Blaž Štempelj in Denis Grabljevec

Seminarska naloga: Ekstrakcija podatkov s spleta

Uvod

Dandanes je svetovni splet največja baza podatkov na svetu. Ker pa so ti podatki predvsem namenjeni ljudem, so večinoma ne-strukturirane narave. To pomeni, da je njihova struktura takšnja, da računalniškim programov, ki poskušajo iz nje pridobiti pomembne informacije, predstavlja težave pri branju in interpretiranju. Za namene branja nestrukturiranih podatkov so se razvile različne metode za pridobivanje podatkov s spleta. Te metode delimo na: *(a) ročne* in *(b) avtomatske*.

Ročne metode zahtevajo, da njihov uporabnik sam opiše katere podatke želi pridobiti in kako jih bo pridobil. Sem spadajo *regularni izrazi* [2,3] in jezik *XPath* [1]. Regularni izrazi predstavljajo gramatiko, ki opiše kako iz besedila pridobiti želene podatke. Jezik XPath pa izkorišča drevesno strukturo XML ali HTML dokumenta tako, da nam omogoči sprehod po njegovi drevesni strukturi do želenega vozlišča (vsebuje nam pomembne informacije) z uporabo različnih kriterijev. Glavna slabost ročnih metod je, da je potrebno v primeru, da se struktura spletne strani, v nekem trenutku spremeni, ponovno potrebno opisati kako pridobiti želene podatke.

Avtomatske metode znajo samodejno iz same strukture HTML pridobiti želene podatke. Sem spada algoritem *RoadRunner* [4], ki deluje tako, da s primerjavo dveh podobnih spletnih strani določi v katerih elementih ali besedilu se strani razlikujeta. Na podlagi teh razlik zgradi regularni izraz, ki generalizira strukturo obeh strani. Ta regularni izraz predstavlja ovojnico podobnih spletnih strani, ki nam omogoča, da iz nestrukturirane vsebine pridobimo želene podatke (v strukturirani obliki).

Implementacija

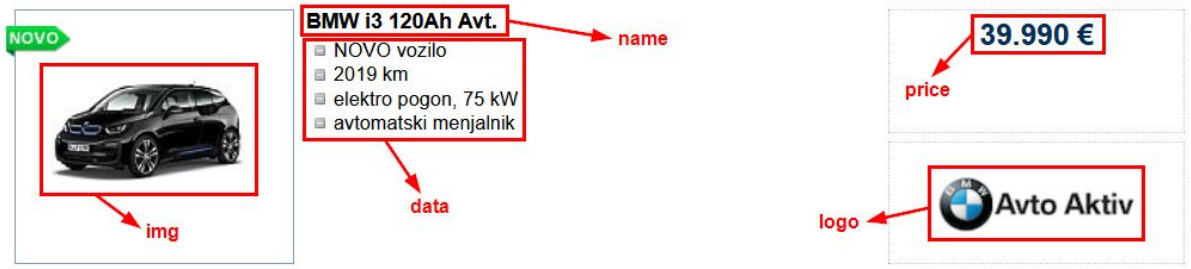
Seminarska naloga je zahtevala ekstrakcijo podatkov iz dveh vnaprej pridobljenih spletnih strani (*rtvslo.si* in *overstock.com*) in dveh spletnih strani po svoji izbiri. Spletne strani morajo vsebovati bodisi seznam podatkov (izdelkov, člankov,..) bodisi morajo biti opisne strani izdelkov, člankov, itd.

Predpriprava

Kot že omenjeno je bilo potrebno poleg že vnaprej pridobljenih spletnih strani, pridobiti štiri strani po lastni izbiri. Za dodatne spletne strani sva izbrala *avto.net* in *slotech.si*.

avto.net

Spletna stran avto.net, služi kupovanju rabljenih avtomobilov. Uporabnik lahko preko različnih kriterijev pridobi rezultate v obliki seznama. Po vsebini je najbolj podobna strani overstock.com. Spodnja slika prikazuje, katere podatke sva pridobila iz seznama rezultatov. Posebno je bilo potrebno paziti na ceno avtomobila, ker se je enkrat pojavljala navadna cena, enkrat pa stara cena skupaj z akcijsko ceno.



slotech.si

Spletna stran slotech.si je namenjena novicam o tehnoloških novitetah in tehnologiji nasploh. Vsebinsko in strukturno je podobna rtvslo.si. Spodnja slika prikazuje, katere podatke sva pridobila iz izbrane novice.

