

# Naloga 2: Ekstrakcija podatkov iz spleta

Blaž Marolt, Rok Šolar, Anže Veršnik

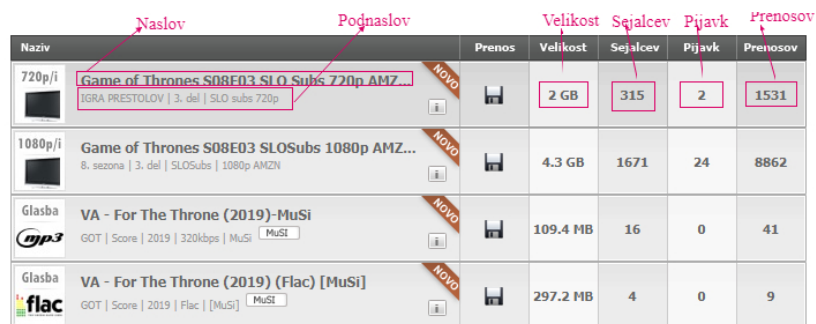
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

## 1 Uvod

Druga domača naloga je zahtevala implementacijo treh različnih pristopov za ekstrakcijo strukturiranih podatkov iz spletnih strani. Implementirali smo ekstrakcijo s uporabo regularnih izrazov, poizvedb XPath ter z algoritmom, osnovanim na ideji algoritma RoadRunner [1]. Na voljo smo imeli po dve strani iz domen `overstock.com` in `rtvslo.si` ter dve strani iz domene po svoji izbiri. Za ekstrakcijo z regularnimi izrazi ter poizvedbami XPath smo napisali funkcije za ekstrakcijo za vsako domeno posebej, algoritem RoadRunner pa z ustrezno pred obdelavo deluje na nekaterih skupinah podobnih si strani. V tem delu bomo najprej predstavili izbrani spletni strani, nato pa še naše implementacije vseh treh pristopov ter rezultate.

## 2 Izbrani spletni strani

Izbrali smo dve spletni strani iz domene `partis.si`, ki predstavljata rezultate iskanja. Vsebujeta seznam istih objektov, z različnimi vrednostmi atributov. Seznama se tudi razlikujeta po dolžini. Iz teh objektov smo ekstrahirali naslov, podnaslov, velikost datoteke, število sejalcov, število pijavk ter število prenosov (slika 1).



The screenshot shows a table of search results on the website Partis.si. The table has columns for 'Naziv', 'Prenos', 'Velikost', 'Sejalcev', 'Pijavk', and 'Prenosov'. The first row is for 'Game of Thrones S08E03 SLO Subs 720p AMZ...', and the second row is for 'Game of Thrones S08E03 SLO Subs 1080p AMZN...'. The third and fourth rows are for 'VA - For The Throne (2019)-MuSi' and 'VA - For The Throne (2019) (Flac) [MuSi]' respectively. Annotations with arrows point to specific data points: 'Naslov' points to the title, 'Podnaslov' points to the subtitle, 'Velikost' points to the file size, 'Sejalcev' points to the number of seeders, 'Pijavk' points to the number of leechers, and 'Prenosov' points to the number of downloads.

Naziv	Prenos	Velikost	Sejalcev	Pijavk	Prenosov
720p/i Game of Thrones S08E03 SLO Subs 720p AMZ... IGRA PRESTOLOV   3. del   SLO subs 720p		2 GB	315	2	1531
1080p/i Game of Thrones S08E03 SLO Subs 1080p AMZN... 8. sezona   3. del   SLO Subs   1080p AMZN		4.3 GB	1671	24	8862
Glasba VA - For The Throne (2019)-MuSi GOT   Score   2019   320kbps   MuSi   MuSi		109.4 MB	16	0	41
Glasba VA - For The Throne (2019) (Flac) [MuSi] GOT   Score   2019   Flac   [MuSi]   MuSi		297.2 MB	4	0	9

Slika 1. Podatki na spletni strani Partis.si

### 3 Implementacija

#### 3.1 Regularni izrazi

HTML kodo za spletne strani iz domene **partis.si** in domene **rtvslo.si** preberemo s kodiranjem UTF-8, kar nam omogoča pravilen izpis šumnikov. Vsem trem funkcijam v obliki parametra podamo pridobljeno HTML kodo, iz katere želimo ekstrahirati podatke. V večini primerov je nato za vsak potreben podatek napisan svoj regularni izraz za ekstrakcijo. V nadaljevanju bodo prikazani uporabljeni regularni izrazi.

##### overstock.com:

Title:

```
<td valign="top">[\s\S]*?<a href="/\"(.*)\"><b>(.)</b></a>
```

Content (razdeljeno v dve vrstici zaradi dolžine):

```
<td valign="top">[\s\S]*?<span class="normal\">([\s\S]*)<br>
```

```
<a href=\".*><span class="tiny\"><b>(.)</b>
```

ListPrice:

```
<s>(.)</s>
```

Price:

```
<span class="bigred\"><b>(.)</b></span>
```

Saving & SavingPercent:

```
<span class="littleorange\">(\$.*) \((.*?d{1,3}%)</span>
```

##### rtvslo.si:

Author & Time:

```
<div class="author-timestamp\">\s*<strong>(.)\s*</strong>\\s*(\w*. \w* \w* \w* \w*:\w*)
```

Title:

```
<h1>(.)</h1>
```

Subtitle:

```
<div class="subtitle\">(.)</div>
```

Lead:

```
<p class="lead\">(.)</p>
```

Content (razdeljeno v dve vrstici zaradi dolžine):

```
(?=(?:<\figure><p>|</p><p( class=\"Body\")?>)(?!<iframe>)(.+)?</p>)|
```

```
<span class=\"icon-photo\"></span>(.*?)</figcaption>
```

**partis.si:**

Naslov:

```
<div class=\"listeklink\">[\s\S]*?<a href=\".*\">(.*?)</a>
```

Podnaslov:

```
<div class=\"liopisl\">(.*?)(<img src=\".*\">)?</div>
```

Velikost, sejalcev, pijavk & prenosov:

```
<div class=\"datat\">(.*?)</div>
```

Kjer je potrebno z uporabo knjižnice `BeautifulSoup` odstranimo vse odvečne HTML značke (`br`, `strong`, `ipd.`), izločimo še vse znake za novo vrstico ali tabulator (`\t`, `\r` in `\n` zamenjamo s presledkom), ter izpišemo rezultate v JSON formatu.

