

Ekstrakcija podatkov iz spleta

2. domača naloga pri predmetu IEPS

Žiga Černigoj
63130028

Marko Lavrinec
63130134

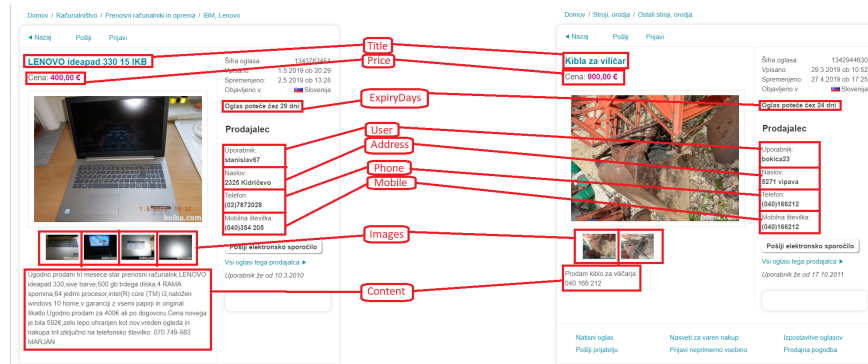
Marec 2019

1 Uvod

Cilj naloge je razviti program, ki s pomočjo XPATH-a, Regularnih izrazov in z Road Runner-jem izvozilo želene podatke iz spletne strani. Za implementacijo bo uporabljen programski jezik Python.

2 Izbira dodatnih dveh spletnih strani

Za dodatni dve strani, ki ju zahteva naloga, smo izbrali 2 naključni strani s prodajajočim se izdelkom na strani bolha.com. Na strani želimo pobirati podatke, ki jih prikazuje Slika 1.



Slika 1: Zajemajoči podatki na strani bolha.com

Na strani se pobirajo podatki Title, Price, DaysUntilExpires, User, Address, Phone, Mobile, Content in MainImage, ki so tipa string, kot tudi Images ki predstavlja seznam nizov.

3 Implementacija

3.1 XPATH

Na začetku se pokliče metoda `process_file`, ki se ji poda stran, ki jo želimo obdelati. Ta nato s pomočjo `lxml` knjižnice formatira vhodni HTML v objekt, nad katerim nato s pomočjo funkcije `xpath` dobimo željene rezultate.

3.1.1 rtvslo.si

Na tej strani smo definirali sledeče poti:

```
1 {
2     'Title': '//*[@id="main-container"]/h1/text()',
3     'SubTitle': '//*[@id="main-container"]/*[@class="
4         subtitle"]/text()',
5     'Lead': '//*[@id="main-container"]/*[@class="lead
6         "]/text()',
7     'Content': '//*[@id="main-container"]/*[@class="
8         article-body"]/*[@class="article"]//descendant:
9         :p/text()',
10    'Author': '//*[@id="main-container"]/*[@class="
11        author-name"]/text()',
12    'PublishedTime': '//*[@id="main-container"]/*[
13        @class="publish-meta"]/text()'
14 }
```

Pri vseh vsebinah, ki jih želimo se sklicujemo na id-je in class-e, za katere menimo, da so najbolj konsistentni in se zgoraj zagotovo ponovijo pri vseh oblikah člankov na tej strani. Poleg njih se usmerimo tudi na html značke (`div`, `p`, `h1`).

Ena od izjem na strani pa je pri polju `Content`. Tam namreč znotraj class-a `article-body` s funkcijo `descendant::p` izvozimo vsebino vseh člankov. Pri izvozu vsebine je bila še posebej pomembna značka `p`, saj bi bila vsebina brez nje polna javascript kode.

3.1.2 overstock.com

Na tej strani smo prvotno hoteli izvesti 2 XPATH ukaza, prvi bi bil:

`'//table[@cellpadding="2"]/tbody/tr[@bgcolor]'` za pridobitev vseh tabel z željeno vsebino, drugi pa potem nadaljna pot do zelenega cilja.

Ta pot bi bila namreč smiselnejša, saj če bi v kateri od tabel manjkalo kakšen od podatkov, bi se pri direktnem dostopu to lahko zamaknilo in rezultat na koncu nebi bil točen.

Ker sta podani strani zadosti konsistentni, da nobeden od podatkov ne manjka, smo lahko definirali naslednje XPATH-e, ki smo jih nato po vrsti razvrstili v naš rezultat:

```

1 {
2   'Title': '//table[@cellpadding="2"]/tbody/tr[
3     @bgcolor]/td[2]/a/b/text()',
4   'Content': '//table[@cellpadding="2"]/tbody/tr[
5     @bgcolor]/td[2]/table//span[@class="normal"]/
6     text()',
7   'ListPrice': '//table[@cellpadding="2"]/tbody/tr[
8     @bgcolor]/td[2]/table//s/text()',
9   'Price': '//table[@cellpadding="2"]/tbody/tr[
10    @bgcolor]/td[2]/table//span[@class="bigred"]/b/
11    text()',
12   'Saving': '//table[@cellpadding="2"]/tbody/tr[
13    @bgcolor]/td[2]/table//span[@class="
14    littleorange"]/text()'
15 }

```

Na tej strani s pomočjo // dobimo vse elemente, ki se nahajajo na istem mestu po tabelah. Tako nam zgornji rezultati javljajo sezname nizov, ki pa jih nato zaporedno pretvorimo v naš rezultat. Glede na vsebino strani se namreč rezultat v tabeli za Title nahaja na istem mestu kot tudi indeksi za ostale strani.

3.1.3 bolha.com

Na strani smo definirali sledeče XPATH-e:

```

1 {
2   'Title': '//*[@id="adDetail"]/div[2]/h1/text()',
3   'Price': '//*[@id="adDetail"]//div[@class="price"]
4     /span/text()',
5   'DaysUntilExpires': '//*[@id="adDetail"]/div[3]/
6     div[1]/p[5]/text()',
7   'User': '//*[@id="sellerInfo"]/div/p[1]/strong/
8     text()',
9   'Address': '//*[@id="sellerInfo"]/div/p[2]/strong/
10    text()',
11   'Phone': '//*[@id="sellerInfo"]/div/p[3]/strong/
12    text()',
13   'Mobile': '//*[@id="sellerInfo"]/div/p[4]/strong/
14    text()',
15   'Content': '//*[@id="adDetail"]//div[@class="
16     content"]/descendant::*/*text()',
17   'MainImage': '//*[@id="gallery"]/table/tbody/tr/td
18     /a/img/@src',
19   'Images': '//*[@id="gal"]//td[@class="thumb"]/a/
20     img/@src',
21 }

```

Tu se križata 2 vrsti vsebine. Prva je običajna vsebina, ki se pobere po strani, druga vrsta pa je seznam vsebine (Images). Ta nam namreč vrne seznam vseh slik, za razliko od ostalih, ki nam vračajo samo dobljene nize na določeni poti.

