# Ekstrakcija podatkov iz spleta

2. domača naloga pri predmetu IEPS

Žiga Černigoj 63130028

Marko Lavrinec 63130134

Marec 2019

## 1 Uvod

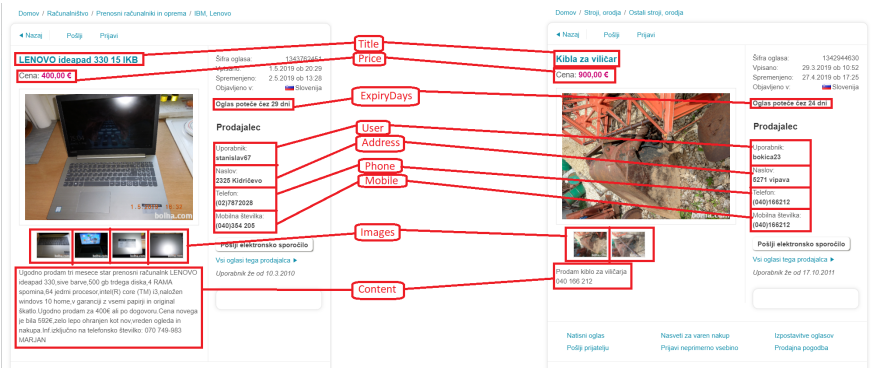
Cilj naloge je razviti program, ki s pomočjo XPATH-a, Regularnih izrazov in z Road Runner-jem izvozilo želene podatke iz spletne strani. Za implementacijo bo uporabljen programski jezik Python.

## 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.

Strani vsebujeta le en podatkovni zapis z atributi Title, Price, DaysUntilExpires, User, Address, Phone, Mobile, Content in MainImage, ki so tipa string in Images ki predstavlja seznam nizov.

Na sliki 1 vidimo, kje na strani se podatki nahajajo.




Slika 1: Zajemajoči podatki na strani bolha.com

# 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:

```
{
    'Title': '//*[@id="main-container"]//h1/text()',
    'SubTitle': '//*[@id="main-container"]//*[@class="
       subtitle"]/text()',
    'Lead': '//*[@id="main-container"]//*[@class="lead
       "]/text()',
    'Content': '//*[@id="main-container"]//*[@class="
       article-body"]/*[@class="article"]//descendant:
       :p/text()',
    'Author': '//*[@id="main-container"]//*[@class="
       author-name"]/text()',
    'PublishedTime': '//*[@id="main-container"]//*[
       @class="publish-meta"]/text()'
}
```

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 posebaj pomemebna 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 želenega cilja.

Ta pot bi bila namreč smiselnejša, saj če bi v kateri od tabel manjkal 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:

```
{
    'Title': '//table[@cellpadding="2"]/tbody/tr[
       @bgcolor]/td[2]/a/b/text()',
    'Content': '//table[@cellpadding="2"]/tbody/tr[
       @bgcolor]/td[2]/table//span[@class="normal"]/
       text()',
    'ListPrice': '//table[@cellpadding="2"]/tbody/tr[
       @bgcolor]/td[2]/table//s/text()',
    'Price': '//table[@cellpadding="2"]/tbody/tr[
       @bgcolor]/td[2]/table//span[@class="bigred"]/b/
       text()',
    'Saving': '//table[@cellpadding="2"]/tbody/tr[
       @bgcolor]/td[2]/table//span[@class="
       littleorange"]/text()'
}
```

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:

```
{
    'Title': '//*[@id="adDetail"]/div[2]/h1/text()',
    'Price': '//*[@id="adDetail"]//div[@class="price"]
       /span/text()',
    'DaysUntilExpires': '//*[@id="adDetail"]/div[3]/
       div[1]/p[5]/text()',
    'User': '//*[@id="sellerInfo"]/div/p[1]/strong/
       text()',
    'Address': '//*[@id="sellerInfo"]/div/p[2]/strong/
       text()',
    'Phone': '//*[@id="sellerInfo"]/div/p[3]/strong/
       text()',
    'Mobile': '//*[@id="sellerInfo"]/div/p[4]/strong/
       text()',
    'Content': '//*[@id="adDetail"]//div[@class="
       content"]/descendant::*/text()',
    'MainImage': '//*[@id="gallery"]/table/tbody/tr/td
       /a/img/@src',
    'Images': '//*[@id="gal"]//td[@class="thumb"]/a/
       img/@src',
}
```

