Programming Assignment 2

Implementing structured data extraction

Jaka Kokošar, Danijel Maraž, and Toni Kocjan

Fakulteta za Računalništvo in Informatiko UL dm9929@student.uni-lj.si, jk0902@student.uni-lj.si, tk3152@student.uni-lj.si

Povzetek The article covers the work done in the scope of the second programming assignment as part of the subject web information extraction and retrieval.

Keywords: Data Extraction Retrieval XPath Regex Roadrunner

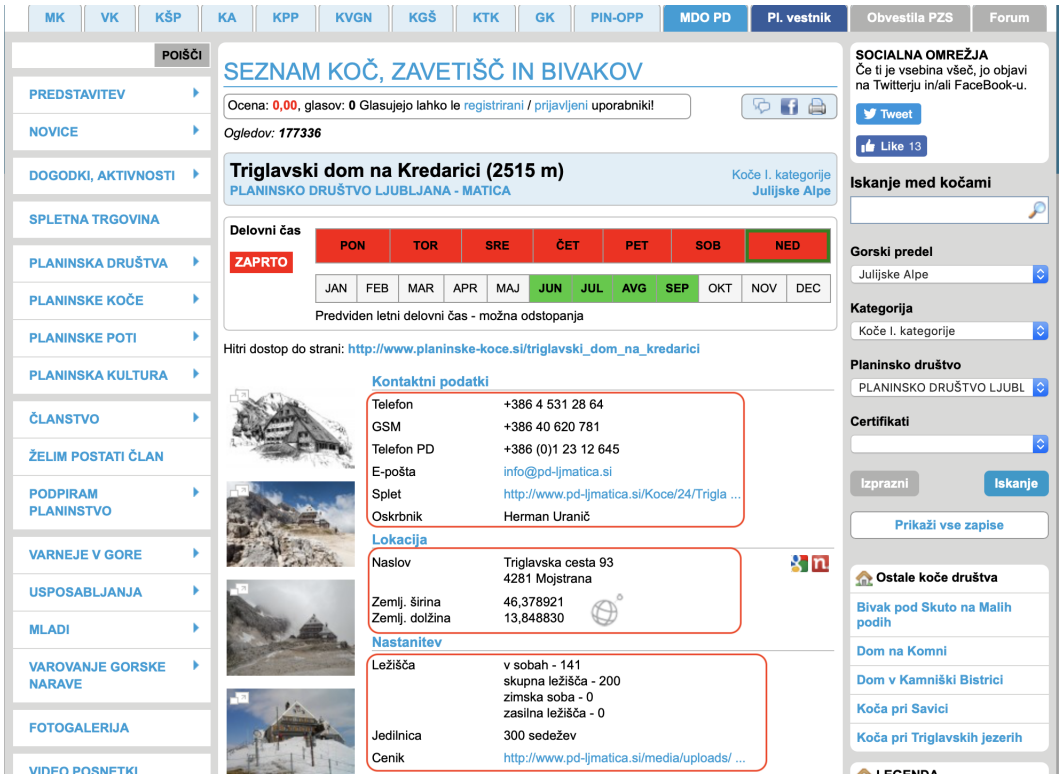
1 Introduction

After having collected the raw data with a crawler the logical next step is to convert it into a more structured format. As web pages do not have a strict shape this poses quite a challenge as any attempt of data extraction must be robust and resistant to various deformities in the html code. The report covers our attempts in implementing processes of structured data extraction in 6 different pages using basic methods such as regular expressions and xpath expressions. Alongside this an attempt was made to implement the RoadRunner algorithm which greatly simplifies the unpredictability aspect of data extraction.

2 Description of chosen web pages and data items

In addition to the mandatory websites, we selected a PZS[1] (*Planinska zveza Slovenije*) site, which is the central register of information about the operation of mountain huts in Slovenia. As shown in Figure 1, we wanted to extract information of accommodations, location and contact details of a given hut. Since there are no other resources or APIs through which we could access this information and use them for any other purpose, we are forced to obtain information directly from this site.