Regularni izrazi

overstock.com

Ekstrakcija podatkov iz strani je bila dokaj enostavna. Vsi podatki, ki smo jih želeli ekstrahirati so oviti vsak v svojo unikatno html značko ali značko in css razred. Največ težav sta predstavljala regularna izraza za pridobitev *prihraneka in odstotka popusta*.

- **naslov:** `<a href="(.*)"><b>(.*)</b></a><br>`
- **osnovna cena:** `<s>([$]\s*[0-9.,]+)</s>`
- **cena:** `<span class="bigred"><b>([$]\s*[0-9.,]+)</b></span>`
- **prihranek in odstotek popusta:** `([$]\s*[0-9.,]+)\s*\((\d+%)\)`
- **vsebina:** `<span class="normal">(.*?)<br>`

rtvslo.si

Ta stran je zahtevala nekoliko več dela kot prejšnja. Potrebno je namreč uporabiti tri različne regularne izraze za pridobitev vsebine članka. Vsi ostali regularni izrazi so uporabljeni predvsem za ekstrakcija teksta med html značkami z izjemo *časa objave,* kjer je bil potreben dodaten regularni izraz za ekstrakcijo datuma in časa.

- **avtor**: `<strong>(.*?)</strong>`
- **čas objave**: `\d{1,2}\.\s*(januar|februar|marec|april|maj|junij|julij|avgust|september|oktober|november|december)\s*\d{4}\s*\w*\s*\d{2}:\d{2}`
- **naslov**: `<h1>(.*?)</h1>`
- **podnaslov**: `<div class="subtitle">(.*?)</div>`
- **lead**: `<p class="lead">(.*?)</p>`
- **vsebina**
    a. izločitev html kode med značkama article: `<article class="article">(.*?)</article>`
    b. izločitev teksta med značkami p: `<p[^>]*>(.*?)<\/p>`
    c. iskanje html značk: `<\/?\w+\s*.*?>`

slotech.com

Podobno kot pri člankih na strani rtvslo.si, je bilo tudi tukaj največ dela z izločitvijo vsebine iz samega članka. Vsi ostali regularni izrazi so podobni tistih na ostalih straneh (iskanje teksta med html značkami). Izjema je ločitev datuma in časa, ter vira na podlagi katerega je bil spisan članek (nahaja se neposredno pred samo vsebino članka).

- **naslov**: `<h3 itemprop="headline"><a href=".*" itemprop="name">(.*?)</a></h3>`
- **avto**: `<span itemprop="name">(.*?)</span>`
- **datum**: `\d{1,2}\.\s(jan|feb|mar|apr|maj|jun|jul|avg|sep|okt|nov|dec)\s\d{4}`
- **čas**: `\d{2}:\d{2}`
- **vir**: `<a\sclass="source".*?>(.*?)<\/a>`
- **kategorija**: `itemprop="articleSection">(.*?)<\/a>`
- **vsebina**: `>\s-\s(.*?)</div>`

avto.net

To je bila najbolj zahtevna stran za ekstrakcijo podatkov. Ponovno kot prej, se je večino zanimivih podatkov nahajalo med html značkami. Zataknilo pa se je pri ekstrakciji cene avtomobila. Ta je namreč lahko v dveh oblikah: *navadna cena* ali *akcijska cena*. Ekstrakcija navadne cene ni težavna. Pri ekstrakciji akcijske cene pa je potrebno paziti, da po pomoti ne zajamemo *stare cene*, ki se nahaja pred akcijsko. Za rešitev te težave je bilo potrebno uporabiti iskanje za nazaj; regularni izraz, ki najde ceno, izpusti le-to, če se pred njo nahaja značka s css razredom “StaraCena”.