### 3.2 XPath

HTML kodo za spletne strani iz domene `partis.si` in domene `rtvslo.si` preberemo s kodiranjem UTF-8, kar nam omogoča pravilen izpis šumnikov. Vsem trem funkcijam v obliki parametra podamo pridobljeno HTML kodo, iz katere želimo ekstrahirati podatke. V večini primerov je nato za vsak potreben podatek napisan svoj XPath izraz za ekstrakcijo. V nadaljevanju bodo prikazani uporabljeni XPath izrazi.

**overstock.com:**

Title:

```
./table[@cellpadding="2"]/tbody/tr/td[2]/a/b/text()
```

ListPrice:

```
./table[@cellpadding="2"]/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s/text()
```

Price:

```
./table[@cellpadding="2"]/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b/text()
```

Saving in saving percent:

```
./table[@cellpadding="2"]/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text()
```

Poleg XPath uporabljen še regularni izraz za ločitev podatkov saving in saving percent.

```
(\$.*?) \((.*?\d{1,3}%\)\\
```

Content:

```
./table[@cellpadding="2"]/tbody/tr/td[2]/table/tbody/tr/td[2]/span/text()
```

Poleg XPath uporabljen še regularni izraz za izločanje posebnih znakov in ločil.

```
'(.*?)' | \"(.*?)\" \\
```

**rtvslo.si:**

Author:

```
//div[@class="author-timestamp"]/strong/text()
```

Time:

```
//div[@class="author-timestamp"]/text()[2]
```

Title:

```
//header[@class="article-header"]/h1/text()
```

Subtitle:

```
//div[@class="subtitle"]/text()
```

Lead:

```
//p[@class="lead"]/text()
```

Content:

```
./article[@class="article"]/p//text()
```

**partis.si:**

Naslov:

```
./div[@class="list"]/div/div[2]/a/text()
```

Podnaslov:

```
./div[@class="list"]/div/div[2]/div/div[1]/text()
```

Velikost:

```
./div[@class="list"]/div/div[4]/text()
```

Sejalcev:

```
./div[@class="list"]/div/div[5]/text()
```

Pijavk:

```
./div[@class="list"]/div/div[6]/text()
```

Prenosov:

```
./div[@class="list"]/div/div[7]/text()
```

Za odstranjevanje posebnih ločil in znakov, kot so tabulator, nova vrstica, ", ', , ", |, ... smo napisali funkcijo, ki nam vse te znake nadomesti. Poleg odstranjevanja posebnih znakov je bilo potrebno pri spletnih straneh iz partis.si domene deliti rezultate po znaku „.

### 3.3 RoadRunner

Pri algoritmu RoadRunner smo se oprli na članek [1]. Poskušali smo slediti napotkom in ga po potrebi prilagajali za naš problem, vendar nam je zmanjkalo časa za podrobnejšo implementacijo.

Algoritem na začetku prebere obe strani ter prvo uporabi kot začetno ovojnico. Obe strani pretvori v format XHTML, odstrani vse odvečne znake, skriptne elemente, komentarje ter uporabi le elemente znotraj značk `<body>`. Nato HTML kodo razčlenimo na žetone (ang. tokens), za kar smo napisali ustrezen razred (*Element*). Vsak žeton predstavlja začetno značko, končno značko ali pa vrednost med značkami, vsi žetoni pa so zapisani v seznamu. Nato se sprehajamo po seznamu žetonov in izvajamo algoritem, kot je podan v psevdokodi spodaj. Če pridemo do neusklajenosti med dvema HTML značkama poskusimo najprej iskanje iteratorjev in če to ne uspe poskusimo z iskanjem opcijskih elementov. Pri tem upoštevamo delovanje, kot je opisano v članku. Rekurzivno iskanje opcijskih elementov znotraj iteratorjev smo uspeli implementirati do te mere, da je delovalo na primerih, podanih v članku [1], na naših primerih pa ni, zato smo ta del izločili iz končne rešitve.

### 3.4 RoadRunner psevdokoda

```

Data: wrapper = prva stran, sample = druga stran
Result: wrapper
to_xhtml(wrapper, sample) ;
normalize_html(wrapper, sample) ;
generate_token(wrapper, sample);
wrapper_index = 0 ;
sample_index = 0;
while (wrapper_index) < max(length(wrapper), length(sample)) do
    if match(wrapper[i], sample[i]) both_string() then
        if match(wrapper[i], sample[i]) both_tag() then
            wrapper[wrapper_index] = wrapper[wrapper_index];
            wrapper_index++;
            sample_index++;
        else
            wrapper[wrapper_index] = #PCDATA;
            wrapper_index++;
            sample_index++;
        end
    else
        if find_iterators(wrapper, sample)[1] == False then
            [wrapper, wrapper_index, sample_index] =
                find_options(wrapper, sample, wrapper_index, sample_index);
        end
    end
end

```

**Algorithm 1:** Glavna zanka algoritma RoadRunner

**Data:** wrapper = prva stran, sample = druga stran  
**Result:** wrapper, isSuccessful  
get\_terminal\_tags();  
**if** *not terminal tags match* **then**  
| return [wrapper, false];  
**end**  
find\_next\_terminal\_tag\_accurence\_sample();  
do\_backtrack\_Sample();  
**if** *backtracking successful* **then**  
| find\_square\_repetitions\_in\_wrapper();  
| wrapper = generalize\_wrapper();  
| return [wrapper, true];  
**else**  
| find\_next\_terminal\_tag\_accurence\_wrapper();  
| do\_backtrack\_wrapper();  
| **if** *backtracking successful* **then**  
| | find\_square\_repetitions\_in\_wrapper();  
| | wrapper = generalize\_wrapper();  
| | return [wrapper, true];  
| **end**  
**end**  
return [wrapper, false];