Pri strani se zanašamo na najbolj razpoznavne class-e in id-je. Pri slikah namesto besedila izvažamo kar povezave do slik, kar dosežemo z @src.

3.2 Regularni izrazi

Na začetku se pokliče metoda `process_file`, ki se ji poda stran, ki jo želimo obdelati. Strane se poda kot parameter funkciji `re.findall` skupaj z regularnimi izrazi specifičnimi za določen del strani. Dobljen rezultat zgornje funkcije predstavlja zahtevano rešitev.

3.2.1 `rtvslo.si`

Ob začetku dela na tej strani smo želeli zagotoviti regularne izraze, ki bi zagotovili delovanje tudi ob dodajanju classov ali drugih spremembah nad gledanimi elementi. Primer takega ukaza bi bil:

```
1 r"<div [^>]*class=\" [^\"]*subtitle [^\"]*\" [^>]*>(.*?)</div [^>]*>"
```

A nam je zgornji primer povzročil več težav kot koristi, zato smo se raje omejili na izraze, ki vrnejo samo vsebino celotnega diva glede na trenutno sestavo:

```
1 r'<div class="subtitle">(.*?)</div>'
```

Problem pri tej strani nam je predstavljala tudi vsebina na tej strani, ki smo jo želeli izvoziti z enim ukazom:

```
1 r'<article class="article">.+(?=<p)<p [^>]*>(.(+?(?=</p))</p>'
```

A ker je na strani pod vsebino več značk p z vsebino, smo se odločili za pristop da najprej izvozimo kar celoten article, tako kot je to označeno v navodilih, naknadno pa ga prečistimo.

Naši regularni izrazi za stran so:

```
1 {
2     'Title': r"<h1 [^>]*>(.*?)</h1 [^>]*>",
3     'SubTitle': r'<div class="subtitle">(.*?)</div>',
4     'Lead': r'<p class="lead">(.*?)</p>',
5     'Content': r'<article class="article">(.(+?(?=</article)))',
6     'Author': r'<div class="author-name">(.*?)</div>',
7     'PublishedTime': r'<div class="publish-meta">(.*?)<br>'
8 }
```

3.2.2 overstock.com

Na tej strani smo prvotno hoteli izvesti dvokratni regularni izraz. Prvi bi bil za pridobitev vseh tabel z željeno vsebino, drugi pa potem nadaljna pot do zelenega cilja.

Skladno z navodili pa smo nato naredili en celoten ukaz, ki celotno pot do zadetka ustvari sam:

```
1 {
2   'Title': r'</table></td><td valign="top"> <a href
      =" [^\"]*" ><b>(.*?)</b>',
3   'Content': '<span class="normal">([<]*)<',
4   'ListPrice': '<s>([<]*)</s>',
5   'Price': '<span class="bigred"><b>([<]*)</b></span>',
6   'Saving': r'<span class="littleorange">([^\s]*)\s[
      ^<]*</span>',
7   'SavingPercent': r'<span class="littleorange">[^\(
      ]*\(( [^<]* )\)</span>'
8 }
```

Pri tem smo se sklicevali na najkrajše možne unikatne nize, ki nas pripeljejo do rezultata. Tako smo se zanašali na class-e, in pa na značke, ki se nahajajo okoli zelenega rezultata.

3.2.3 bolha.com

Na strani smo definirali sledeče regularne izraze:

```
1 {
2   'Title': r"<h1 [^>]*>(.*?)</h1 [^>]*>",
3   'Price': '<div class="price">Cena: <span>([<]*)</span></div>',
4   'DaysUntilExpires': r'<p class="validTo">Oglas
      poteče čez ([^\s]*) dni </p>',
5   'User': '<p><label>Uporabnik:</label><strong>([<]*
      *)</strong></p>',
6   'Address': '<p><label>Naslov:</label><strong>([<]*
      *)<',
7   'Phone': '<p><label>Telefon:</label><strong>([<]*
      *)<',
8   'Mobile': '<p><label>Mobilna številka:</label><
      strong>([<]*)<',
9   'Content': '<div class="content">(.*?(?=</div))</div>',
10  'MainImage': '<td class="imgHolder">.*.+?(?=<img)<img src
              ='([^\"]*)''
12 }

```

Pri večini regularnih izrazov se zanašamo na vsebino, ki se nahaja med HTML značakami ``, pri zadnjih dveh za slike pa na vsebino med narekovaji atributa `src`. Za določeno vsebino se zanašamo tudi na predhodno vsebino labela.

3.3 Road Runner

4 Dobljeni rezultati

4.1 XPATH

4.1.1 rtvslo.si

rtvslo.si/Audi A6 50 TDI quattro_ nemir v premijskem razredu - RTVSLO.si.html:

```

1 {
2   "Title": "Audi A6 50 TDI quattro: nemir v premijskem
           razredu",
3   "SubTitle": "Test nove generacije",
4   "Lead": "To je novi audi A6. V razred najdražjih in
           najbolj premijskih žrebcev je vnesel nemir, že
           preden je sploh zapeljal na parkirni prostor,
           rezerviran za izvržnega direktorja. ",
5   "Content": "Samo pogledajte njegovo masko € to ogromno
           satovje z radarji na takem položaju, da se ti
           na avtocesti tudi pri 120 km/h vsi spožtljivo
           umikajo, saj so prepričani, da gre za Pahorjev
           ali žarčev avto. Seveda, novi A6 lahko cesto in
           promet skenira s kar petimi radarji, petimi
           kamerami, infrardečo kamero za nočni vid,
           dvanajstimi ultrazvočnimi senzorji in laserskim
           čitalnikom € lidarjem. V glavnem vojaška
           tehnologija v službi varnosti za fante, ki smo
           radi gledali Top Gun, Bonda in druge možakarja s
           finimi igračami. Vozniški delovni prostor je
           novo poglavje digitalne dobe, z dvema ogromnima
           zaslonoma, ki tako kot naprednejši telefoni
           dregnejo blazinice vaših prstov, kot se
           sprehajate po steklu. A že bolj se nam zdi
           pomembno, da so osnovna stikala tam, kjer jih
           pričakujete. Najprej so torej zagotovili
           enostavno osnovo, tisti bolj \"advanced\"
           vozniki pa si lahko nato vse skupaj že veliko
           bolj prilagodijo. Velik korak naprej pri