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:

```
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:

```
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:

```
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:

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

### 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 želenega cilja.

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

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

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 želenega rezltata.

### 3.2.3 bolha.com

Na strani smo definirali sledeče regularne izraze:

```
{
    'Title': r"<h1[^>]*>(.*)<\/h1[^>]*>",
    'Price': '<div class="price">Cena: <span>([^<]*)</
       span></div>',
    'DaysUntilExpires': r'<p class="validTo">Oglas
       poteče čez ([^\s]*) dni </p>',
    'User': '<p><label>Uporabnik:</label><strong>([^<]
       *)</strong></p>',
    'Address': '<p><label>Naslov:</label><strong>([^<]
       *)<',
    'Phone': '<p><label>Telefon:</label><strong>([^<]
       *)<',
    'Mobile': '<p><label>Mobilna žtevilka:</label><
       strong>([^<]*)<',
    'Content': '<div class="content">(.+?(?=</div))</
       div>',
    'MainImage': '<td class="imgHolder">.*<img src="([
       ^"]*)"',
```

```
    'Images': '<td class="thumb">.+?(?=<img)<img src
        ="([^"]*)"'
}
```

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

Implementacije algoritma smo se lotili na več načinov, vendar noben ni implementiran do stopnje, ko bi za izbrane tipe strani vračal pravilen rezultat. Pri vseh načinih smo si pomagali s knjižnico BeautifulSoup, s katero smo razčlenili HTML kodo in jo predstavili z različnimi podatkovnimi strukturami.

Prvi način se nahaja v datoteki `./implementation/src/road.py`. Načrt za prvi način je bil sočasno sprehajanje po DOM drevesih obeh spletnih strani in primerjanje vozlišč na istem nivoju. Pri tem so vozlišča vsebovala le podatke o imenu HTML značke, njenih atributih ter besedilo, če ga je značka vsebovala. Za dostop do podatkov naslednikov smo tako morali napredovati po nivojih. Navdih za tak način smo dobili v magistrski nalogi Erika Schlyterja z naslovom Structured data extraction [4], kjer je v sorodnih delih opisan sprehod po drevesu. Implementacijo tega načina smo opustili zaradi zahtevnosti napredovanja po nivojih do vrednosti in potratnosti sprehajanja po vseh otrocih vseh vozlišč na istem nivoju kot trenutno izbrano vozlišče.

Drugi način se nahaja v datoteki `./implementation/src/road2.py`. Pri tem načinu smo DOM drevo samo drugače predstavili - s tabelo. Vrhnje HTML elemente smo dali v tabelo. Vsak element tabele je tako vseboval celoten HTML element - objekt, ki je vseboval celoten HTML element, shranjen kot niz znakov, HTML značko in njene atribute ter tabelo naslednikov. Tako smo lahko neposredno na najvišjem nivoju primerjali večje dele HTML kode. Vendar se je zopet pokazala težavnost pri napredovanju na nadaljnje nivoje, tako da smo tudi ta način opustili.

Tretji način se nahaja v datoteki `./implementation/src/road3.py`. Pri tem načinu smo sledili opisu implementacije s prosojnic, avtorjev algoritma RoadRunner [3, 2, 1], vendar nam je zmanjkalo časa, da bi implementirali zahtevnejše procesiranje bolj kompleksne HTML kode.

## 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:

```
{
```

```
  "Title":"Audi A6 50 TDI quattro: nemir v premijskem razredu",
  "SubTitle":"Test nove generacije",
  "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. ",
  "Content":"Samo poglejte 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 vojažka tehnologija v službi varnosti za fante, ki smo radi gledali Top Gun, Bonda in druge možakarja s finimi igračami. Voznički delovni prostor je novo poglavje digitalne dobe, z dvema ogromnima zaslonoma, ki tako kot naprednejži telefoni dregnejo blazinice važih 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 vozničko 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 krajži in bolj agilen. Vse nažteto smo preskužali v družbi 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 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 štiri 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 najboljse v razredu - RTVSLO.si.htm:


```
{
   "Title":"Volvo XC 40 D4 AWD momentum: suvereno med
      najboljže v razredu",
   "SubTitle":"Test novega modela",
   "Lead":"XC 40 je najmanjži 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 nižnimi 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 premišljeno funkcionalnost ter
       na dotik prijetne materiale.  že posebej
      hvalimo žtevilo, iznajdljivost in velikost razli
      čnih odlagalnih prostorov ter široke, čvrste in
      zelo udobne sedeže. Intuitivno in enostavno logi
      čno 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 uporabnižkem 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 nevtralno in
 predvidljivo lego. V premeru preskužanega
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. Preskužani 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 žtevilnimi asistenčnimi sistemi in
izstopajočim skandinavskim dizajnom v premiž
ljenem 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
žtiri 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",
```

```
    "Author":"Miha Merljak",
    "PublishedTime":"   25. januar 2019 ob 15:23"
}
```

#### 4.1.2 overstock.com

overstock.com/jewelry01.html:

```
[
    {
        "Title":"10-kt. Seven Diamond Ladies Heart Ring
            (0.08 TW)",
        "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.
            ",
        "ListPrice":"$149.00",
        "Price":"$69.99",
        "Saving":"$79.01",
        "SavingPercent":"(53%)"
    },
    {
        "Title":"10-Kt. Diamond Ring (.25 TW)",
        "Content":"Nineteen round diamonds accent this 1
            0-karat yellow gold ring with filigree
            accents.",
        "ListPrice":"$250.00",
        "Price":"$74.90",
        "Saving":"$175.10",
        "SavingPercent":"(70%)"
    },
    {
        "Title":"10-kt. Pearl and Diamond Butterfly
            Earrings",
        "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.",
        "ListPrice":"$149.00",
        "Price":"$42.99",
        "Saving":"$106.01",
        "SavingPercent":"(71%)"
    },
```

```
    {
        "Title":"14-kt. Diamond 'S' Tennis Bracelet (2.00 TW)",
        "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.",
        "ListPrice":"$1,539.99",
        "Price":"$499.99",
        "Saving":"$1,040.00",
        "SavingPercent":"(67%)"
    },
    {
        "Title":"10-kt. Diamond Band Fashion Ring (.11 TW)",
        "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.",
        "ListPrice":"$179.99",
        "Price":"$79.99",
        "Saving":"$100.00",
        "SavingPercent":"(55%)"
    },
    {
        "Title":"14-kt. White Gold, Pearl and Diamond Ring",
        "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.",
        "ListPrice":"$419.99",
        "Price":"$149.99",
        "Saving":"$270.00",
        "SavingPercent":"(64%)"
    },
    {
        "Title":"14-kt. Gold Diamond Present Future Pendant (.25TW)",
        "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).  ",
    "ListPrice":"$299.00",
    "Price":"$149.99",
    "Saving":"$149.01",
    "SavingPercent":"(49%)"
  },
  {
    "Title":"14-kt. Diamond Solitaire Pendant (.33
        TW)",
    "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.",
    "ListPrice":"$1,019.99",
    "Price":"$319.99",
    "Saving":"$700.00",
    "SavingPercent":"(68%)"
  },
  {
    "Title":"14-kt. Diamond Solitaire Earrings (0.33
        TW)",
    "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.",
    "ListPrice":"$639.99",
    "Price":"$199.99",
    "Saving":"$440.00",
    "SavingPercent":"(68%)"
  },
  {
    "Title":"14-kt. Diamond Cross Pendant (.06 TW)",
    "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).",
    "ListPrice":"$305.00",
    "Price":"$119.99",
    "Saving":"$185.01",
    "SavingPercent":"(60%)"
  },
  {
    "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",
```

```
        "Price":"$409.99",
        "Saving":"$960.00",
        "SavingPercent":"(70%)"
    },
    {
        "Title":"14-kt White Gold Diamond Band (0.50 TW)
            ",
        "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.",
        "ListPrice":"$1,635.00",
        "Price":"$609.99",
        "Saving":"$1,025.01",
        "SavingPercent":"(62%)"
    }
]
```