**Algorithm 2:** Iskanje iteratorjev

**Data:** wrapper=prva stran,sample = druga stran,wrapper\_index,sample\_index  
**Result:** Označi vse opsijske elemente  
sample\_tag = sample[sample\_index];  
wrapper\_tag = wrapper[wrapper\_index];  
index\_in\_wrapper = find(sample\_tag,wrapper);  
**if** (*index\_in\_wrapper != -1*) **then**  
| wrapper[wrapper\_count].setOptional();  
| wrapper\_count++;  
**else**  
| index\_in\_sample = find(wrapper\_tag,sample);  
| temp = sample[index\_in\_sample].setOptional();  
| wrapper[wrapper\_count] = temp;  
| wrapper\_count++;  
| sample\_count++;  
**end**

**Algorithm 3:** Iskanje opsijskih elementov

## 4 Izhodi

### 4.1 Regularni izrazi

overstock.com 1:

```
[
  {
    "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. Click here to purchase.",
    "ListPrice": "$149.00",
    "Price": "$69.99",
    "Saving": "$79.01",
    "SavingPercent": "53%"
  },
  {
    "Title": "10-Kt. Diamond Ring (.25 TW)",
    "Content": "Nineteen round diamonds accent this 10-karat yellow gold ring with filigree accents. Click here to purchase.",
    "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. Click here to purchase.",
    "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. Click here to purchase.",
    "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. Click here to purchase.",
      "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. Click here to purchase.",
      "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). Click here to purchase.",
      "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. Click here to purchase.",
      "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.
    Click here to purchase.",
    "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). Click here to purchase.",
    "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. Click here to purchase.",
    "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. Click here to purchase.",
    "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.
    Click here to purchase.",

```

```

    "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. Click here to purchase.",
    "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. Click here to purchase.",
    "ListPrice": "$1,635.00",
    "Price": "$609.99",
    "Saving": "$1,025.01",
    "SavingPercent": "62%"
  }
]

```

overstock.com 2:

```

[
  {
    "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. Click here to purchase.",
    "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. Click here to purchase.",
  }
]

```

```

        "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. Click here to purchase.",
        "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. Click here to purchase.",
        "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. Click here to purchase.",
        "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. Click here to purchase.",
        "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. Click here to purchase.",
      "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. Click here to purchase.",
      "ListPrice": "$100.00",
      "Price": "$35.99",
      "Saving": "$64.01",
      "SavingPercent": "64%"
    }
  ]
}

```

rtvslo.si 1:

```

{
  "Autor": "Miha Merljak",
  "PublishedTime": "28. december 2018 ob 08:51",
  "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, še preden je sploh zapeljal na parkirni prostor, rezerviran za izvršnega direktorja. ",
  "Content": "Audi A6 je avtomobil za direktorje, ki se mu na avtocesti spoštljivo umikajo, kot bi šlo za Pahorjev ali Šarčev avtomobil. Foto: David Šavli Audi A6 je avtomobil za direktorje, ki se mu na avtocesti spoštljivo umikajo, kot bi šlo za Pahorjev ali Šarčev avtomobil. Foto: David Šavli 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 Šarčev 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. Novo poglavje 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 še 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 še 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 športno podvozje in vsekakor za štirikolesno krmiljenje. S tem postane A6 med ovinki v občutku na volanu še veliko krajši in bolj agilen. Vse naštetost 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. Ključni tehnični podatki:- na testu Audi A6 50 TDI quattro tiptronic  
Mere:- dolžina: 4,9 m- medosna razdalja: 2,9 m- obračalni krog: 12,1 m- prtljažnik: 530 l- masa: 1.900 kg  
Pogon:- trilitrski šestvaljni 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 CO2: 147 g/km  
Stroški pri 15.000 km in 5-letni uporabi:- nakupna cena: 69.080 EUR- stroški finančnega lizinga: 4.463 EUR/5 let- stroški registracije: 10.829 EUR/5 let- stroški vzdrževanja: 1.926 EUR/5 let- stroški goriva: 6.702 EUR/75.000 km- strošek 1 kompleta pnevmatik: 716 EUR- vrednosti po 5 letih po Eurotaxu: 33.964 EUR- stroški skupaj: 1.001 EUR/mesec"

}

rtvslo.si 2:

{

"Autor": "Miha Merljak",  
"PublishedTime": "25. januar 2019 ob 15:23",  
"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": "XC40 se v mestu odlično zlije z okolico. Foto: David Šavli      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 premišljeno funkcionalnost ter na dotik prijetne materiale. Še 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 večfunkcijskim 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 rešitvami 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 še nekoliko bolj trdo, se je to samo še 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 navduši 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 številnimi asistenčnimi sistemi in izstopajočim skandinavskim dizajnom v preišljenem trenutku vstopil na trg modnih mestnih terencev, v katerem se brez ene same sence dvoma suvereno postavi med najdražje in najbolj premijske v mestu. Ključni tehnični podatki:- 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 Stroški pri 15.000 km in 5-letni uporabi:- nakupna cena: 43.619 EUR- stroški finančnega leasinga: 3.268 EUR/5 let- stroški registracije: 8.701 EUR/5 let- stroški vzdrževanja: 2.320 EUR/5 let- stroški goriva: 6.190 EUR/75.000 km- strošek 1 kompleta pnevmatik: 923 EUR- vrednosti po 5 letih po Eurotaxu: 18.886 EUR- stroški skupaj: 774 EUR/mesec"

}

partis.si 1:

[