```

kabinskem udobju zaznavajo tudi na zadnji klopi, tam je prostora v vseh smereh precej več. če vam pogled na Audijev spisek dodatne opreme ne odvzame volje do življenja, potem vsekakor toplo priporočamo nakup zračnega vzmetenja, saj dobi z njim A6 več različnih in vozniško zelo uporabnih karakterjev. Enako velja za seksi luči z inteligentno matrično osvetlitvijo, pa za ž portno podvozje in vsekakor za žtirikolesno krmiljenje. S tem postane A6 med ovinki v obč utku na volanu že veliko krajši in bolj agilen. Vse našteto smo preskušali v družbi agregata 50 TDI, ki je v resnici klasični trilitrski dizel, podkrepjen z elektromotorjem. Ja, ta audi je mehki hibrid z izjemnim navorom in dovolj moči kadar koli in kjer koli. Si pa mislimo, da bo največji del trga zadovoljil že učinkovit dvolitrski mehki hibrid z močjo 150 kilovatov. - na testu Audi A6 50 TDI quattro tiptronic - dol žina: 4,9 m - medosna razdalja: 2,9 m - obrač alni krog: 12,1 m - prtljažnik: 530 l - masa: 1.900 kg - trilitrski žestvaljni dizelski motor - moč: 210 kW - navor: 620 Nm - 8-stopenjski samodejni menjalnik - pogon na vsa žtiri kolesa - pnevmatike: 225/60 R17 - poraba: 6,6 l/100 km = 8,9 EUR/100 km - posoda za gorivo: 73 l - doseg: 1.106 km - izpusti CO : 147 g/km - nakupna cena: 69.080 EUR - strožki finančnega lizinga: 4.463 EUR/5 let - strožki registracije: 10.829 EUR/5 let - strožki vzdrževanja: 1.926 EUR/5 let - strožki goriva: 6.702 EUR/75.000 km - strožek 1 kompleta pnevmatik: 716 EUR - vrednosti po 5 letih po Eurotaxu: 33.964 EUR - strožki skupaj: 1.001 EUR/mesec",

"Author": "Miha Merljak",

"PublishedTime": " 28. december 2018 ob 08:51"

rtvslo.si/Volvo XC 40 D4 AWD momentum_ suvereno med najboljše v razredu
- RTVSLO.si.htm:

```
{
  "Title": "Volvo XC 40 D4 AWD momentum: suvereno med
    najboljše v razredu",
  "SubTitle": "Test novega modela",
  "Lead": "XC 40 je najmanjši Volvov SUV, ki se
```

oblikovno skoraj v celotni naslanja na oba večja predhodnika. že samo s tem so mu vrata do denarnic tistih kupcev, ki iščejo izstopajočo, a hkrati visoko kultivirano in prečiščeno dizajnersko govorico, na pol odprta.",

"Content": "Volvo se je nižjih srednjih razredov v preteklosti izogibal ali pa je vanje vstopal z zelo nižnjimi produkti, ki niso pustili večjega tržnega pečata. V primeru XC 40 ni težko napovedati, da bo ta tradicija prekinjena. Ponuja namreč visoko kakovost končne izdelave in v kabini odlično premižljeno funkcionalnost ter na dotik prijetne materiale. že posebej hvalimo žtevilo, iznajdljivost in velikost različnih odlagalnih prostorov ter žiroke, čvrste in zelo udobne sedeže. Intuitivno in enostavno logično je upravljanje z velikim vmesnikom, ki z več funkcijskim zaslonom na dotik kraljuje na z roko lahko dostopnem mestu na sredinski armaturi. Razočaranj ne bo niti v velikosti in uporabnosti prtljažnega prostora, ki s 460 litri prostornine sicer ni med večjimi v razredu, a se v uporabniškem smislu odkupi z dobro urejenostjo ter domiselnimi režitvami pregrajevanja. XC 40 je od tal odmaknjen konkretnih 21 cm, a sta vzmetenje in krmilni mehanizem tako nastavljena, da ponuja tudi v hitro odpeljanih ovinkih zelo dolgo nevtrarno in predvidljivo lego. V premeru preskušane modela, ki je imel v paketu R design vzmetenje že nekoliko bolj trdo, se je to samo že bolj potrdilo, a je v tem primeru treba računati na manj udobno vožnjo čez različne asfaltne grbine. Podoben razmislek velja opraviti tudi pri izbiri motorja. Preskušani 2-litrski dizel s 190 KM predstavlja vrh ponudbe, ki z močjo, udobjem in tudi povprečno porabo navduži predvsem pri avtocestnih dolgoprogaških izzivih, v počasni mestni vožnji ter pri pogostih postankih in speljevanjih pa deluje preveč robusten. XC 40 je s čvrsto gradnjo, funkcionalno in udobno kabino ter žtevilnimi asistenčnimi sistemi in izstopajočim skandinavskim dizajnom v premižljenem trenutku vstopil na trg modnih mestnih terencev, v katerem se brez ene same sence dvoma suvereno postavi med najdražje in najbolj


```

premijske v mestu. - na testu Volvo XC40 2.0 TD
avt awd momentum Mere: - dolžina: 4,4 m -
medosna razdalja: 2,7 m - obračalni krog: 11,4 m
- oddaljenost od tal: 21 cm - prtljažnik: 432 l
- masa: 2.250 kg Pogon: - 2-litrski 4-valjni
bencinski motor - moč: 140 kW - navor: 400 Nm -
8-stopenjski samodejni menjalnik - pogon na vsa
štiri kolesa - pnevmatike: 235/50 R19 - poraba:
6,3 l/100 km = 8,2 EUR/100km - posoda za gorivo
: 54 l - doseg: 857 km - izpusti CO2: 133 g/km
Strožki pri 15.000 km in 5-letni uporabi: -
nakupna cena: 43.619 EUR - strožki finančnega
leasinga: 3.268 EUR/5 let - strožki registracije
: 8.701 EUR/5 let - strožki vzdrževanja: 2.320
EUR/5 let - strožki goriva: 6.190 EUR/75.000 km
- strožek 1 kompleta pnevmatik: 923 EUR -
vrednosti po 5 letih po Eurotaxu: 18.886 EUR -
strožki skupaj: 774 EUR/mesec",
6 "Author": "Miha Merljak",
7 "PublishedTime": "25. januar 2019 ob 15:23"
8 }

```

4.1.2 overstock.com

overstock.com/jewelry01.html:

```

1 [
2   {
3     "Title": "10-kt. Seven Diamond Ladies Heart Ring
4     (0.08 TW)",
5     "Content": "This ladies fashion ring dazzles with
6     hearts and diamonds. The gold band is
7     crafted into delicate, open hearts. Seven
8     brilliant-cut diamonds add a bit of sparkle.
9     ",
10    "ListPrice": "$149.00",
11    "Price": "$69.99",
12    "Saving": "$79.01",
13    "SavingPercent": "(53%)"
14  },
15  {
16    "Title": "10-Kt. Diamond Ring (.25 TW)",
17    "Content": "Nineteen round diamonds accent this 1
18    0-karat yellow gold ring with filigree
19    accents.",
20    "ListPrice": "$250.00",