overstock.com/jewelry02.html:

```
[
    {
        "Title":"14-kt. Green Jade Hoops",
        "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.",
        "ListPrice":"$90.00",
        "Price":"$46.99",
        "Saving":"$43.01",
        "SavingPercent":"47%"
    },
    {
        "Title":"14-kt. Jade Doughnut Pendant",
        "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.",
        "ListPrice":"$150.00",
        "Price":"$48.99",
        "Saving":"$101.01",
        "SavingPercent":"67%"
    },
    {
```

```
        "Title":"14-kt. Charcoal Jade and Ruby Elephant Pendant",
        "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.",
        "ListPrice":"$100.00",
        "Price":"$28.99",
        "Saving":"$71.01",
        "SavingPercent":"71%"
    },
    {
        "Title":"14-kt. Carved Lavender Jade Earrings",
        "Content":"Luscious 8-mm lavender jade balls, carved with intricate Asian style, dangle from a 14-karat yellow gold French hook.",
        "ListPrice":"$80.00",
        "Price":"$39.99",
        "Saving":"$40.01",
        "SavingPercent":"50%"
    },
    {
        "Title":"14-kt. Jade Cross Pendant",
        "Content":"Green jade and gold create this beautiful cross pendant. Cylindrical bars of green jade feature caps and center of 14-karat yellow gold.",
        "ListPrice":"$150.00",
        "Price":"$49.99",
        "Saving":"$100.01",
        "SavingPercent":"66%"
    },
    {
        "Title":"14-kt. Multicolored Jade Earrings",
        "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.",
        "ListPrice":"$375.00",
        "Price":"$99.99",
        "Saving":"$275.01",
        "SavingPercent":"73%"
```

```
    },
    {
        "Title":"14-kt. Multicolored Jade Ring",
        "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.",
        "ListPrice":"$250.00",
        "Price":"$56.99",
        "Saving":"$193.01",
        "SavingPercent":"77%"
    },
    {
        "Title":"14-kt. Onyx and Ruby Elephant Pendant",
        "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.",
        "ListPrice":"$100.00",
        "Price":"$35.99",
        "Saving":"$64.01",
        "SavingPercent":"64%"
    }
]
```

#### 4.1.3 bolha.com

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

```
{
   "Title":"LENOVO ideapad 330 15 IKB",
   "Price":"400,00 €",
   "DaysUntilExpires":"Oglas poteče čez 29 dni ",
   "User":"stanislav67",
   "Address":"2325 Kidričevo",
   "Phone":"(02)7872028",
   "Mobile":"(040)354 205",
   "Content":"Ugodno prodam tri mesece star prenosni
      računalnk LENOVO ideapad 330,sive barve,500 gb
      trdega diska,4 RAMA spomina,64 jedrni procesor,
      intel(R) core (TM) i3,naložen windovs 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",
   "MainImage":"./LENOVO ideapad 330 15 IKB __ bolha.
      com_files/LENOVOidejapadIKB-640x640-1000.png",
   "Images":[
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1001.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1002.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1003.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1004.png"
   ]
}
```


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


```
{
   "Title":"Kibla za viličar",
   "Price":"900,00 €",
   "DaysUntilExpires":"Oglas poteče čez 24 dni ",
   "User":"bokica23",
   "Address":"5271 vipava",
   "Phone":"(040)166212",
   "Mobile":"(040)166212",
   "Content":"Prodam kiblo za viličarja  040 166 212",
   "MainImage":"./Kibla za viličar __ bolha.com_files/
      Kiblazaviliar-640x640-1000.png",
   "Images":[
      "./Kibla za viličar __ bolha.com_files/
         Kiblazaviliar-64x48-1001.png",
      "./Kibla za viličar __ bolha.com_files/
         Kiblazaviliar-64x48-1002.png"
   ]
}
```