```
{
  "Naslov": "Game of Thrones S08E03 SLO Subs 720p AMZ...",
  "Podnaslov": "IGRA PRESTOLOV | 3. del | SLO subs 720p",
  "Velikost": "2 GB",
  "Sejalcev": "309",
  "Pijavk": "6",
  "Prenosov": "1318"
},
{
  "Naslov": "Game of Thrones S08E03 SLO Subs 1080p AMZ...",
```

```

        "Podnaslov": "8. sezona | 3. del | SLOSubs | 1080p AMZN",
        "Velikost": "4.3 GB",
        "Sejalcev": "1640",
        "Pijavk": "26",
        "Prenosov": "7911"
    },
    {
        "Naslov": "Game of Thrones S08E02 SLOSubs 1080p AMZ...",
        "Podnaslov": "8. sezona | 2. del | 1080p AMZN",
        "Velikost": "3 GB",
        "Sejalcev": "1126",
        "Pijavk": "6",
        "Prenosov": "8947"
    },
    {
        "Naslov": "Game of Thrones S08E01 SLOSubs 1080p AMZ...",
        "Podnaslov": "IGRA PRESTOLOV | Nova sezona | 1.del | SLO...",
        "Velikost": "2.8 GB",
        "Sejalcev": "1256",
        "Pijavk": "9",
        "Prenosov": "11820"
    },
    {
        "Naslov": "Game Of Thrones Audiobooks 1-5 ",
        "Podnaslov": "GoT | 5 knjig | George R. R. Martin",
        "Velikost": "9.1 GB",
        "Sejalcev": "30",
        "Pijavk": "3",
        "Prenosov": "225"
    },
    {
        "Naslov": "Mastermix Pro Disc Plus The Ones That Go...",
        "Podnaslov": "Mix | 2000-2014 | 320kbps | [MuSi] ",
        "Velikost": "26.6 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "32"
    },
    {
        "Naslov": "Britains Got Talent S12E01 XviD-AFG",
        "Podnaslov": "1. Del - Nova sezona",
        "Velikost": "1.8 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "114"
    }

```



```

},
{
  "Naslov": "Britains Got Talent S12E01 HDTV x264-PLU...",
  "Podnaslov": "1. Del - Nova sezona",
  "Velikost": "582.2 MB",
  "Sejalcev": "3",
  "Pijavk": "0",
  "Prenosov": "171"
},
{
  "Naslov": "Britains Got Talent S12E01 720p HDTV x26...",
  "Podnaslov": "1. Del - Nova sezona",
  "Velikost": "1.7 GB",
  "Sejalcev": "3",
  "Pijavk": "0",
  "Prenosov": "134"
},
{
  "Naslov": "Game of Thrones S07 SLOSubs 720p AMZN WE...",
  "Podnaslov": "IGRA PRESTOLOV | Celotna 7.sezona | SLO po...",
  "Velikost": "9.8 GB",
  "Sejalcev": "140",
  "Pijavk": "18",
  "Prenosov": "8981"
},
{
  "Naslov": "Game of Thrones S07 Complete SLOSubs 108...",
  "Podnaslov": "Celotna 7. sezona | 1080p | SLO podnapisi ",
  "Velikost": "27.1 GB",
  "Sejalcev": "95",
  "Pijavk": "19",
  "Prenosov": "4486"
},
{
  "Naslov": "Game of Thrones S07E07 The Dragon and th...",
  "Podnaslov": "7.del | 1080p AMZN WEB-DL",
  "Velikost": "4.7 GB",
  "Sejalcev": "22",
  "Pijavk": "4",
  "Prenosov": "4578"
},
{
  "Naslov": "Game of Thrones S07E07 The Dragon and th...",
  "Podnaslov": "7.del | 720p AMZN WEB-DL",
  "Velikost": "1.7 GB",

```

```

        "Sejalcev": "14",
        "Pijavk": "0",
        "Prenosov": "2981"
    },
    {
        "Naslov": "Game of Thrones S07E06 Beyond the Wall S...",
        "Podnaslov": "IGRA PRESTOLOV S07, 6.del | SLOSubs | 720p",
        "Velikost": "1.8 GB",
        "Sejalcev": "14",
        "Pijavk": "1",
        "Prenosov": "2976"
    },
    {
        "Naslov": "Game of Thrones S07E06 Beyond the Wall 1...",
        "Podnaslov": "6.del | 1080p AMZN WEB-DL",
        "Velikost": "4.7 GB",
        "Sejalcev": "2",
        "Pijavk": "0",
        "Prenosov": "943"
    },
    {
        "Naslov": "Game of Thrones S07E05 SLOSubs Eastwatch...",
        "Podnaslov": "IGRA PRESTOLOV | 5.del | SLO podnapisi | 7...",
        "Velikost": "1.2 GB",
        "Sejalcev": "35",
        "Pijavk": "6",
        "Prenosov": "8007"
    },
    {
        "Naslov": "Game of Thrones S07E05 Eastwatch 1080p A...",
        "Podnaslov": "5.del | 1080p AMZN WEB-DL",
        "Velikost": "3.3 GB",
        "Sejalcev": "23",
        "Pijavk": "0",
        "Prenosov": "5332"
    },
    {
        "Naslov": "Game of Thrones S07E04 1080p AMZN WEBRip...",
        "Podnaslov": "Igra prestolov S07E04 | 1080p AMZN",
        "Velikost": "3 GB",
        "Sejalcev": "7",
        "Pijavk": "0",
        "Prenosov": "2089"
    },
    {

```