```

```

14     "Price": "$74.90",
15     "Saving": "$175.10",
16     "SavingPercent": "(70%)"
17   },
18   {
19     "Title": "10-kt. Pearl and Diamond Butterfly
20       Earrings",
21     "Content": "Perfectly proportioned 5.5- to 6-mm
22       cultured pearls on 10-karat yellow gold
23       settings highlight these petite earrings. A
24       dainty rhodium-plated gold butterfly studded
25       with a diamond (0.02 total carat weight, J-K
26       color, I-2 clarity) rests atop each pearl.",
27     "ListPrice": "$149.00",
28     "Price": "$42.99",
29     "Saving": "$106.01",
30     "SavingPercent": "(71%)"
31   },
32   {
33     "Title": "14-kt. Diamond 'S' Tennis Bracelet (2.0
34       0 TW)",
35     "Content": "Invest in a swirl of light with this
36       diamond 'S' tennis bracelet. Crafted in 14-
37       karat gold, the piece features 49 diamonds
38       for two full carats. The 7.25-inch bracelet
39       closes with a pressure clasp.",
40     "ListPrice": "$1,539.99",
41     "Price": "$499.99",
42     "Saving": "$1,040.00",
43     "SavingPercent": "(67%)"
44   },
45   {
46     "Title": "10-kt. Diamond Band Fashion Ring (.11
47       TW)",
48     "Content": "Crafted in white and yellow gold,
49       this ring displays a band of seven round
50       diamonds. Order your new gold and diamond
51       fashion ring today at our low, online price."
52     ,
53     "ListPrice": "$179.99",
54     "Price": "$79.99",
55     "Saving": "$100.00",
56     "SavingPercent": "(55%)"
57   },
58   {
59     "Title": "14-kt. White Gold, Pearl and Diamond

```

```

Ring",
44   "Content": "Show your romantic side with this 14-
      karat diamond and pearl ring. Set in a domed
      band of 14-karat white gold, the ring
      features a 7-mm cultured pearl. Curved rows
      of diamonds flank the pearl.",
45   "ListPrice": "$419.99",
46   "Price": "$149.99",
47   "Saving": "$270.00",
48   "SavingPercent": "(64%)"
49 },
50 {
51   "Title": "14-kt. Gold Diamond Present Future
      Pendant (.25TW)",
52   "Content": "Designed with three large, sparkling
      diamonds to represent past, present, and
      future, this stunning pendant is set in
      gleaming 14-karat gold. It incorporates a
      total of nine diamonds (0.25 total carat
      weight, K color, I-2 to I-3 clarity). ",
53   "ListPrice": "$299.00",
54   "Price": "$149.99",
55   "Saving": "$149.01",
56   "SavingPercent": "(49%)"
57 },
58 {
59   "Title": "14-kt. Diamond Solitaire Pendant (.33
      TW)",
60   "Content": "In this simple, yet elegant pendant,
      a round brilliant diamond (0.33 total carat
      weight, H-J color, I-1 to I-2 clarity) is
      prong-set in 14-karat white gold.",
61   "ListPrice": "$1,019.99",
62   "Price": "$319.99",
63   "Saving": "$700.00",
64   "SavingPercent": "(68%)"
65 },
66 {
67   "Title": "14-kt. Diamond Solitaire Earrings (0.33
      TW)",
68   "Content": "Dazzle your way into her heart, with
      these classic diamond solitaire earrings. Two
      brilliant-cut diamonds (0.33 total carat
      weight, G-H color, I-1 to I-2 clarity) are
      set in four prongs of 14-karat white gold.",
69   "ListPrice": "$639.99",

```

```

70     "Price": "$199.99",
71     "Saving": "$440.00",
72     "SavingPercent": "(68%)"
73   },
74   {
75     "Title": "14-kt. Diamond Cross Pendant (.06 TW)",
76     "Content": "Over a cleanly sculpted Roman cross
                  of 14-karat white gold drapes a slender
                  banner containing three bright prong-set
                  round diamonds (0.06 total carat weight, H-I
                  color, I clarity).",
77     "ListPrice": "$305.00",
78     "Price": "$119.99",
79     "Saving": "$185.01",
80     "SavingPercent": "(60%)"
81   },
82   {
83     "Title": "14-kt. Diamond Solitaire Stud Earrings
              (.50 TW)",
84     "Content": "Every jewelry collection needs a
                  classic pair of diamond solitaire earrings.
                  Set in 14-karat gold, these diamond stud
                  earrings (0.50 total carat weight) have post
                  backs with butterfly clasps.",
85     "ListPrice": "$999.99",
86     "Price": "$359.99",
87     "Saving": "$640.00",
88     "SavingPercent": "(64%)"
89   },
90   {
91     "Title": "14-kt. Cultured Pearl Diamond Earrings",
92     "Content": "Create an elegant appearance with
                  these pearl and diamond stud earrings. Set in
                  14-karat yellow gold, each earring features
                  an 8 to 8.5-mm cultured white pearl. Prong-
                  set round diamonds accent the pearls. Posts
                  with butterfly clasps secure the earrings.",
93     "ListPrice": "$508.99",
94     "Price": "$179.99",
95     "Saving": "$329.00",
96     "SavingPercent": "(64%)"
97   },
98   {
99     "Title": "14-kt. Diamond 7.5-8 mm Pearl Pendant",
100    "Content": "Add a classic to your jewelry

```

```

    collection with this 14-karat gold, diamond,
    and pearl necklace. The 7.5-8 mm cultured
    white pearl creates the focal point of the
    pendant, while a diamond (0.10 TW) adds
    sparkle.",
101   "ListPrice": "$196.99",
102   "Price": "$69.99",
103   "Saving": "$127.00",
104   "SavingPercent": "(64%)"
105 },
106 {
107   "Title": "14-kt. Diamond Solitaire Earrings (.50
    TW)",
108   "Content": "This earring set has two brilliant-
    cut diamonds (0.50 total carat weight, G-H
    color, I-1 to I-2 clarity) set in four prongs
    of 14-karat white gold.",
109   "ListPrice": "$1,369.99",
110   "Price": "$409.99",
111   "Saving": "$960.00",
112   "SavingPercent": "(70%)"
113 },
114 {
115   "Title": "14-kt White Gold Diamond Band (0.50 TW)
    ",
116   "Content": "Crafted of 14-karat white gold, this
    stylish ring features a bright row of 20
    channel-set, princess-cut baguette diamonds.
    Treat her like royalty and save when you buy
    jewelry treasures at Overstock.com.",
117   "ListPrice": "$1,635.00",
118   "Price": "$609.99",
119   "Saving": "$1,025.01",
120   "SavingPercent": "(62%)"
121 }
122 ]