- **ime**: `<a\sclass="Adlink".*?>\n?<span>(.*?)</span>\n?</a>`
- **slika**: `<div\sclass="ResultsAdPhotoTop">\n?\s*.*?<img\ssrc=\"(.*?)\"`
- **logo**: `<div\sclass="ResultsAdLogo">\n?\s*.*?<img\ssrc=\"(.*?)\"`
- **cena**: `ResultsAdPrice[^>]+>.*?(?<!StaraCena\">)(\d{2}\.\d{3})`
- **podatki**: regex: `<div\sclass="ResultsAdDataTop">.*?<ul>(.*?)</ul>`

XPath

rtvslo.si

Spletna stran rtvslo.si je dobro strukturirana, html elementi in css razredi so logično poimenovani ter uporabljajo nove HTML5 elemente (header, article). Kar omogoča, da se podatke preko metode XPath, pridobi enostavno.

- **avtor** - pridobivanje avtorja je bilo zaradi css razreda *author-name* enostavno
    - `//div[@class="author-name"]/text()`
- **čas objave** - podatki, ki opisujejo čas objave članka se skrivajo za css razredom *publish-meta*
    - `//div[@class="publish-meta"]/text()`
- **naslov** - naslov članka se skriva v header znački s css razredom *article-header*
    - `//header[@class="article-header"]/h1/text()`

- **podnaslov** - podnaslov se skriva v istem header vozlišču kot naslov, le pod drugim otrokom
    - `//header[@class="article-header"]/div[@class="subtitle"]/text()`
- **lead** - Tudi lead se skriva v istem vozlišču kot naslov ter podnaslov
    - `//header[@class="article-header"]/p[@class="lead"]/text()`
- **vsebina** - vsebina članka se skriva pod css razredom *article-body*, ter otrokom elementom article
    - `//div[@class="article-body"]/article[@class="article"]/p/text()`

overstock.com

Spletna stran overstock.com je večinoma narejena s tabelami (gre za prakse starejših spletnih strani), zato je bilo potrebno paziti, da smo pri pridobivanju podatkov upoštevali pravo zaporedje html elementov (table, tr, td, itd).
Vsi potrebni podatki vsebujejo enako vozlišče, zato sem za osnovni XPath vzel drugo *td* značko v najvišje postavljeni tabeli: ***`//td[2][@valign="top"]`***

- **naslov** - naslov je bil edini podatek v najvišje postavljeni tabeli v html drevesu
    - `osnovni XPath + /a/b/text()`
- **vsebina** - vsebina izdelka se skriva v tabeli globje pod drugim stolpcem
    - `osnovni XPath + /tabel + osnovni XPath + /span[@class="normal"]/text()`
- **osnovna cena** - osnovna cena se skriva v tretji tabeli globoko, kot prva vrstica tabele
    - `osnovni XPath + /table//td[1][@valign="top"]//tr[1]/td[2]/s/text()`
- **cena** - cena se prav tako skriva v tretji tabeli, le kot druga vrstica tabele
    - `osnovni XPath + /table//td[1][@valign="top"]//tr[2]/td[2]/span/b/text()`
- **prihranek in odstotek popusta** - podatka se prav tako skrivata v tretji tabeli globoko kot tretja vrstica tabele (podatki so zahtevali še nekaj Python procesiranja)
    - `/table//td[1][@valign="top"]//tr[3]/td[2]/span[@class="littleorange"]/text()`

slotech.si

Spletna stran slotech.si je za pridobitev podatkov iz metodo XPath zahtevala že nekaj bolj naprednega znanja XPath. Prav tako kot pri overstock.com spletni strani, so tudi tukaj vsi elementi spadali pod neko osnovno vejo HTML drevesa (osnovni XPath): **`//div[@id="content"]//article/`**

- **naslov** - naslov se je nahajal v elementu header, pod itemprop atributom z vrednostjo headline
    - `osnovni XPath + header/h3[@itemprop="headline"]/a/text()`
- **avtor** - podatek o avtorju se nahaja v seznamu pod elementom za povezavo z atributom *itemprop* z vrednostjo *author*
    - `Osnovni XPath + header/ul[@class="info"]//a[@itemprop="author"]/span/text()`