## 4.2 Regularni izrazi

### 4.2.1 rtvslo.si

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


```
{
```

```
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 poglejte 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 vojažka
        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
        vozniżko 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 krajži
        in bolj agilen. Vse nažteto smo preskužali v dru
        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",
6    "Author":"Miha Merljak",
7    "PublishedTime":"   28. december 2018 ob 08:51"
8 }
```


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


```
1 {
2    "Title":"Volvo XC 40 D4 AWD momentum: suvereno med
        najboljže 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 nižnimi 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 premišljeno funkcionalnost ter
          na dotik prijetne materiale. že posebej hvalimo
          število, iznajdljivost in velikost različnih
         odlagalnih prostorov ter široke, 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 uporabnižkem 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 nevtralno in
          predvidljivo lego. V premeru preskužanega
         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.Preskužani 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 premišljenem
         trenutku vstopil na trg modnih mestnih terencev,
```

```
        v katerem se brez ene same sence dvoma suvereno
        postavi med najdražje in najbolj premijske v
       mestu.",
   "Author":"Miha Merljak",
   "PublishedTime":"   25. januar 2019 ob 15:23"
}
```

#### 4.2.2 overstock.com

overstock.com/jewelry01.html:

```
[
   {
      "Title":"10-kt. Seven Diamond Ladies Heart Ring
         (0.08 TW)",
      "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.
         ",
      "ListPrice":"$149.00",
      "Price":"$69.99",
      "Saving":"$79.01",
      "SavingPercent":"53%"
   },
   {
      "Title":"10-Kt. Diamond Ring (.25 TW)",
      "Content":"Nineteen round diamonds accent this 1
         0-karat yellow gold ring with filigree
         accents.",
      "ListPrice":"$250.00",
      "Price":"$74.90",
      "Saving":"$175.10",
      "SavingPercent":"70%"
   },
   {
      "Title":"10-kt. Pearl and Diamond Butterfly
         Earrings",
      "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.",
      "ListPrice":"$149.00",
      "Price":"$42.99",
```

```
        "Saving":"$106.01",
        "SavingPercent":"71%"
    },
    {
        "Title":"14-kt. Diamond 'S' Tennis Bracelet (2.00 TW)",
        "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.",
        "ListPrice":"$1,539.99",
        "Price":"$499.99",
        "Saving":"$1,040.00",
        "SavingPercent":"67%"
    },
    {
        "Title":"10-kt. Diamond Band Fashion Ring (.11 TW)",
        "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.",
        "ListPrice":"$179.99",
        "Price":"$79.99",
        "Saving":"$100.00",
        "SavingPercent":"55%"
    },
    {
        "Title":"14-kt. White Gold, Pearl and Diamond Ring",
        "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.",
        "ListPrice":"$419.99",
        "Price":"$149.99",
        "Saving":"$270.00",
        "SavingPercent":"64%"
    },
    {
        "Title":"14-kt. Gold Diamond Present Future Pendant (.25TW)",
        "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).  ",
        "ListPrice":"$299.00",
        "Price":"$149.99",
        "Saving":"$149.01",
        "SavingPercent":"49%"
    },
    {
        "Title":"14-kt. Diamond Solitaire Pendant (.33
            TW)",
        "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.",
        "ListPrice":"$1,019.99",
        "Price":"$319.99",
        "Saving":"$700.00",
        "SavingPercent":"68%"
    },
    {
        "Title":"14-kt. Diamond Solitaire Earrings (0.33
             TW)",
        "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.",
        "ListPrice":"$639.99",
        "Price":"$199.99",
        "Saving":"$440.00",
        "SavingPercent":"68%"
    },
    {
        "Title":"14-kt. Diamond Cross Pendant (.06 TW)",
        "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).",
        "ListPrice":"$305.00",
        "Price":"$119.99",
        "Saving":"$185.01",
        "SavingPercent":"60%"
```

```
    },
    {
        "Title":"14-kt. Diamond Solitaire Stud Earrings (.50 TW)",
        "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.",
        "ListPrice":"$999.99",
        "Price":"$359.99",
        "Saving":"$640.00",
        "SavingPercent":"64%"
    },
    {
        "Title":"14-kt. Cultured Pearl Diamond Earrings",
        "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.",
        "ListPrice":"$508.99",
        "Price":"$179.99",
        "Saving":"$329.00",
        "SavingPercent":"64%"
    },
    {
        "Title":"14-kt. Diamond 7.5-8 mm Pearl Pendant",
        "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.",
        "ListPrice":"$196.99",
        "Price":"$69.99",
        "Saving":"$127.00",
        "SavingPercent":"64%"
    },
    {
        "Title":"14-kt. Diamond Solitaire Earrings (.50 TW)",
        "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.",
      "ListPrice":"$1,369.99",
      "Price":"$409.99",
      "Saving":"$960.00",
      "SavingPercent":"70%"
   },
   {
      "Title":"14-kt White Gold Diamond Band (0.50 TW)
         ",
      "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.",
      "ListPrice":"$1,635.00",
      "Price":"$609.99",
      "Saving":"$1,025.01",
      "SavingPercent":"62%"
   }
]
```