```

overstock.com/jewelry02.html:

```

1 [
2   {
3     "Title": "14-kt. Green Jade Hoops",
4     "Content": "Hoops of cool green jade rest between
    14-karat yellow gold endpieces. The hoops
    graduate in thickness from 3 mm at the ends
    to 6 mm in the center, with approximately 29

```

```

        mm overall diameter.",
5      "ListPrice": "$90.00",
6      "Price": "$46.99",
7      "Saving": "$43.01",
8      "SavingPercent": "47%"
9    },
10   {
11     "Title": "14-kt. Jade Doughnut Pendant",
12     "Content": "The 25-mm disk hangs delicately from
        a 14-karat gold chain. The disk features a
        dramatic gold Chinese character in the center
        , accompanied by four stylized gold bees.",
13     "ListPrice": "$150.00",
14     "Price": "$48.99",
15     "Saving": "$101.01",
16     "SavingPercent": "67%"
17   },
18   {
19     "Title": "14-kt. Charcoal Jade and Ruby Elephant
        Pendant",
20     "Content": "Carved of rich dark grey jade, this
        elephant pendant has 14-karat yellow gold
        applied to mark the feet, tusk, tail, and
        blanket. A 2-mm round faceted ruby in a gold
        bezel setting forms the eye. The pendant
        hangs from an 18-inch chain.",
21     "ListPrice": "$100.00",
22     "Price": "$28.99",
23     "Saving": "$71.01",
24     "SavingPercent": "71%"
25   },
26   {
27     "Title": "14-kt. Carved Lavender Jade Earrings",
28     "Content": "Luscious 8-mm lavender jade balls,
        carved with intricate Asian style, dangle
        from a 14-karat yellow gold French hook.",
29     "ListPrice": "$80.00",
30     "Price": "$39.99",
31     "Saving": "$40.01",
32     "SavingPercent": "50%"
33   },
34   {
35     "Title": "14-kt. Jade Cross Pendant",
36     "Content": "Green jade and gold create this
        beautiful cross pendant. Cylindrical bars of
        green jade feature caps and center of 14-

```

```

    karat yellow gold.",
37     "ListPrice": "$150.00",
38     "Price": "$49.99",
39     "Saving": "$100.01",
40     "SavingPercent": "66%"
41 },
42 {
43     "Title": "14-kt. Multicolored Jade Earrings",
44     "Content": "A delicate wrapping of 14-karat
        yellow gold wire holds six 6 x 4 pear shapes
        of jade in various shades: brilliant green,
        orange, lavender, black, pale yellow, and
        white. The post earrings have butterfly backs
        .",
45     "ListPrice": "$375.00",
46     "Price": "$99.99",
47     "Saving": "$275.01",
48     "SavingPercent": "73%"
49 },
50 {
51     "Title": "14-kt. Multicolored Jade Ring",
52     "Content": "A delicate wrapping of 14-karat
        yellow gold wire holds six 6 x 4 ovals of
        jade in various shades: brilliant green,
        orange, lavender, black, pale yellow, and
        white. A narrow gold band divides to support
        the setting.",
53     "ListPrice": "$250.00",
54     "Price": "$56.99",
55     "Saving": "$193.01",
56     "SavingPercent": "77%"
57 },
58 {
59     "Title": "14-kt. Onyx and Ruby Elephant Pendant",
60     "Content": "Carved of rich black onyx, this
        elephant pendant has 14-karat yellow gold
        applied to mark the feet, tusk, tail, and
        blanket. A 2-mm round faceted ruby in a gold
        bezel setting forms the eye. The pendant
        hangs from an 18-inch chain.",
61     "ListPrice": "$100.00",
62     "Price": "$35.99",
63     "Saving": "$64.01",
64     "SavingPercent": "64%"
65 }
66 ]

```

4.1.3 bolha.com

bolha.com/LENOVO ideapad 330 15 IKB __ bolha.com.html:

```
1 {
2   "Title": "LENOVO ideapad 330 15 IKB",
3   "Price": "400,00 €",
4   "DaysUntilExpires": "Oglas poteče čez 29 dni ",
5   "User": "stanislav67",
6   "Address": "2325 Kidričevo",
7   "Phone": "(02)7872028",
8   "Mobile": "(040)354 205",
9   "Content": "Ugodno prodam tri mesece star prenosni
    računalnik LENOVO ideapad 330, sive barve, 500 gb
    trdega diska, 4 RAMA spomina, 64 jedrni procesor,
    intel(R) core (TM) i3, naložen windows 10 home, v
    garancijski z vsemi papirji in originalni žaklo.
    Ugodno prodam za 400€ ali po dogovoru. Cena
    novega je bila 592€, zelo lepo ohranjen kot nov,
    vreden ogleda in nakupa. Inf. izključno na
    telefonsko številko: 070 749-983 MARJAN",
10  "MainImage": "./LENOVO ideapad 330 15 IKB __ bolha.
    com_files/LENOVOideapadIKB-640x640-1000.png",
11  "Images": [
12    "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
    LENOVOideapadIKB-64x48-1001.png",
13    "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
    LENOVOideapadIKB-64x48-1002.png",
14    "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
    LENOVOideapadIKB-64x48-1003.png",
15    "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
    LENOVOideapadIKB-64x48-1004.png"
16  ]
17 }
```

bolha.com/Kibla za viličar __ bolha.com.html:

```
1 {
2   "Title": "Kibla za viličar",
3   "Price": "900,00 €",
4   "DaysUntilExpires": "Oglas poteče čez 24 dni ",
5   "User": "bokica23",
6   "Address": "5271 vipava",
7   "Phone": "(040)166212",
8   "Mobile": "(040)166212",
```



```

9  "Content": "Prodaj kiblo za viličarja 040 166 212",
10 "MainImage": "./Kibla za viličar __ bolha.com_files/
    Kiblazaviliar-640x640-1000.png",
11 "Images": [
12     "./Kibla za viličar __ bolha.com_files/
        Kiblazaviliar-64x48-1001.png",
13     "./Kibla za viličar __ bolha.com_files/
        Kiblazaviliar-64x48-1002.png"
14 ]
15 }