[1] https://www.pzs.si/



Slika 1. Data items we want to extract from pzs.si website

3 Regular expressions implementation

We defined three functions *parse_rtv_content*, *parse_overstock_content*, *parse_pzs_content*. Every function is responsible for a different page type but they are structured similarly. We extract data in three steps:

1. read the input file
2. pattern matching
3. clean the results

We only define one regular expression per page type. We extract all of the data items with a single expression in which we define all of the desired groups. After we obtain the results we clean the data of leftover html tags and strip newlines, whitespaces and tabulators. Result of the extraction process is show in appendix section A.

4 XPath implementation

4.1 jewelry

An initial Xpath expression is called on the input html which selects a *tbody* element. The children of this element represent our jewelry and other items on the page. We count the number of children and with a for loop and extra Xpath expression access each data item that interests us for each child. Any children that are malformed or do not contain the items we're interested in (such as price etc.) encounter an exception as the Xpath expression throws an error and are automatically not processed. Afterwards some minor post-processing of the acquired strings is done and the json is created.

Output

Output

4.2 cars

The items Title, SubTitle, Author, PublishedTime and Lead are all collected via seperate XPath expressions on the input html as both pages have the same structure when it comes to them. The main difference between the two pages is the structure of the article body from which we've derived our Content item. We handle this by first extracting an article body tag which contains all the contents we need. Afterwards we give the extracted element as an argument to our function *intr* which recursively iterates through the tag structure and appends all encountered text into a *nonlocal* string. The string is then briefly processed and added to our json output.

output1

output2

4.3 koce

output1

output2

5 RoadRunner like implementation

```
import numpy as np

def incmatrix(genl1,genl2):
    m = len(genl1)
    n = len(genl2)
    M = None #to become the incidence matrix
    VT = np.zeros((n*m,1), int)  #dummy variable

    #compute the bitwise xor matrix
    M1 = bitxormatrix(genl1)
    M2 = np.triu(bitxormatrix(genl2),1)

    for i in range(m-1):
        for j in range(i+1, m):
            [r,c] = np.where(M2 == M1[i,j])
            for k in range(len(r)):
                VT[(i)*n + r[k]] = 1;
                VT[(i)*n + c[k]] = 1;
                VT[(j)*n + r[k]] = 1;
                VT[(j)*n + c[k]] = 1;

                if M is None:
                    M = np.copy(VT)
                else:
                    M = np.concatenate((M, VT), 1)

                VT = np.zeros((n*m,1), int)

    return M
```

Appendix

A Regular expression outputs.


```
{
    "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 žnajdrajih in
        najbolj premijskih žrebcev je vnesel nemir, še
        preden je sploh zapeljal na parkirni prostor,
        rezerviran za šizvrnega direktorja. ",
    "Content": "Samo poglejte njegovo masko – to ogromno
        satovje z radarji na takem žpoloaju, da se ti na
        avtocesti tudi pri 120 km/h vsi šspotljivo
        umikajo, saj so čprepriani, da gre za Pahorjev
        ali Ščarev avto. Seveda, novi A6 lahko cesto in
        promet skenira s kar petimi radarji, petimi
        kamerami, činfrardeo kamero za čnoni vid,
        dvanajstimi čultrazvonimi senzorji in laserskim
        čitalnikom – lidarjem. V glavnem švojaka
        tehnologija v žslubi varnosti za fante, ki smo
        radi gledali Top Gun, Bonda in druge žmoakarja s
        finimi čigraami.Novo špoglavjeVozniki delovni
        prostor je novo poglavje digitalne dobe, z dvema
        ogromnima zaslonoma, ki tako kot šnapredneji
        telefoni dregnejo blazinice švaih prstov, kot se
        sprehajate po steklu. A še bolj se nam zdi
        pomembno, da so osnovna stikala tam, kjer jih
        čpriakujete. 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 čveČ.e vam pogled
        na Audijev spisek dodatne opreme ne odvzame volje
        do življenja, potem vsekakor toplo čpriporoamo
        nakup čzranega vzmetenja, saj dobi z njim A6 čve
```

```
        črazlinih in švozniko zelo uporabnih
        karakterjev.Enako velja za seksi člui z
        inteligentno čmatrino osvetlitvijo, pa za športno
        podvozje in vsekakor za štirikolesno krmiljenje.
        S tem postane A6 med ovinki v čobutku na volanu še
        veliko škraji in bolj agilen. Vse šnateto smo
        špreskuali v ždrubi agregata 50 TDI, ki je v
        resnici čklasini trilitrski dizel, podkrepljen z
        elektromotorjem. Ja, ta audi je mehki hibrid z
        izjemnim navorom in dovolj čmoi kadar koli in
        kjer koli. Si pa mislimo, da bo čnajveji del trga
        zadovoljil že čuinkovit dvolitrski mehki hibrid z
        čmojo 150 kilovatov.čKljuni čtehnini podatki:– na
        testu Audi A6 50 TDI quattro tiptronicMere:–
        ždolina: 4,9 m– medosna razdalja: 2,9 m– čobraalni
        krog: 12,1 m– žprtljanik: 530 l– masa: 1.900
        kgPogon:– trilitrski šestvaljni dizelski motor–
        čmo: 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/škmStroki pri 15.000
        km in 5–letni uporabi:– nakupna cena: 69.080 EUR–
        šstroki čfinannega lizinga: 4.463 EUR/5 let–
        šstroki registracije: 10.829 EUR/5 let– šstroki
        žvzdrevanja: 1.926 EUR/5 let– šstroki goriva:
        6.702 EUR/75.000 km– šstroek 1 kompleta
        pnevmatik: 716 EUR– vrednosti po 5 letih po
        Eurotaxu: 33.964 EUR– šstroki skupaj: 1.001
        EUR/mesec"
}

{
    "Title": "Volvo XC 40 D4 AWD momentum: suvereno med
        šnajbolje v razredu",
    "SubTitle": "Test novega modela",
    "Author": "Miha Merljak",
    "PublishedTime": "25. januar 2019 ob 15:23",
```