```

"Naslov": "Game of Thrones S07E04 The Spoils of War...",
"Podnaslov": "4.del | 720p",
"Velikost": "1.2 GB",
"Sejalcev": "9",
"Pijavk": "0",
"Prenosov": "2063"
},
{
"Naslov": "Game of Thrones S07E03 The Queens Justic...",
"Podnaslov": "3.del | 1080p",
"Velikost": "4.2 GB",
"Sejalcev": "10",
"Pijavk": "0",
"Prenosov": "3499"
},
{
"Naslov": "Game of Thrones S07E02 Stormborn 1080p A...",
"Podnaslov": "2.del | 1080p AMZN WEBRip",
"Velikost": "3.6 GB",
"Sejalcev": "14",
"Pijavk": "0",
"Prenosov": "5069"
},
{
"Naslov": "Game of Thrones S07E02 Stormborn 720p AM...",
"Podnaslov": "2.del | 720p AMZN WEBRip",
"Velikost": "1.2 GB",
"Sejalcev": "48",
"Pijavk": "4",
"Prenosov": "8933"
},
{
"Naslov": "Britains Got Talent S11E01 XviD-AFG",
"Podnaslov": "1. Del - Nova sezona",
"Velikost": "1.9 GB",
"Sejalcev": "0",
"Pijavk": "0",
"Prenosov": "179"
},
{
"Naslov": "Britains Got Talent S11E01 720p HDTV x26...",
"Podnaslov": "1. Del - Nova sezona",
"Velikost": "1.7 GB",
"Sejalcev": "0",
"Pijavk": "0",

```

```

        "Prenosov": "256"
    },
    {
        "Naslov": "Jane Got a Gun 2016 SloSubs BluRay 1080p...",
        "Podnaslov": "OBOROŽENA JANE | 1080p | DrSi",
        "Velikost": "11.7 GB",
        "Sejalcev": "1",
        "Pijavk": "0",
        "Prenosov": "577"
    },
    {
        "Naslov": "Jane Got a Gun 2015 SrbSubs 1080p BluRay...",
        "Podnaslov": "Western | Natalie Portman",
        "Velikost": "9 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "262"
    },
    {
        "Naslov": "VA-Bazenar's Top 13 Vol 015 (2013) [by b...",
        "Podnaslov": "mislim da ste tole čakal. Sami Hiti Hudo H...",
        "Velikost": "1.1 GB",
        "Sejalcev": "17",
        "Pijavk": "1",
        "Prenosov": "10616"
    }
}
]

```

**partis.si 2:**

```

[
    {
        "Naslov": "Microsoft office 2003 Professional Enter...",
        "Podnaslov": "Microsoft office 2003 angleska verzija",
        "Velikost": "428.1 MB",
        "Sejalcev": "4",
        "Pijavk": "0",
        "Prenosov": "8166"
    },
    {
        "Naslov": "Microsoft Office 2003 Slo",
        "Podnaslov": "Microsoft Office 2003 Slo",
        "Velikost": "362.8 MB",
        "Sejalcev": "14",
        "Pijavk": "0",
    }
]

```

```

        "Prenosov": "29009"
    },
    {
        "Naslov": "MS Office 2003 slo+serial",
        "Podnaslov": "Microsoft Word 2003",
        "Velikost": "428.2 MB",
        "Sejalcev": "120",
        "Pijavk": "1",
        "Prenosov": "73364"
    }
]

```

## 4.2 XPath

overstock 1:

```

[
    {
        "Title": "10-kt. Seven Diamond Ladies Heart Ring (0.08 TW)",
        "Content": "This ladies fashion ring dazzleswith 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 10-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- to6-mm cultured pearls on 10-karat yellow gold settings highlight these petiteearrings. A dainty rhodium-plated gold butterfly studded with a diamond (0.02total 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 withthis diamond S tennis bracelet.
        Crafted in 14-karat gold, the piece features49 diamonds for two full carats. The
        7.25-inch bracelet closes with a pressureclasp.",
        "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 anddiamond 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 this14-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 thepearl.",
        "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, sparklingdiamonds to represent past,
        present, and future, this stunning pendant isset in gleaming 14-karat gold. It
        incorporates a total of nine diamonds (0.25total 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-2clarity) 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 fourprongs 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 Romancross of 14-karat white gold drapes a
slender banner containing three brightprong-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 needsa classic pair of diamond solitaire
earrings. Set in 14-karat gold, thesediamond stud earrings (0.50 total carat
weight) have post backs with butterflyclasps.",
    "ListPrice": "$999.99",
    "Price": "$359.99",
    "Saving": "$640.00",
    "SavingPercent": "64%"
  },
  {
    "Title": "14-kt. Cultured Pearl Diamond Earrings",
    "Content": "Create an elegant appearance withthese pearl and diamond stud
earrings. Set in 14-karat yellow gold, eachearring features an 8 to 8.5-mm
cultured white pearl. Prong-set round diamondsaccent 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 jewelrycollection with this 14-karat gold,
        diamond, and pearl necklace. The 7.5-8mm 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-cutdiamonds (0.50 total carat
        weight, G-H color, I-1 to I-2 clarity) set infour 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 baguettediamonds. Treat her like royalty and
        save when you buy jewelry treasuresat Overstock.com.",
        "ListPrice": "$1,635.00",
        "Price": "$609.99",
        "Saving": "$1,025.01",
        "SavingPercent": "62%"
    }
]

```

overstock 2:

```

[
    {
        "Title": "14-kt. Green Jade Hoops",
        "Content": "Hoops of cool green jade restbetween 14-karat yellow gold endpieces.
        The hoops graduate in thickness from3 mm at the ends to 6 mm in the center, with
        approximately 29 mm overallldiameter.",
        "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: brilliantgreen, 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: brilliantgreen, 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%"
    }
]

```

rtvslo 1:

```

{
    "Title": "Audi A6 50 TDI quattro: nemir v premijskem razredu",
    "Subtitle": "Test nove generacije",
    "Author": "Miha Merljak",
    "PublishedTime": "28. december 2018 ob 08:51",
    "Lead": "To je novi audi A6. V razred najdražjih in najbolj premijskih žrebcev
    je vnesel nemir, še preden je sploh zapeljal na parkirni prostor, rezerviran za
    izvršnega direktorja. ",
    "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 Šarčev 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., Novo poglavje, 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 še 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 še 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 športno podvozje in vsekakor za štirikolesno krmiljenje. S tem postane A6 med ovinki 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., Ključni tehnični podatki:, - na testu Audi A6 50 TDI quattro tiptronic, Mere:, - dolžina: 4,9 m, - medosna razdalja: 2,9 m, - obračalni krog: 12,1 m, - prtljažnik: 530 l, - masa: 1.900 kg, Pogon:, - trilitrski šestvaljni 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, 2, : 147 g/km, Stroški pri 15.000 km in 5-letni uporabi:, - nakupna cena: 69.080 EUR, - stroški finančnega lizinga: 4.463 EUR/5 let, - stroški registracije: 10.829 EUR/5 let, - stroški vzdrževanja: 1.926 EUR/5 let, - stroški goriva: 6.702 EUR/75.000 km, - strošek 1 kompleta pnevmatik: 716 EUR, - vrednosti po 5 letih po Eurotaxu: 33.964 EUR, - stroški skupaj: 1.001 EUR/mesec"

}

rtvslo 2:

{

"Title": "Volvo XC 40 D4 AWD momentum: suvereno med najboljše v razredu",  
 "Subtitle": "Test novega modela",  
 "Author": "Miha Merljak",  
 "PublishedTime": "25. januar 2019 ob 15:23",  
 "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 nišnimi 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. , Še 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 večfunkcijskim 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 rešitvami 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 še nekoliko bolj trdo, se je to samo še 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 navduši predvsem pri avtocestnih dolgotrajnih 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 izstopajočim 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., Ključni tehnični podatki:, - 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, Stroški pri 15.000 km in 5-letni uporabi:, - nakupna cena: 43.619 EUR, - stroški finančnega leasinga: 3.268 EUR/5 let, - stroški registracije: 8.701 EUR/5 let, - stroški vzdrževanja: 2.320 EUR/5 let, - stroški goriva: 6.190 EUR/75.000 km, - strošek 1 kompleta pnevmatik: 923 EUR, - vrednosti po 5 letih po Eurotaxu: 18.886 EUR, - stroški skupaj: 774 EUR/mesec"

}

partis.com 1:

[

```
{
  "Naslov": "Game of Thrones S08E03 SLO Subs 720p AMZ...",
  "Podnaslov": "IGRA PRESTOLOV 3. del SLO subs 720p",
  "Velikost": "2 GB",
  "Sejalcev": "309",
  "Pijavk": "6",
  "Prenosov": "1318"
},
{
  "Naslov": "Game of Thrones S08E03 SLOSubs 1080p AMZ...",
  "Podnaslov": "8. sezona 3. del SLOSubs 1080p AMZN",
  "Velikost": "4.3 GB",
  "Sejalcev": "1640",
  "Pijavk": "26",
```

```

        "Prenosov": "7911"
    },
    {
        "Naslov": "Game of Thrones S08E02 SLOSubs 1080p AMZ...",
        "Podnaslov": "8. sezona 2. del 1080p AMZN",
        "Velikost": "3 GB",
        "Sejalcev": "1126",
        "Pijavk": "6",
        "Prenosov": "8947"
    },
    {
        "Naslov": "Game of Thrones S08E01 SLOSubs 1080p AMZ...",
        "Podnaslov": "IGRA PRESTOLOV Nova sezona 1.del SLO...",
        "Velikost": "2.8 GB",
        "Sejalcev": "1256",
        "Pijavk": "9",
        "Prenosov": "11820"
    },
    {
        "Naslov": "Game Of Thrones Audiobooks 1-5 ",
        "Podnaslov": "GoT 5 knjig George R. R. Martin",
        "Velikost": "9.1 GB",
        "Sejalcev": "30",
        "Pijavk": "3",
        "Prenosov": "225"
    },
    {
        "Naslov": "Mastermix Pro Disc Plus The Ones That Go...",
        "Podnaslov": "Mix 2000-2014 320kbps MuSi ",
        "Velikost": "26.6 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "32"
    },
    {
        "Naslov": "Britains Got Talent S12E01 XviD-AFG",
        "Podnaslov": "1. Del - Nova sezona",
        "Velikost": "1.8 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "114"
    },
    {
        "Naslov": "Britains Got Talent S12E01 HDTV x264-PLU...",
        "Podnaslov": "1. Del - Nova sezona",

```

```

        "Velikost": "582.2 MB",
        "Sejalcev": "3",
        "Pijavk": "0",
        "Prenosov": "171"
    },
    {
        "Naslov": "Britains Got Talent S12E01 720p HDTV x26...",
        "Podnaslov": "1. Del - Nova sezona",
        "Velikost": "1.7 GB",
        "Sejalcev": "3",
        "Pijavk": "0",
        "Prenosov": "134"
    },
    {
        "Naslov": "Game of Thrones S07 SLOSubs 720p AMZN WE...",
        "Podnaslov": "IGRA PRESTOLOV Celotna 7.sezona SLO po...",
        "Velikost": "9.8 GB",
        "Sejalcev": "140",
        "Pijavk": "18",
        "Prenosov": "8981"
    },
    {
        "Naslov": "Game of Thrones S07 Complete SLOSubs 108...",
        "Podnaslov": "Celotna 7. sezona 1080p SLO podnapisi ",
        "Velikost": "27.1 GB",
        "Sejalcev": "95",
        "Pijavk": "19",
        "Prenosov": "4486"
    },
    {
        "Naslov": "Game of Thrones S07E07 The Dragon and th...",
        "Podnaslov": "7.del 1080p AMZN WEB-DL",
        "Velikost": "4.7 GB",
        "Sejalcev": "22",
        "Pijavk": "4",
        "Prenosov": "4578"
    },
    {
        "Naslov": "Game of Thrones S07E07 The Dragon and th...",
        "Podnaslov": "7.del 720p AMZN WEB-DL",
        "Velikost": "1.7 GB",
        "Sejalcev": "14",
        "Pijavk": "0",
        "Prenosov": "2981"
    },
    },