- **datum članka** - datum članka je na spletni strani točno ob uri vendar v drugem listu HTML drevesa, zato je bilo datum ločeno od ure enostavno dobiti
    - `osnovni XPath + header/ul[@class="info"]//span[@class="date"]/time/a/text()`
- **ura članka** - Ura članka je bila kot že omenjeno pod datumom, pri uri je bilo potrebno odstraniti predpono ob (npr. ob 12:35). Za rešitev tega problema sem si pomagal z enakim regularnim izrazom, kot je bil tisti, ki je pobiral uro v metodi regularni izraz.
    - `osnovni XPath + header/ul[@class="info"]//span[@class="date"]/time/text()`
    - regularni izraz za pomoč: `\d{2}:\d{2}`
- **kategorija** - kategorija pod elementom seznama z css razredom *categories*
    - `osnovni XPath + header/ul[@class="info"]/li[@class="categories"]/a/text()`
- **vir:** `osnovni XPath + div[@itemprop="articleBody"]/a[1]/text()`
- **vsebina članka** - Pri vsebini članka je bilo potrebno biti pazljiv, saj se je pojavljala pod elementom div z atributom *itemprop* z vrednostjo *articlebody* in enim korenom nižje (elementom a). Potrebno je bilo uporabiti *operator ali* in dva XPath-a. Pri vsebini, ki se je nahajala pod elementom a, pa smo morali biti pozorni tudi, da je bil element a vsaj drugi. Prvi element a je namreč vseboval podatek za vir.
    - `div[@itemprop="articleBody"]/text() | div[@itemprop="articleBody"]/a[position()>1]/text()`

Rezultati

overstock.com | diamonds


```
{
  'items': [
    {
      'content': 'This ladies fashion ring dazzles\n'
'with hearts and diamonds. The gold band is crafted '
'into delicate, open hearts.\n'
'Seven brilliant-cut diamonds add a bit of sparkle. ',
      'list_price': '$149.00',
      'price': '$69.99',
      'saving': '$79.01',
      'saving_percent': '53%',
      'title': '10-kt. Seven Diamond Ladies Heart Ring (0.08 TW)'
    },
    {
      'content': 'Nineteen round diamonds accent this 10-karat yellow '
'gold ring with filigree accents.',
      'list_price': '$250.00',
      'price': '$74.90',
      'saving': '$175.10',
      'saving_percent': '70%',
      'title': '10-Kt. Diamond Ring (.25 TW)'
    },
}
```


overstock.com | pendants


```
{
  'items': [
      {
        'content': 'Hoops of cool green jade rest\n'
'between 14-karat yellow gold endpieces. The hoops '
'graduate in thickness from\n'
'3 mm at the ends to 6 mm in the center, with '
'approximately 29 mm overall\n'
'diameter.',
        'list_price': '$90.00',
        'price': '$46.99',
        'saving': '$43.01',
        'saving_percent': '47%',
        'title': '14-kt. Green Jade Hoops'
      },
```

```
      {
        'content': 'The 25-mm disk hangs delicately\n'
'from a 14-karat gold chain. The disk features a '
'dramatic gold Chinese character\n'
'in the center, accompanied by four stylized gold '
'bees.',
        'list_price': '$150.00',
        'price': '$48.99',
        'saving': '$101.01',
        'saving_percent': '67%',
        'title': '14-kt. Jade Doughnut Pendant'
      },
}
```