```

4.2 Regularni izrazi

4.2.1 rtvslo.si

rtvslo.si/Audi A6 50 TDI quattro_ nemir v premijskem razredu - RTVSLO.si.html:

```

1  {
2    "Title": "Audi A6 50 TDI quattro: nemir v premijskem
        razredu",
3    "SubTitle": "Test nove generacije",
4    "Lead": "To je novi audi A6. V razred najdražjih in
        najbolj premijskih žrebcev je vnesel nemir, že
        preden je sploh zapeljal na parkirni prostor,
        rezerviran za izvržnega direktorja. ",
5    "Content": "Samo pogledajte njegovo masko € to ogromno
        satovje z radarji na takem položaju, da se ti
        na avtocesti tudi pri 120 km/h vsi spoštljivo
        umikajo, saj so prepričani, da gre za Pahorjev
        ali žarčev avto. Seveda, novi A6 lahko cesto in
        promet skenira s kar petimi radarji, petimi
        kamerami, infrardečo kamero za nočni vid,
        dvanajstimi ultrazvočnimi senzorji in laserskim
        čitalnikom € lidarjem. V glavnem vojaška
        tehnologija v službi varnosti za fante, ki smo
        radi gledali Top Gun, Bonda in druge možakarja s
        finimi igračami. Če vam pogled na Audijev spisek
        dodatne opreme ne odvzame volje do življenja,
        potem vsekakor toplo priporočamo nakup zračnega
        vzmetenja, saj dobi z njim A6 več različnih in
        vozniško zelo uporabnih karakterjev. Enako velja
        za seksi luči z inteligentno matrično
        osvetlitvijo, pa za žportno podvozje in vsekakor
        za žtirikolesno krmiljenje. S tem postane A6
        med ovinki v občutku na volanu že veliko krajši
        in bolj agilen. Vse naštetost smo preskušali v dru

```

```

6   žbi agregata 50 TDI, ki je v resnici klasični
    trilitrski dizel, podkrepljen z elektromotorjem.
    Ja, ta audi je mehki hibrid z izjemnim navorom
    in dovolj moči kadar koli in kjer koli. Si pa
    mislimo, da bo največji del trga zadovoljil že u
    činkovit dvolitrski mehki hibrid z močjo 150
    kilovatov.- na testu Audi A6 50 TDI quattro
    tiptronic- strožki skupaj: 1.001 EUR/mesec",
7   "Author": "Miha Merljak",
    "PublishedTime": "28. december 2018 ob 08:51"
8 }

```

rtvslo.si/Volvo XC 40 D4 AWD momentum- suvereno med najboljše v razredu
- RTVSLO.si.htm:

```

1 {
2   "Title": "Volvo XC 40 D4 AWD momentum: suvereno med
    najboljše v razredu",
3   "SubTitle": "Test novega modela",
4   "Lead": "XC 40 je najmanjši Volvov SUV, ki se
    oblikovno skoraj v celotni naslanja na oba večja
    predhodnika. Že samo s tem so mu vrata do
    denarnic tistih kupcev, ki iščejo izstopajočo, a
    hkrati visoko kultivirano in prečiščeno
    dizajnersko govorico, na pol odprta.",
5   "Content": "Volvo se je nižjih srednjih razredov v
    preteklosti izogibal ali pa je vanje vstopal z
    zelo niznimi produkti, ki niso pustili večjega
    tržnega pečata. V primeru XC 40 ni težko
    napovedati, da bo ta tradicija prekinjena.
    Ponuja namreč visoko kakovost končne izdelave in
    v kabini odlično premižljeno funkcionalnost ter
    na dotik prijetne materiale. Že posebej hvalimo
    žtevilo, iznajdljivost in velikost različnih
    odlagalnih prostorov ter žiroke, čvrste in zelo
    udobne sedeže. Intuitivno in enostavno logično
    je upravljanje z velikim vmesnikom, ki z več
    funkcijskim zaslonom na dotik kraljuje na z roko
    lahko dostopnem mestu na sredinski armaturi.
    Razočaranj ne bo niti v velikosti in uporabnosti
    prtljažnega prostora, ki s 460 litri
    prostornine sicer ni med večjimi v razredu, a se
    v uporabniškem smislu odkupi z dobro
    urejenostjo ter domiselnimi režitvami
    pregrajevanja. XC 40 je od tal odmaknjen
    konkretnih 21 cm, a sta vzmetenje in krmilni

```

```

    mehanizem tako nastavljena, da ponuja tudi v
    hitro odpeljanih ovinkih zelo dolgo nevtralnno in
    predvidljivo lego. V premeru preskušanega
    modela, ki je imel v paketu R design vzmetenje ž
    e nekoliko bolj trdo, se je to samo že bolj
    potrdilo, a je v tem primeru treba računati na
    manj udobno vožnjo čez različne asfaltne grbine.
    Podoben razmislek velja opraviti tudi pri
    izbiri motorja. Preskušani 2-litrski dizel s 190
    KM predstavlja vrh ponudbe, ki z močjo, udobjem
    in tudi povprečno porabo navduži predvsem pri
    avtocestnih dolgoprogažkih izzivih, v počasni
    mestni vožnji ter pri pogostih postankih in
    speljevanjih pa deluje preveč robusten. XC 40 je
    s čvrsto gradnjo, funkcionalno in udobno kabino
    ter številnimi asistenčnimi sistemi in izstopajo
    čim skandinavskim dizajnom v prečižljenem
    trenutku vstopil na trg modnih mestnih terencev,
    v katerem se brez ene same sence dvoma suvereno
    postavi med najdražje in najbolj premijske v
    mestu.",
6   "Author": "Miha Merljak",
7   "PublishedTime": "25. januar 2019 ob 15:23"
8 }

```

4.2.2 overstock.com

overstock.com/jewelry01.html:

```

1  [
2    {
3      "Title": "10-kt. Seven Diamond Ladies Heart Ring
         (0.08 TW)",
4      "Content": "This ladies fashion ring dazzles with
         hearts and diamonds. The gold band is
         crafted into delicate, open hearts. Seven
         brilliant-cut diamonds add a bit of sparkle.
         ",
5      "ListPrice": "$149.00",
6      "Price": "$69.99",
7      "Saving": "$79.01",
8      "SavingPercent": "53%"
9    },
10   {
11     "Title": "10-Kt. Diamond Ring (.25 TW)",

