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

Abstract. 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

3 Regular expressions implementation

3.1 jewelry

output1

output2

3.2 cars

output1

output2

3.3 koce

output1

output2

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 1


```
{
    "title": "10-kt. Seven Diamond Ladies Heart Ring (0.08 TW)",
    "content": "This ladies fashion ring dazzles\nwith hearts and diamonds. The
    "listprice": "$149.00",
    "price": "$69.99",
    "saving": "$79.01",
    "savprcnt": "(53%)"
}
{
    "title": "10-Kt. Diamond Ring (.25 TW)",
    "content": "Nineteen round diamonds accent this 10-karat yellow gold ring wi
    "listprice": "$250.00",
    "price": "$74.90",
    "saving": "$175.10",
    "savprcnt": "(70%)"
}
{
    "title": "10-kt. Pearl and Diamond Butterfly Earrings",
    "content": "Perfectly proportioned 5.5- to\n6-mm cultured pearls on 10-karat
    "listprice": "$149.00",
    "price": "$42.99",
    "saving": "$106.01",
    "savprcnt": "(71%)"
}
{
    "title": "14-kt. Diamond 'S' Tennis Bracelet (2.00 TW)",
    "content": "Invest in a swirl of light with\nthis diamond 'S' tennis bracele
    "listprice": "$1,539.99",
    "price": "$499.99",
```

```
    "saving": "$1,040.00",
    "savprcnt": "(67%)"
}
{
    "title": "10-kt. Diamond Band Fashion Ring (.11 TW)",
    "content": "Crafted in white and yellow gold,\nthis ring displays a band of
    "listprice": "$179.99",
    "price": "$79.99",
    "saving": "$100.00",
    "savprcnt": "(55%)"
}
{
    "title": "14-kt. White Gold, Pearl and Diamond Ring",
    "content": "Show your romantic side with this\n14-karat diamond and pearl ri
    "listprice": "$419.99",
    "price": "$149.99",
    "saving": "$270.00",
    "savprcnt": "(64%)"
}
{
    "title": "14-kt. Gold Diamond Present Future Pendant (.25TW)",
    "content": "Designed with three large, sparkling\ndiamonds to represent past
",
    "listprice": "$299.00",
    "price": "$149.99",
    "saving": "$149.01",
    "savprcnt": "(49%)"
}
{
    "title": "14-kt. Diamond Solitaire Pendant (.33 TW)",
    "content": "In this simple, yet elegant pendant,\na round brilliant diamond
    "listprice": "$1,019.99",
    "price": "$319.99",
    "saving": "$700.00",
    "savprcnt": "(68%)"
}
{
    "title": "14-kt. Diamond Solitaire Earrings (0.33 TW)",
    "content": "Dazzle your way into her heart,\nwith these classic diamond soli
    "listprice": "$639.99",
    "price": "$199.99",
    "saving": "$440.00",
    "savprcnt": "(68%)"
}
{
```

```
    "title": "14-kt. Diamond Cross Pendant (.06 TW)",
    "content": "Over a cleanly sculpted Roman\ncross of 14-karat white gold drap
    "listprice": "$305.00",
    "price": "$119.99",
    "saving": "$185.01",
    "savprcnt": "(60%)"
}
{
    "title": "14-kt. Diamond Solitaire Stud Earrings (.50 TW)",
    "content": "Every jewelry collection needs\na classic pair of diamond solita
    "listprice": "$999.99",
    "price": "$359.99",
    "saving": "$640.00",
    "savprcnt": "(64%)"
}
{
    "title": "14-kt. Cultured Pearl Diamond Earrings",
    "content": "Create an elegant appearance with\nthese pearl and diamond stud
    "listprice": "$508.99",
    "price": "$179.99",
    "saving": "$329.00",
    "savprcnt": "(64%)"
}
{
    "title": "14-kt. Diamond 7.5-8 mm Pearl Pendant",
    "content": "Add a classic to your jewelry\ncollection with this 14-karat gol
    "listprice": "$196.99",
    "price": "$69.99",
    "saving": "$127.00",
    "savprcnt": "(64%)"
}
{
    "title": "14-kt. Diamond Solitaire Earrings (.50 TW)",
    "content": "This earring set has two brilliant-cut\ndiamonds (0.50 total car
    "listprice": "$1,369.99",
    "price": "$409.99",
    "saving": "$960.00",
    "savprcnt": "(70%)"
}
{
    "title": "14-kt White Gold Diamond Band (0.50 TW)",
    "content": "Crafted of 14-karat white gold,\nthis stylish ring features a br
    "listprice": "$1,635.00",
    "price": "$609.99",
    "saving": "$1,025.01",
```

```
    "savprcnt": "(62%)"
}
```


Output 2


```
{
    "title": "14-kt. Green Jade Hoops",
    "content": "Hoops of cool green jade rest\nbetween 14-karat yellow gold endp
    "listprice": "$90.00",
    "price": "$46.99",
    "saving": "$43.01",
    "savprcnt": "(47%)"
}
{
    "title": "14-kt. Jade Doughnut Pendant",
    "content": "The 25-mm disk hangs delicately\nfrom a 14-karat gold chain. The
    "listprice": "$150.00",
    "price": "$48.99",
    "saving": "$101.01",
    "savprcnt": "(67%)"
}
{
    "title": "14-kt. Charcoal Jade and Ruby Elephant Pendant",
    "content": "Carved of rich dark grey jade,\nthis elephant pendant has 14-kar
    "listprice": "$100.00",
    "price": "$28.99",
    "saving": "$71.01",
    "savprcnt": "(71%)"
}
{
    "title": "14-kt. Carved Lavender Jade Earrings",
    "content": "Luscious 8-mm lavender jade balls, carved with intricate Asian s
    "listprice": "$80.00",
    "price": "$39.99",
    "saving": "$40.01",
    "savprcnt": "(50%)"
}
{
    "title": "14-kt. Jade Cross Pendant",
    "content": "Green jade and gold create this\nbeautiful cross pendant. Cylinc
    "listprice": "$150.00",
    "price": "$49.99",
    "saving": "$100.01",
    "savprcnt": "(66%)"
}
```

```
{
    "title": "14-kt. Multicolored Jade Earrings",
    "content": "A delicate wrapping of 14-karat\nyellow gold wire holds six 6 x
    "listprice": "$375.00",
    "price": "$99.99",
    "saving": "$275.01",
    "savprcnt": "(73%)"
}
{
    "title": "14-kt. Multicolored Jade Ring",
    "content": "A delicate wrapping of 14-karat\nyellow gold wire holds six 6 x
    "listprice": "$250.00",
    "price": "$56.99",
    "saving": "$193.01",
    "savprcnt": "(77%)"
}
{
    "title": "14-kt. Onyx and Ruby Elephant Pendant",
    "content": "Carved of rich black onyx, this\nelephant pendant has 14-karat y
    "listprice": "$100.00",
    "price": "$35.99",
    "saving": "$64.01",
    "savprcnt": "(64%)"
}
```


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
```

References