrste

Slika 3 izhodni wrapper za Overstock,RTV in Mimovrste