"Lead": "XC 40 je šnajmanji Volvov SUV, ki se oblikovno skoraj v celotni naslanja na oba čveja predhodnika. Že samo s tem so mu vrata do denarnic tistih kupcev, ki ščiejo čizstopajoo, a hkrati visoko kultivirano in čščpreieno dizajnersko govorico, na pol odprta.",
"Content": "Volvo se je žnijih srednjih razredov v preteklosti izogibal ali pa je vanje vstopal z zelo šninimi produkti, ki niso pustili čvejega žtrnega čpeata. V primeru XC 40 ni žteko napovedati, da bo ta tradicija prekinjena. Ponuja čnamre visoko kakovost čkonne izdelave in v kabini čodlino špremiljeno funkcionalnost ter na dotik prijetne materiale. Še posebej hvalimo število, iznajdljivost in velikost črazlinih odlagalnih prostorov ter široke, čvrste in zelo udobne žsedee. Intuitivno in enostavno člogino je upravljanje z velikim vmesnikom, ki z čvefunkcijskim zaslonom na dotik kraljuje na z roko lahko dostopnem mestu na sredinski armaturi. čRazoaranj ne bo niti v velikosti in uporabnosti žprtljanega prostora, ki s 460 litri prostornine sicer ni med čvejimi v razredu, a se v šuporabnikem smislu odkupi z dobro urejenostjo ter domiselnimi šreitvami pregrajevanja.Foto: David ŠavliXC 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 špreskuanega 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 čraunati na manj udobno žvonjo čez črazline asfaltne grbine. Podoben razmislek velja opraviti tudi pri izbiri motorja.šPreskuani 2−litrski dizel s 190 KM predstavlja vrh ponudbe, ki z čmojo, udobjem in tudi čpovpreno porabo šnavdui predvsem pri avtocestnih šdolgoprogakih

```
        izzivih , v čpoasni mestni žvonji ter pri pogostih
        postankih in speljevanjih pa deluje čpreve
        robusten.XC 40 je s čvrsto gradnjo , funkcionalno
        in udobno kabino ter številnimi časistennimi
        sistemi in čizstopajoim skandinavskim dizajnom v
        špremiljenem trenutku vstopil na trg modnih
        mestnih terencev , v katerem se brez ene same
        sence dvoma suvereno postavi med žnajdraje in
        najbolj premijske v mestu.čKljuni čtehnini
        podatki:- na testu Volvo XC40 2.0 TD avt awd
        momentumMere:- ždolina: 4,4 m- medosna razdalja:
        2,7 m- čobraalni krog: 11,4 m- oddaljenost od
        tal: 21 cm- žprtljanik: 432 l- masa: 2.250
        kgPogon:- 2-litrski 4-valjni bencinski motor-
        čmo: 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/škmStroki pri 15.000 km in
        5-letni uporabi:- nakupna cena: 43.619 EUR-
        šstroki čfinannega leasinga: 3.268 EUR/5 let-
        šstroki registracije: 8.701 EUR/5 let- šstroki
        žvzdrevanja: 2.320 EUR/5 let- šstroki goriva:
        6.190 EUR/75.000 km- šstroek 1 kompleta
        pnevmatik: 923 EUR- vrednosti po 5 letih po
        Eurotaxu: 18.886 EUR- šstroki skupaj: 774
        EUR/mesec"
}
```