```

```

12     "Content": "Nineteen round diamonds accent this 1
        0-karat yellow gold ring with filigree
        accents.",
13     "ListPrice": "$250.00",
14     "Price": "$74.90",
15     "Saving": "$175.10",
16     "SavingPercent": "70%"
17 },
18 {
19     "Title": "10-kt. Pearl and Diamond Butterfly
        Earrings",
20     "Content": "Perfectly proportioned 5.5- to 6-mm
        cultured pearls on 10-karat yellow gold
        settings highlight these petite earrings. A
        dainty rhodium-plated gold butterfly studded
        with a diamond (0.02 total carat weight, J-K
        color, I-2 clarity) rests atop each pearl.",
21     "ListPrice": "$149.00",
22     "Price": "$42.99",
23     "Saving": "$106.01",
24     "SavingPercent": "71%"
25 },
26 {
27     "Title": "14-kt. Diamond 'S' Tennis Bracelet (2.0
        0 TW)",
28     "Content": "Invest in a swirl of light with this
        diamond 'S' tennis bracelet. Crafted in 14-
        karat gold, the piece features 49 diamonds
        for two full carats. The 7.25-inch bracelet
        closes with a pressure clasp.",
29     "ListPrice": "$1,539.99",
30     "Price": "$499.99",
31     "Saving": "$1,040.00",
32     "SavingPercent": "67%"
33 },
34 {
35     "Title": "10-kt. Diamond Band Fashion Ring (.11
        TW)",
36     "Content": "Crafted in white and yellow gold,
        this ring displays a band of seven round
        diamonds. Order your new gold and diamond
        fashion ring today at our low, online price."
        ,
37     "ListPrice": "$179.99",
38     "Price": "$79.99",
39     "Saving": "$100.00",

```

```

40     "SavingPercent": "55%"
41   },
42   {
43     "Title": "14-kt. White Gold, Pearl and Diamond
44       Ring",
45     "Content": "Show your romantic side with this 14-
46       karat diamond and pearl ring. Set in a domed
47       band of 14-karat white gold, the ring
48       features a 7-mm cultured pearl. Curved rows
49       of diamonds flank the pearl.",
50     "ListPrice": "$419.99",
51     "Price": "$149.99",
52     "Saving": "$270.00",
53     "SavingPercent": "64%"
54   },
55   {
56     "Title": "14-kt. Gold Diamond Present Future
57       Pendant (.25TW)",
58     "Content": "Designed with three large, sparkling
59       diamonds to represent past, present, and
60       future, this stunning pendant is set in
61       gleaming 14-karat gold. It incorporates a
62       total of nine diamonds (0.25 total carat
63       weight, K color, I-2 to I-3 clarity). ",
64     "ListPrice": "$299.00",
65     "Price": "$149.99",
66     "Saving": "$149.01",
67     "SavingPercent": "49%"
68   },
69   {
70     "Title": "14-kt. Diamond Solitaire Pendant (.33
71       TW)",
72     "Content": "In this simple, yet elegant pendant,
73       a round brilliant diamond (0.33 total carat
74       weight, H-J color, I-1 to I-2 clarity) is
75       prong-set in 14-karat white gold.",
76     "ListPrice": "$1,019.99",
77     "Price": "$319.99",
78     "Saving": "$700.00",
79     "SavingPercent": "68%"
80   },
81   {
82     "Title": "14-kt. Diamond Solitaire Earrings (0.33
83       TW)",
84     "Content": "Dazzle your way into her heart, with
85       these classic diamond solitaire earrings. Two

```

```

        brilliant-cut diamonds (0.33 total carat
        weight, G-H color, I-1 to I-2 clarity) are
        set in four prongs of 14-karat white gold.",
69     "ListPrice": "$639.99",
70     "Price": "$199.99",
71     "Saving": "$440.00",
72     "SavingPercent": "68%"
73   },
74   {
75     "Title": "14-kt. Diamond Cross Pendant (.06 TW)",
76     "Content": "Over a cleanly sculpted Roman cross
        of 14-karat white gold drapes a slender
        banner containing three bright prong-set
        round diamonds (0.06 total carat weight, H-I
        color, I clarity).",
77     "ListPrice": "$305.00",
78     "Price": "$119.99",
79     "Saving": "$185.01",
80     "SavingPercent": "60%"
81   },
82   {
83     "Title": "14-kt. Diamond Solitaire Stud Earrings
        (.50 TW)",
84     "Content": "Every jewelry collection needs a
        classic pair of diamond solitaire earrings.
        Set in 14-karat gold, these diamond stud
        earrings (0.50 total carat weight) have post
        backs with butterfly clasps.",
85     "ListPrice": "$999.99",
86     "Price": "$359.99",
87     "Saving": "$640.00",
88     "SavingPercent": "64%"
89   },
90   {
91     "Title": "14-kt. Cultured Pearl Diamond Earrings"
92     ,
93     "Content": "Create an elegant appearance with
        these pearl and diamond stud earrings. Set in
        14-karat yellow gold, each earring features
        an 8 to 8.5-mm cultured white pearl. Prong-
        set round diamonds accent the pearls. Posts
        with butterfly clasps secure the earrings.",
94     "ListPrice": "$508.99",
95     "Price": "$179.99",
96     "Saving": "$329.00",
97     "SavingPercent": "64%"

```

```

97     },
98     {
99         "Title": "14-kt. Diamond 7.5-8 mm Pearl Pendant",
100        "Content": "Add a classic to your jewelry
                    collection with this 14-karat gold, diamond,
                    and pearl necklace. The 7.5-8 mm cultured
                    white pearl creates the focal point of the
                    pendant, while a diamond (0.10 TW) adds
                    sparkle.",
101        "ListPrice": "$196.99",
102        "Price": "$69.99",
103        "Saving": "$127.00",
104        "SavingPercent": "64%"
105    },
106    {
107        "Title": "14-kt. Diamond Solitaire Earrings (.50
                    TW)",
108        "Content": "This earring set has two brilliant-
                    cut diamonds (0.50 total carat weight, G-H
                    color, I-1 to I-2 clarity) set in four prongs
                    of 14-karat white gold.",
109        "ListPrice": "$1,369.99",
110        "Price": "$409.99",
111        "Saving": "$960.00",
112        "SavingPercent": "70%"
113    },
114    {
115        "Title": "14-kt White Gold Diamond Band (0.50 TW)",
116        "Content": "Crafted of 14-karat white gold, this
                    stylish ring features a bright row of 20
                    channel-set, princess-cut baguette diamonds.
                    Treat her like royalty and save when you buy
                    jewelry treasures at Overstock.com.",
117        "ListPrice": "$1,635.00",
118        "Price": "$609.99",
119        "Saving": "$1,025.01",
120        "SavingPercent": "62%"
121    }
122 ]

```

overstock.com/jewelry02.html:

```

1 [
2   {
3     "Title": "14-kt. Green Jade Hoops",