overstock.com/jewelry02.html:


```
[
   {
      "Title":"14-kt. Green Jade Hoops",
      "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.",
      "ListPrice":"$90.00",
      "Price":"$46.99",
      "Saving":"$43.01",
      "SavingPercent":"47%"
   },
   {
      "Title":"14-kt. Jade Doughnut Pendant",
      "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.",
      "ListPrice":"$150.00",
      "Price":"$48.99",
      "Saving":"$101.01",
```

```
        "SavingPercent":"67%"
    },
    {
        "Title":"14-kt. Charcoal Jade and Ruby Elephant
            Pendant",
        "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.",
        "ListPrice":"$100.00",
        "Price":"$28.99",
        "Saving":"$71.01",
        "SavingPercent":"71%"
    },
    {
        "Title":"14-kt. Carved Lavender Jade Earrings",
        "Content":"Luscious 8-mm lavender jade balls,
            carved with intricate Asian style, dangle
            from a 14-karat yellow gold French hook.",
        "ListPrice":"$80.00",
        "Price":"$39.99",
        "Saving":"$40.01",
        "SavingPercent":"50%"
    },
    {
        "Title":"14-kt. Jade Cross Pendant",
        "Content":"Green jade and gold create this
            beautiful cross pendant. Cylindrical bars of
            green jade feature caps and center of 14-
            karat yellow gold.",
        "ListPrice":"$150.00",
        "Price":"$49.99",
        "Saving":"$100.01",
        "SavingPercent":"66%"
    },
    {
        "Title":"14-kt. Multicolored Jade Earrings",
        "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
            .",
        "ListPrice":"$375.00",
```

```
        "Price":"$99.99",
        "Saving":"$275.01",
        "SavingPercent":"73%"
    },
    {
        "Title":"14-kt. Multicolored Jade Ring",
        "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.",
        "ListPrice":"$250.00",
        "Price":"$56.99",
        "Saving":"$193.01",
        "SavingPercent":"77%"
    },
    {
        "Title":"14-kt. Onyx and Ruby Elephant Pendant",
        "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.",
        "ListPrice":"$100.00",
        "Price":"$35.99",
        "Saving":"$64.01",
        "SavingPercent":"64%"
    }
]
```


#### 4.2.3 bolha.com

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


```
{
   "Title":"LENOVO ideapad 330 15 IKB",
   "Price":"400,00 €",
   "DaysUntilExpires":"29",
   "User":"stanislav67",
   "Address":"2325 Kidričevo",
   "Phone":"(02)7872028",
   "Mobile":"(040)354 205",
   "Content":"        <p>Ugodno prodam tri mesece star
       prenosni računalnk LENOVO ideapad 330,sive
```

```
      barve,500 gb trdega diska,4 RAMA spomina,64
      jedrni procesor,intel(R) core (TM) i3,naložen
      windovs 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>    ",
   "MainImage":"./LENOVO ideapad 330 15 IKB __ bolha.
      com_files/Lenovoideapadigbgeforcegbssd-640x640-1
      000.png",
   "Images":[
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1001.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1002.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1003.png",
      "./LENOVO ideapad 330 15 IKB __ bolha.com_files/
         LENOVOidejapadIKB-64x48-1004.png"
   ]
}
```