Listing 1.1. Results for rtvslo website


```
{
    "listing": [
        {
            "Title": "10-kt. Seven Diamond Ladies Heart
                Ring (0.08 TW)",
            "listPrice": "$149.00",
            "Price": "$69.99",
            "Saving": "$79.01",
```

```
    "SavingPercent": "53%",
    "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."
},
{
    "Title": "10-Kt. Diamond Ring (.25 TW)",
    "listPrice": "$250.00",
    "Price": "$74.90",
    "Saving": "$175.10",
    "SavingPercent": "70%",
    "Content": "Nineteen round diamonds accent
       this 10-karat yellow gold ring with
       filigree accents."
},
{
    "Title": "10-kt. Pearl and Diamond Butterfly
       Earrings",
    "listPrice": "$149.00",
    "Price": "$42.99",
    "Saving": "$106.01",
    "SavingPercent": "71%",
    "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."
},
{
    "Title": "14-kt. Diamond 'S' Tennis Bracelet
       (2.00 TW)",
    "listPrice": "$1,539",
    "Price": "$499.99",
    "Saving": "$100.00",
```

```
        "SavingPercent": "55%",
        "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."
    },
    {
        "Title": "14-kt. White Gold, Pearl and
            Diamond Ring",
        "listPrice": "$419.99",
        "Price": "$149.99",
        "Saving": "$270.00",
        "SavingPercent": "64%",
        "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."
    },
    {
        "Title": "14-kt. Gold Diamond Present Future
            Pendant (.25TW)",
        "listPrice": "$299.00",
        "Price": "$149.99",
        "Saving": "$149.01",
        "SavingPercent": "49%",
        "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)."
    },
    {
        "Title": "14-kt. Diamond Solitaire Pendant
            (.33 TW)",
```

```
    "listPrice": "$1,019",
    "Price": "$319.99",
    "Saving": "$700.00",
    "SavingPercent": "68%",
    "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."
},
{
    "Title": "14–kt. Diamond Solitaire Earrings
       (0.33 TW)",
    "listPrice": "$639.99",
    "Price": "$199.99",
    "Saving": "$440.00",
    "SavingPercent": "68%",
    "Content": "Dazzle your way into her heart,
       with these classic diamond solitaire
       earrings. Two brilliant–cut diamonds
       (0.33 total carat weight, G–H color, I–1
       to I–2 clarity) are set in four prongs of
       14–karat white gold."
},
{
    "Title": "14–kt. Diamond Cross Pendant (.06
       TW)",
    "listPrice": "$305.00",
    "Price": "$119.99",
    "Saving": "$185.01",
    "SavingPercent": "60%",
    "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)."
},
{
```

```
        "Title": "14-kt. Diamond Solitaire Stud
            Earrings (.50 TW)",
        "listPrice": "$999.99",
        "Price": "$359.99",
        "Saving": "$640.00",
        "SavingPercent": "64%",
        "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."
    },
    {
        "Title": "14-kt. Cultured Pearl Diamond
            Earrings",
        "listPrice": "$508.99",
        "Price": "$179.99",
        "Saving": "$329.00",
        "SavingPercent": "64%",
        "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."
    },
    {
        "Title": "14-kt. Diamond 7.5-8 mm Pearl
            Pendant",
        "listPrice": "$196.99",
        "Price": "$69.99",
        "Saving": "$127.00",
        "SavingPercent": "64%",
        "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."
        },
        {
            "Title": "14-kt. Diamond Solitaire Earrings
                (.50 TW)",
            "listPrice": "$1,369",
            "Price": "$409.99",
            "Saving": "$960.00",
            "SavingPercent": "70%",
            "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."
        }
    ]
}
{
    "listing": [
        {
            "Title": "14-kt. Green Jade Hoops",
            "listPrice": "$90.00",
            "Price": "$46.99",
            "Saving": "$43.01",
            "SavingPercent": "47%",
            "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."
        },
        {
            "Title": "14-kt. Jade Doughnut Pendant",
            "listPrice": "$150.00",
            "Price": "$48.99",
            "Saving": "$101.01",
```

```
        "SavingPercent": "67%",
        "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."
    },
    {
        "Title": "14-kt. Charcoal Jade and Ruby
            Elephant Pendant",
        "listPrice": "$100.00",
        "Price": "$28.99",
        "Saving": "$71.01",
        "SavingPercent": "71%",
        "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."
    },
    {
        "Title": "14-kt. Carved Lavender Jade
            Earrings",
        "listPrice": "$80.00",
        "Price": "$39.99",
        "Saving": "$40.01",
        "SavingPercent": "50%",
        "Content": "Luscious 8-mm lavender jade
            balls, carved with intricate Asian style,
            dangle from a 14-karat yellow gold French
            hook."
    },
    {
        "Title": "14-kt. Jade Cross Pendant",
        "listPrice": "$150.00",
        "Price": "$49.99",
```

```
    "Saving": "$100.01",
    "SavingPercent": "66%",
    "Content": "Green jade and gold create this
       beautiful cross pendant. Cylindrical bars
       of green jade feature caps and center of
       14-karat yellow gold."
},
{
    "Title": "14-kt. Multicolored Jade Earrings",
    "listPrice": "$375.00",
    "Price": "$99.99",
    "Saving": "$275.01",
    "SavingPercent": "73%",
    "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."
},
{
    "Title": "14-kt. Multicolored Jade Ring",
    "listPrice": "$250.00",
    "Price": "$56.99",
    "Saving": "$193.01",
    "SavingPercent": "77%",
    "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."
},
{
    "Title": "14-kt. Onyx and Ruby Elephant
       Pendant",
    "listPrice": "$100.00",
    "Price": "$35.99",
```

```
            "Saving": "$64.01",
            "SavingPercent": "64%",
            "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."
        }
    ]
}
```