rtvslo.si | Audi

```
{
  'author': 'Miha Merljak',
  'content': 'Samo poglejte njegovo masko \x80\x93 to ogromno satovje '
'z radarji na takem položaju, da se ti na avtocesti tudi pri 120 '
'km/h vsi spoštljivo umikajo, saj so prepri\x8dani, da gre za '
'Pahorjev ali \xa0ar\x8dev avto. Seveda, novi A6 lahko cesto in '
'promet skenira s kar petimi radarji, petimi kamerami, '
'infrarde\x8do kamero za no\x8dni vid, dvanajstimi '
'ultrazvo\x8dnimi senzorji in laserskim \x8ditalnikom \x80\x93 '
'lidarjem. V glavnem vojaška tehnologija v službi varnosti za '
'fante, ki smo radi gledali Top Gun, Bonda in druge možakarja s '
'finimi igra\x8dami.Novo poglavjeVozniški 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 pri\x8dakujete. 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 ve\x8d.\x8ce vam '
'pogled na Audijev spisek dodatne opreme ne odvzame volje do '
'življenja, potem vsekakor toplo priporo\x8damo nakup '
'zra\x8dnega vzmetenja, saj dobi z njim A6 ve\x8d razli\x8dnih '
'in vozniško zelo uporabnih karakterjev.Enako velja za seksi '
'lu\x8di z inteligentno matri\x8dno osvetlitvijo, pa za '
'športno podvozje in vsekakor za štirikolesno krmiljenje. S tem '
'postane A6 med ovinki v ob\x8dutku na volanu še veliko krajši '
'in bolj agilen. Vse našteto smo preskušali v družbi agregata '
'50 TDI, ki je v resnici klasi\x8dni trilitrski dizel, '
'podkrepljen z elektromotorjem. Ja, ta audi je mehki hibrid z '
```

```
'izjemnim navorom in dovolj mo\x8di kadar koli in kjer koli. Si '
'pa mislimo, da bo najve\x8dji del trga zadovoljil že '
'u\x8dinkovit dvolitrski mehki hibrid z mo\x8djo 150 '
'kilovatov.Klju\x8dni tehni\x8dni podatki:- na testu Audi A6 50 '
'TDI quattro tiptronicMere:- dolžina: 4,9 m- medosna razdalja: '
'2,9 m- obra\x8dalni krog: 12,1 m- prtljažnik: 530 l- masa: '
'1.900 kgPogon:- trilitrski šestvaljni dizelski motor- mo\x8d: '
'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/kmStroški pri 15.000 km in 5-letni uporabi:- '
'nakupna cena: 69.080 EUR- stroški finan\x8dnega 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',
  '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. ',
  'published_time': '28. december 2018 ob 08:51',
  'subtitle': 'Test nove generacije',
  'title': 'Audi A6 50 TDI quattro: nemir v premijskem razredu'
}
```


rtvslo.si | Volvo


```
{
  'author': 'Miha Merljak',
  'content': 'Volvo se je nižjih srednjih razredov v preteklosti izogibal'
'ali pa je vanje vstopal z zelo nišnimi produkti, ki niso pustili '
've\x8djega tržnega pe\x8data. V primeru XC 40 ni težko '
'napovedati, da bo ta tradicija prekinjena. Ponuja namre\x8d '
'visoko kakovost kon\x8dne izdelave in v kabini odli\x8dno '
'premišljeno funkcionalnost ter na dotik prijetne materiale. '
'\xa0e posebej hvalimo število, iznajdljivost in velikost '
'razli\x8dnih odlagalnih prostorov ter široke, \x8dvrste in '
'zelo udobne sedeže. Intuitivno in enostavno logi\x8dno je '
'upravljanje z velikim vmesnikom, ki z ve\x8dfunkcijskim '
'zaslonom na dotik kraljuje na z roko lahko dostopnem mestu na '
'sredinski armaturi. Razo\x8daranj ne bo niti v velikosti in '
'uporabnosti prtljažnega prostora, ki s 460 litri prostornine '
'sicer ni med ve\x8djimi 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 nevtralno in predvidljivo '
```

```
'lego. V premeru preskušanega 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 ra\x8dunati na manj udobno '
'vožnjo \x8dez razli\x8dne asfaltne grbine. Podoben razmislek '
'velja opraviti tudi pri izbiri motorja.Preskušani 2-litrski '
'dizel s 190 KM predstavlja vrh ponudbe, ki z mo\x8djo, udobjem '
'in tudi povpre\x8dno porabo navduši predvsem pri avtocestnih '
'dolgoprogaških izzivih, v po\x8dasni mestni vožnji ter pri '
'pogostih postankih in speljevanjih pa deluje preve\x8d '
'robusten.XC 40 je s \x8dvrsto gradnjo, funkcionalno in udobno '
'kabino ter številnimi asisten\x8dnimi sistemi in '
'izstopajo\x8dim 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.Klju\x8dni tehni\x8dni podatki:- na testu '
'Volvo XC40 2.0 TD avt awd momentumMere:- dolžina: 4,4 m- '
'medosna razdalja: 2,7 m- obra\x8dalni krog: 11,4 m- oddaljenost '
'od tal: 21 cm- prtljažnik: 432 l- masa: 2.250 kgPogon:- '
'2-litrski 4-valjni bencinski motor- mo\x8d: 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/kmStroški pri 15.000 km in 5-letni uporabi:- nakupna '
'cena: 43.619 EUR- stroški finan\x8dnega 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',
  'lead': 'XC 40 je najmanjši Volvov SUV, ki se oblikovno skoraj v celotni
'
'naslanja na oba ve\x8dja predhodnika. Že samo s tem so mu vrata '
'do denarnic tistih kupcev, ki iš\x8dejo izstopajo\x8do, a hkrati '
'visoko kultivirano in pre\x8diš\x8deno dizajnersko govorico, na '
'pol odprta.',
  'published_time': '25. januar 2019 ob 15:23',
  'subtitle': 'Test novega modela',
  'title': 'Volvo XC 40 D4 AWD momentum: suvereno med najboljše v razredu'
}
```