```

```

{
  "Naslov": "Game of Thrones S07E06 Beyond the Wall S...",
  "Podnaslov": "IGRA PRESTOLOV S07, 6.del  SLOSubs  720p",
  "Velikost": "1.8 GB",
  "Sejalcev": "14",
  "Pijavk": "1",
  "Prenosov": "2976"
},
{
  "Naslov": "Game of Thrones S07E06 Beyond the Wall 1...",
  "Podnaslov": "6.del  1080p  AMZN WEB-DL",
  "Velikost": "4.7 GB",
  "Sejalcev": "2",
  "Pijavk": "0",
  "Prenosov": "943"
},
{
  "Naslov": "Game of Thrones S07E05 SLOSubs Eastwatch...",
  "Podnaslov": "IGRA PRESTOLOV  5.del  SLO podnapisi  7...",
  "Velikost": "1.2 GB",
  "Sejalcev": "35",
  "Pijavk": "6",
  "Prenosov": "8007"
},
{
  "Naslov": "Game of Thrones S07E05 Eastwatch 1080p A...",
  "Podnaslov": "5.del  1080p  AMZN WEB-DL",
  "Velikost": "3.3 GB",
  "Sejalcev": "23",
  "Pijavk": "0",
  "Prenosov": "5332"
},
{
  "Naslov": "Game of Thrones S07E04 1080p AMZN WEBRip...",
  "Podnaslov": "Igra prestolov S07E04  1080p  AMZN",
  "Velikost": "3 GB",
  "Sejalcev": "7",
  "Pijavk": "0",
  "Prenosov": "2089"
},
{
  "Naslov": "Game of Thrones S07E04 The Spoils of War...",
  "Podnaslov": "4.del  720p",
  "Velikost": "1.2 GB",
  "Sejalcev": "9",

```

```

        "Pijavk": "0",
        "Prenosov": "2063"
    },
    {
        "Naslov": "Game of Thrones S07E03 The Queens Justic...",
        "Podnaslov": "3.del 1080p",
        "Velikost": "4.2 GB",
        "Sejalcev": "10",
        "Pijavk": "0",
        "Prenosov": "3499"
    },
    {
        "Naslov": "Game of Thrones S07E02 Stormborn 1080p A...",
        "Podnaslov": "2.del 1080p AMZN WEBRip",
        "Velikost": "3.6 GB",
        "Sejalcev": "14",
        "Pijavk": "0",
        "Prenosov": "5069"
    },
    {
        "Naslov": "Game of Thrones S07E02 Stormborn 720p AM...",
        "Podnaslov": "2.del 720p AMZN WEBRip",
        "Velikost": "1.2 GB",
        "Sejalcev": "48",
        "Pijavk": "4",
        "Prenosov": "8933"
    },
    {
        "Naslov": "Britains Got Talent S11E01 XviD-AFG",
        "Podnaslov": "1. Del - Nova sezona",
        "Velikost": "1.9 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "179"
    },
    {
        "Naslov": "Britains Got Talent S11E01 720p HDTV x26...",
        "Podnaslov": "1. Del - Nova sezona",
        "Velikost": "1.7 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "256"
    },
    {
        "Naslov": "Jane Got a Gun 2016 SloSubs BluRay 1080p...",

```



```

        "Podnaslov": "OBOROŽENA JANE 1080p DrSi",
        "Velikost": "11.7 GB",
        "Sejalcev": "1",
        "Pijavk": "0",
        "Prenosov": "577"
    },
    {
        "Naslov": "Jane Got a Gun 2015 SrbSubs 1080p BluRay...",
        "Podnaslov": "Western Natalie Portman",
        "Velikost": "9 GB",
        "Sejalcev": "0",
        "Pijavk": "0",
        "Prenosov": "262"
    },
    {
        "Naslov": "VA-Bazenars Top 13 Vol 015 (2013) by b...",
        "Podnaslov": "mislim da ste tole čakal. Sami Hiti Hudo H...",
        "Velikost": "1.1 GB",
        "Sejalcev": "17",
        "Pijavk": "1",
        "Prenosov": "10616"
    }
]

```

**partis.com2:**

```

[
    {
        "Naslov": "Microsoft office 2003 Professional Enter...",
        "Podnaslov": "Microsoft office 2003 angleska verzija",
        "Velikost": "428.1 MB",
        "Sejalcev": "4",
        "Pijavk": "0",
        "Prenosov": "8166"
    },
    {
        "Naslov": "Microsoft Office 2003 Slo",
        "Podnaslov": "Microsoft Office 2003 Slo",
        "Velikost": "362.8 MB",
        "Sejalcev": "14",
        "Pijavk": "0",
        "Prenosov": "29009"
    },
    {
        "Naslov": "MS Office 2003 slo+serial",

```

```

        "Podnaslov": "Microsoft Word 2003",
        "Velikost": "428.2 MB",
        "Sejalcev": "120",
        "Pijavk": "1",
        "Prenosov": "73364"
    }
]

```

### 4.3 RoadRunner

Izhod algoritma je definiran enako, kot so ga navedli avtorji članka [1]. Torej iterativne elemente predstavimo z `(elementi)+`, opcijske z `(element)?`, podatke pa z `#PCDATA`. Algoritem smo uspeli uspešno implementirati za primer iz članka ter nekaj lažjih primerov, na podanih straneh pa ne deluje v popolnosti.

Delovanje je bilo ustrezno na straneh domene `partis.si`, in `overstock.com`, v kolikor smo jih priredili. Ročno smo morali izvzeti ven glavni seznam, ki vsebuje iterativne elemente ter odstraniti tiste pomožne elemente, ki se nahajajo znotraj iterativnih elementov. Prilagamo rezultate za prirejene strani.