Listing 1.2. Results for jewlary website

```
{
    "Telefon": "+386 4 531 28 64",
    "GSM": "+386 40 620 781",
    "TelefonPD": "+386 (0)1 23 12 645",
    "eMail": "info@pd-ljmatica.si",
    "Splet":
        "http://www.pd-ljmatica.si/Koce/24/Triglavski-dom-na-Kredarici",
    "Oskrbnik": "Herman čUrani",
    "Naslov": "Triglavska cesta 93, 4281 Mojstrana",
    "ZemljepisnaSirina": "46,378921",
    "ZemljepisnaDolzina": "13,848830",
    "Lezisca": "v sobah - 141, skupna žščleia - 200,
        zimska soba - 0, zasilna žščleia - 0",
    "Jedilnica": "300 žsedeev",
    "Cenik":
        "http://www.pd-ljmatica.si/media/uploads/ceniki/cenik_kredarica_2017.pdf"
}
{
    "Telefon": "+386 5 7264 529",
    "GSM": "+386 51 373 900",
    "TelefonPD": "+386 (0)41 688 696",
    "eMail": "vojkova.koca@pdpostojna.si",
    "Splet":
        "http://www.pdpostojna.si/index.php?option=com_content&amp;view=article&a
    "Oskrbnik": "Lilijana ččValeni",
```

```
    "Naslov": "Nanos 11, 5271 Vipava",
    "ZemljepisnaSirina": "45,772063",
    "ZemljepisnaDolzina": "14,053140",
    "Lezisca": "v sobah - 4, skupna žščleia - 48, zimska
        soba - 0, zasilna žščleia - 0",
    "Jedilnica": "82 žsedeev",
    "Cenik": "http://www.vozni-red.si/"
}
```


Listing 1.3. Results for PZS website