slotech.si | Tesla


```
{
  'author': 'Matej Hu┤',
  'category': 'To¥be',
  'content': 'Elon Musk se je poravnal z ameri┤ko agencijo za trg
vrednostnih '
```

'papirjev (SEC), ki je avgusta zaÞela preiskavo zaradi sporne '
'objave na Twitterju. Musk, ki je tedaj brez utemeljitve trdil, '
'da bodo Teslo odkupili in da je financiranje po ceni 420 '
'dolarjev na delnico ¥e zagotovljeno, je s tem najprej poskrbel '
'za rast cene delnice Tesle, kar je imelo finanÞne posledice za '
'prodajalce na kratko, kasneje pa je delnica izgubila ┤e veÞ '
'vrednosti, ko se je razvedela neoprijemljivost Muskovih trditev. '
'Elon Musk bo plaÞal 20 milijonov dolarjev globe, a bo lahko '
'ostal na Þelu podjetja. Poravnava je bila priÞakovana, saj je '
'SEC Muska sodno preganjal, ko ta ni pristal na prvo poravnavo, '
'v takih postopkih pa SEC praviloma zmaga. Poravnava doloÞa, '
'da Musku ni treba priznati nobenega nezakonitega ravnanja, da '
'morata Tesla in Musk plaÞati vsak 20 milijonov dolarjev globe, '
'da Musk lahko ostane izvr┤ni direktor, ne sme pa biti v '
'naslednjih treh letih predsednik upravnega odbora. Prav tako '
'mora upravni odbor imenovati dva nova Þlana. Glavni oÞitek '
'podjetju samemu je neobstoj postopkov in kontrolnih mehanizmov '
'za komuniciranje z javnostmi prek Muskovega Twitterjevega '
'raÞuna. Z drugimi besedami: Tesla ne bi bila smela dovoliti '
'Musku, da brez odobritve in vÚdenja podjetja tvita, kar mu pade '
'na pamet. Zaradi tega bo Tesla morala uvesti postopke in '
'mehanizme glede Muskovega (oziroma Þigarkoli) komuniciranja z '
'javnostjo na dru¥benih omre¥jih. Poravnavo mora potrditi ┤e '
'sodi┤Þe.   ',
  'date': '27. apr 2019',
  'source': 'Slo-Tech',
  'time': '21:47',
  'title': 'Musk zaradi afere s tvitom oglobljen z 20 milijoni dolarjev, '
'odstopa kot prvi mo¥ upravnega odbora'
}