#### **overstock.com**

```

<body>
<td>
<br>
</br>
<b>
<a>
Jewelry, Watches & Gifts
</a>
>
<a>
Jewelry
</a>
>
<a>
#PCDATA
</a>
> View All
</b>
<br>
</br>
<table>
<tbody>
<tr>
<td>

```

```

<br>
</br>
<table>
<tbody>
<tr>
<td>
<table>
<tbody>
<tr>
<td>
<span>
<b>
List Sorted By:
</b>
</span>
</td>
<td>
<b>
Top Sellers
</b>
|
<a>
Discount
</a>
|
<a>
Newest First
</a>
|
<a>
Price
</a>
|
<a>
Quantity
</a>
|
<a>
Markdowns
</a>
</td>
</tr>
(<tr>)?
(<td>)?
(<span>)?

```

```

(<b>)?
(Items:)?
(</b>)?
(</span>)?
(</td>)?
(<td>)?
(<b>)?
(16)?
(</b>)?
( - )?
(<b>)?
(30)?
(</b>)?
( of )?
(<b>)?
(296)?
(</b>)?
( | )?
(<a>)?
(First Page)?
(</a>)?
( | )?
(<a>)?
(Previous 15)?
(</a>)?
( | )?
(<a>)?
(Next 15)?
(</a>)?
(</td>)?
(</tr>)?
</tbody>
</table>
</td>
</tr>
</tbody>
</table>
</td>
</tr>
<tr>
<td>
<table>
<tbody>
<tr>
<td>

```

```

<table>
<tbody>
  (<tr>
    <td>
      <table>
        <tbody>
          <tr>
            <td>
              <a>
                <img>
              </img>
            </a>
          </td>
        </tr>
        <tr>
          <td>
            <a>
              More Info...
            </a>
          </td>
        </tr>
      </tbody>
    </table>
  </td>
  <td>
    <a>
      <b>
        #PCDATA
      </b>
    </a>
    <br>
  </br>
  <table>
    <tbody>
      <tr>
        <td>
          <table>
            <tbody>
              <tr>
                <td>
                  <b>
                    List Price:
                  </b>
                </td>
              </tr>
            </tbody>
          </table>
        </td>
      </tr>
    </tbody>
  </table>

```

```

<s>
#PCDATA
</s>
</td>
</tr>
<tr>
<td>
<b>
Price:
</b>
</td>
<td>
<span>
<b>
#PCDATA
</b>
</span>
</td>
</tr>
<tr>
<td>
<b>
You Save:
</b>
</td>
<td>
<span>
#PCDATA
</span>
</td>
</tr>
</tbody>
</table>
</td>
<td>
<span>
#PCDATA
<br>
</br>
<a>
<span>
<b>
Click here to purchase.
</b>
</span>

```

```

</a>
</span>
<br>
</br>
</td>
</tr>
</tbody>
</table>
</td>
</tr>
<tr>
<td>
<img>
</img>
</td>
</tr>)+
</tbody>
</table>
</td>
</tr>
</tbody>
</table>
</td>
</tr>
<tr>
<td>
<br>
</br>
<table>
<tbody>
<tr>
<td>
<table>
<tbody>
<tr>
<td>
<span>
<b>
List Sorted By:
</b>
</span>
</td>
<td>
<b>
Top Sellers

```

```

</b>
|
<a>
Discount
</a>
|
<a>
Newest First
</a>
|
<a>
Price
</a>
|
<a>
Quantity
</a>
|
<a>
Markdowns
</a>
</td>
</tr>
(<tr>)?
(<td>)?
(<span>)?
(<b>)?
(Items:)?
(</b>)?
(</span>)?
(</td>)?
(<td>)?
(<b>)?
(16)?
(</b>)?
( - )?
(<b>)?
(30)?
(</b>)?
( of )?
(<b>)?
(296)?
(</b>)?
( | )?
(<a>)?

```



```
(First Page)?
(</a>)?
( | )?
(<a>)?
(Previous 15)?
(</a>)?
( | )?
(<a>)?
(Next 15)?
(</a>)?
(</td>)?
(</tr>)?
</tbody>
</table>
</td>
</tr>
</tbody>
</table>
</td>
</tr>
</tbody>
</table>
</td>
</body>
```

partis.si

```
<body>
<div>
  (<div>
    <div>
      <div>
        </div>
      </div>
    <div>
      <a>
        #PCDATA
      </a>
    <div>
      <div>
        #PCDATA
      </div>
      <div>
        <div>
```

```

<a>
<img>
</img>
<span>
<span>
</span>
<span>
#PCDATA
<br>
</br>
#PCDATA
<br>
</br>
#PCDATA
<br>
</br>
#PCDATA
<br>
</br>
#PCDATA
</span>
<span>
</span>
</span>
</a>
</div>
</div>
</div>
</div>
<div>
<a>
<img>
</img>
</a>
</div>
<div>
#PCDATA
</div>
<div>
#PCDATA
</div>
<div>
#PCDATA
</div>
<div>

```

```
#PCDATA  
</div>  
</div>)+  
</div>  
</body>
```

## 5 Zaključek

Uporaba izrazov XPath ter regularnih izrazov se je na majhni množici spletnih strani izkazala za zanesljivo, vendar zahteva da za vsako skupino oziroma tip strani napišemo nove izraze, kar je neučinkovito. Težava je tudi robustnost, saj metodi prenehata delovati že ob majhnih spremembah na spletnih straneh (predvsem metoda z uporabo regularnih izrazov, pri kateri se sklicujemo neposredno na kodo oz. besedilo). Na podani množici strani smo sicer z nekaj obdelave podatkov po ekstrakciji uspeli pridobiti vse zahtevane podatke, v ustrezni obliki. Algoritem RoadRunner se je izkazal za večji izziv kot smo pričakovali. Implementirali smo iskanje iteratorjev ter opsijskih elementov, kot to opisuje članek ter uspeli pridobiti podatke iz strani, če te ustrezno prej uredimo. Potrebno bi bilo zagotoviti še ustrezno rekurzivno iskanje iteratorjev in opsijskih elementov znotraj obstoječih kandidatov za iteratorje.

## Literatura

- [1] Crescenzi, V., Mecca, G., Merialdo, P., et al. (2001). Roadrunner: Towards automatic data extraction from large web sites. In *VLDB*, volume 1, pages 109–118.