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


```
{
   "Title":"Kibla za viličar",
   "Price":"900,00 €",
   "DaysUntilExpires":"24",
   "User":"bokica23",
   "Address":"5271 vipava",
   "Phone":"(040)166212",
   "Mobile":"(040)166212",
   "Content":"        <p>Prodam kiblo za viličarja<br>
       040 166 212</p>    ",
   "MainImage":"./Kibla za viličar __ bolha.com_files/
      OBNOVA-VILICARJEV--GRADBENE-IN-KMETIJSKE-
      MEHANIZACIJE_507c6ad2f0908.jpg",
   "Images":[
      "./Kibla za viličar __ bolha.com_files/
         Kiblazaviliar-64x48-1001.png",
      "./Kibla za viličar __ bolha.com_files/
         Kiblazaviliar-64x48-1002.png"
   ]
}
```

### 4.3 Road Runner

#### 4.3.1 rtvslo.si

Izpis za spletni strani rtvslo.si se nahaja v datoteki `./output/rtvslo-si.txt`.

#### 4.3.2 overstock.com

Izpis za spletni strani overstock.com se nahaja v datoteki `./output/overstock-com.txt`.

#### 4.3.3 bolha.com

Izpis za spletni strani bolha.com se nahaja v datoteki `./output/bolha-com.txt`.

## 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

Potek algoritma sledi opisu prej omenjenega algoritma RoadRunner [3, 2, 1] z nekaj dodatnimi operacijami.

## 7 Zaključek

V tej seminarski nalogi smo razvili delujoči program za ekstrakcijo vsebin spletnih strani. Razvili smo tri različne pristope, in sicer XPATH, regularni izraz in algoritem RoadRunner. Dobljeni rezultati za XPATH in regularni izraz se nam medsebojno ujemajo glede na različne tipe strani. Algoritem RoadRunner žal ni v celoti implementiran.

## References


[1] Valter Crescenzi and Giansalvatore Mecca. Automatic information extraction from large websites. *Journal of the ACM (JACM)*, 51(5):731–779, 2004.

[2] Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Automatic web information extraction in the roadrunner system. In *International Conference on Conceptual Modeling*, pages 264–277. Springer, 2001.

[3] Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo, et al. Roadrunner: Towards automatic data extraction from large web sites. In *VLDB*, volume 1, pages 109–118, 2001.

```
Data: HTML code of both sites
Result: regular expression
parsing of code with BeautifulSoup, code formatting, tag attributes
 removal;
separation of tags and data strings into own elements in arrays;
initialization of arrays for storing repeating sections and optional sections;
while not at end of both arrays do
    compare content of current elements;
    if string mismatch then
        mark data as "interesting"
    end
    else if tag mismatch then
        find section that represents a whole element and determine in
         which array;
        check if section is repeating;
        if repeating then
            save section in array for repeating sections;
            if section not in array that represents a wrapper then
                add section to array that represents a wrapper;
            end
        end
    end
    else
        check if section is optional;
        if optional then
            save section in array for optional sections
            if section not in array that represents a wrapper then
                add section to array that represents a wrapper;
            end
        end
    end
end
join array that represents a wrapper into one string;
forall sections in array of repeating sections do
    index = index of first occurrence of section in wrapper string;
    remove all occurrences of a section in wrapper string;
    add "(section)+" to the wrapper string on a wrapper string at index;
end
forall sections in array of optional sections do
    index = index of first occurrence of section in wrapper string;
    remove occurrence of a section in wrapper string;
    add "(section)?" to the wrapper string on a wrapper string at index;
end
return wrapper string;
```

**Algorithm 1:** Pseudocode of RoadRunner

[4] Erik Schlyter. Structured data extraction. 2007.