slotech.com | Nvidia

{
  'author': 'Jurij Kristan',
  'category': 'Grafi ├¿ne kartice',
  'content': 'Nvidia je zelo na hitro poslala v trgovine novega Geforca -
GTX '
'1650; sicer najmanj┬┤ega predstavnika arhitekture Turing. '
' ├êeprav je bilo zaradi uhajanja informacij ┬¥e nekaj dni jasno, '
'da bo zeleni tabor kmalu dal na trg ┬┤e naj┬┤ibkej┬┤o ├¿lanico '
'svoje trenutne serije grafi ├¿nih kartic, je lansiranje Geforca '
'GTX 1650 zainteresirano javnost ujelo nepripravljeno, saj '
'Nvidia predstavnikom medijev ni poslala ni ├¿esar v predhodno '
'testiranje. Malo zato, ker so ┬┤ele danes pri┬┤li posodobljeni '

```
'gonilniki, ┬┤e bolj pa zato, ker gre pa ├¿ za izdelek za spodnji '
'del trga.  Kartica GTX 1650 je zgrajena na ├¿ipu TU117, ki je '
'obtesan TU116 iz  GTXov 1660; in sicer za tretjino. To pomeni '
'896 jedrc CUDA, 128-bitno  pomnilni┬┤ko vodilo in ┬┤tiri '
'gigabajte RAMa GDDR5. Skladno z var├¿nostjo  te sorte kartic ne '
'potrebuje dodatnega napajanja, je pa nekaj dra┬¥ja,  kot je bila '
'ob nastopu GTX 1050, katere zamenjava naj bi bila: 150  dolarjev '
'brez davka v ZDA, kar bo pri nas obvezno naneslo prek 200  '
'evrov. Pri tej ceni se neposredno bode z Radeonom RX 570, ki ga '
'sode ├¿ po  prvih testih  gleda v vzvratne lu├¿i. Je pa treba '
'poleg omeniti, da gre v bistvu za  malo okrnjeno ina├¿ico  ├¿ipa '
'TU117 in da je polna, najverjetneje v obliki GTXa 1650 Ti, '
'br┬¥kone tudi ┬¥e za vogalom. ',
  'date': '27. apr 2019',
  'source': 'AnandTech',
  'time': '21:47',
  'title': 'Nvidia lansirala Geforce GTX 1650'
}
```


avto.net | BMW i3


```
{
  'cars': [
    {
      'data': [
        'NOVO vozilo',
        '2019 km',
        'elektro pogon, 75 kW ',
        'avtomatski menjalnik'
      ],
      'img':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20BMW_files/1058558_160.jpg',
      'logo':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20BMW_files/14462.gif',
      'name': 'BMW i3 120Ah  Avt.',
      'price': '39.990€'
    },
    {
      'data': [
        'NOVO vozilo',
        '10 km',
        'elektro pogon, 75 kW ',
        'avtomatski menjalnik'
      ],
      'img':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20BMW_files/1057881_160.jpg',
```

```
      'logo':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20BMW_files/14462.gif',
      'name': 'BMW i3 120Ah  Avt.',
      'price': '30.990€'
    }
  ]
}
```


avto.net | Volkswagen Arteon


```
{
  'cars': [
    {
      'data': [
        'Letnik 1.registracije:2017',
        '49688 km',
        'diesel motor, 1968 ccm, 110 kW  / 150 KM',
        'avtomatski menjalnik / tiptronic'
       ],
      'img':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20Volkswagen_files/1042892_160.
jpg',
      'logo':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20Volkswagen_files/15572.gif',
      'name': 'Volkswagen Arteon R-Line 2.0 TDI DSG-POL...',
      'price': '34.990€'
    },
    {
      'data': [
        'Letnik 1.registracije:2017',
        '34400 km',
        'diesel motor, 1968 ccm, 179 kW  / 244 KM',
        'avtomatski menjalnik / tiptronic'
      ],
      'img':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20Volkswagen_files/1085361_160.
jpg',
      'logo':
'www.Avto.net%20%20Najve%C4%8Dja%20ponudba%20Volkswagen_files/13586.gif',
      'name': 'Volkswagen Arteon 2.0 TDI Elegance 4M DS...',
      'price': '37.990€'
    }]}
```

Zaključek

Ekstrakcijo podatkov s spleta se lahko opravi preko na podlagi metod, ki jih delimo v dve katergoriji: ročne metode (regularni izrazi in XPath) in avtomatske metode (RoadRunner). V seminarski nalogi sva na različnih spletnih straneh uporabila prej omenjeni ročni metodi. Z obema metodama je bilo mogoče dobiti podatke na enak način, kar sva pokazala z enakimi JSON izhodi.

Literatura

[1] XPath Tutorial. (april 2019) [Online]. Dosegljivo: https://www.w3schools.com/xml/xpath_intro.asp
[2] Regex One. (april 2019) [Online]: https://regexone.com/
[3] Regular Expressions. (april 2019) [Online]: https://www.regular-expressions.info/
[4] V. Crescenzi, G. Mecca, P. Merialdo, “RoadRunner: Towards Automatic Data Extraction fromLarge Web Sites”. (maj 2019) [Online]. Dosegljivo: http://vldb.org/conf/2001/P109.pdf