```

```

4      "Content": "Hoops of cool green jade rest between
          14-karat yellow gold endpieces. The hoops
          graduate in thickness from 3 mm at the ends
          to 6 mm in the center, with approximately 29
          mm overall diameter.",
5      "ListPrice": "$90.00",
6      "Price": "$46.99",
7      "Saving": "$43.01",
8      "SavingPercent": "47%"
9  },
10 {
11     "Title": "14-kt. Jade Doughnut Pendant",
12     "Content": "The 25-mm disk hangs delicately from
          a 14-karat gold chain. The disk features a
          dramatic gold Chinese character in the center
          , accompanied by four stylized gold bees.",
13     "ListPrice": "$150.00",
14     "Price": "$48.99",
15     "Saving": "$101.01",
16     "SavingPercent": "67%"
17 },
18 {
19     "Title": "14-kt. Charcoal Jade and Ruby Elephant
          Pendant",
20     "Content": "Carved of rich dark grey jade, this
          elephant pendant has 14-karat yellow gold
          applied to mark the feet, tusk, tail, and
          blanket. A 2-mm round faceted ruby in a gold
          bezel setting forms the eye. The pendant
          hangs from an 18-inch chain.",
21     "ListPrice": "$100.00",
22     "Price": "$28.99",
23     "Saving": "$71.01",
24     "SavingPercent": "71%"
25 },
26 {
27     "Title": "14-kt. Carved Lavender Jade Earrings",
28     "Content": "Luscious 8-mm lavender jade balls,
          carved with intricate Asian style, dangle
          from a 14-karat yellow gold French hook.",
29     "ListPrice": "$80.00",
30     "Price": "$39.99",
31     "Saving": "$40.01",
32     "SavingPercent": "50%"
33 },
34 {

```



```

35     "Title": "14-kt. Jade Cross Pendant",
36     "Content": "Green jade and gold create this
                beautiful cross pendant. Cylindrical bars of
                green jade feature caps and center of 14-
                karat yellow gold.",
37     "ListPrice": "$150.00",
38     "Price": "$49.99",
39     "Saving": "$100.01",
40     "SavingPercent": "66%"
41 },
42 {
43     "Title": "14-kt. Multicolored Jade Earrings",
44     "Content": "A delicate wrapping of 14-karat
                yellow gold wire holds six 6 x 4 pear shapes
                of jade in various shades: brilliant green,
                orange, lavender, black, pale yellow, and
                white. The post earrings have butterfly backs
                .",
45     "ListPrice": "$375.00",
46     "Price": "$99.99",
47     "Saving": "$275.01",
48     "SavingPercent": "73%"
49 },
50 {
51     "Title": "14-kt. Multicolored Jade Ring",
52     "Content": "A delicate wrapping of 14-karat
                yellow gold wire holds six 6 x 4 ovals of
                jade in various shades: brilliant green,
                orange, lavender, black, pale yellow, and
                white. A narrow gold band divides to support
                the setting.",
53     "ListPrice": "$250.00",
54     "Price": "$56.99",
55     "Saving": "$193.01",
56     "SavingPercent": "77%"
57 },
58 {
59     "Title": "14-kt. Onyx and Ruby Elephant Pendant",
60     "Content": "Carved of rich black onyx, this
                elephant pendant has 14-karat yellow gold
                applied to mark the feet, tusk, tail, and
                blanket. A 2-mm round faceted ruby in a gold
                bezel setting forms the eye. The pendant
                hangs from an 18-inch chain.",
61     "ListPrice": "$100.00",
62     "Price": "$35.99",

```

```

63     "Saving": "$64.01",
64     "SavingPercent": "64%"
65 }
66 ]

```

4.2.3 bolha.com

bolha.com/LENOVO ideapad 330 15 IKB __ bolha.com.html:

```

1 {
2     "Title": "LENOVO ideapad 330 15 IKB",
3     "Price": "400,00 €",
4     "DaysUntilExpires": "29",
5     "User": "stanislav67",
6     "Address": "2325 Kidričevo",
7     "Phone": "(02)7872028",
8     "Mobile": "(040)354 205",
9     "Content": "
        <p>Ugodno prodam tri mesece star
        prenosni računalnik LENOVO ideapad 330, sive
        barve, 500 gb trdega diska, 4 RAMA spomina, 64
        jedrni procesor, intel(R) core (TM) i3, naložen
        windows 10 home, v garanciji z vsemi papirji in
        original žkatlo. Ugodno prodam za 400€ ali po
        dogovoru. Cena novega je bila 592€, zelo lepo
        ohranjen kot nov, vreden ogleda in nakupa. Inf.
        izključno na telefonsko žtevilko: 070 749-983
        MARJAN</p>
        ",
10    "MainImage": "./LENOVO ideapad 330 15 IKB __ bolha.
        com_files/Lenovoideapadigbgeforcegbssd-640x640-1
        000.png",
11    "Images": [
12        "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
        LENOVOideapadIKB-64x48-1001.png",
13        "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
        LENOVOideapadIKB-64x48-1002.png",
14        "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
        LENOVOideapadIKB-64x48-1003.png",
15        "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
        LENOVOideapadIKB-64x48-1004.png"
16    ]
17 }

```

bolha.com/Kibla za viličar __ bolha.com.html:

```

1 {
2     "Title": "Kibla za viličar",

```

```

3  "Price": "900,00 €",
4  "DaysUntilExpires": "24",
5  "User": "bokica23",
6  "Address": "5271 vipava",
7  "Phone": "(040)166212",
8  "Mobile": "(040)166212",
9  "Content": "          <p>Prodam kablo za viličarja<br>
          040 166 212</p>          ",
10 "MainImage": "./Kibla za viličar __ bolha.com_files/
      OBNOVA-VILICARJEV--GRADBENE-IN-KMETIJSKE-
      MEHANIZACIJE_507c6ad2f0908.jpg",
11 "Images": [
12     "./Kibla za viličar __ bolha.com_files/
        Kiblazaviliar-64x48-1001.png",
13     "./Kibla za viličar __ bolha.com_files/
        Kiblazaviliar-64x48-1002.png"
14 ]
15 }

```

4.3 Road Runner

5 Zagon algoritma

Zagon programa se lahko izvede preko ukazne vrstice z vpisom ukaza `python crawler.py <approach> <page>`, kjer je `approach` parameter, ki določa vrsto algoritma za izvoz podatkov, `page` pa stran iz kje se naj izvozijo.

6 Psevdokoda Road Runner-ja

7 Zaključek

V tej seminarski nalogi smo razvili delujoči program za pobiranje vsebin strani. Razvili smo tri različne pristope za pobiranje iste vsebine. Te tipi so XPATH, regularni izraz in algoritem Road Runner. Dobljeni rezultati se nam medsebojno ujemajo glede na različne tipe strani.

References