IEPS: Data extraction from the Web

Avtorja: Julijan Jug, Jaka Jenko

I. Introduction

For the second project of the Web Information Extraction and Retrieval course, we had to implement three different web data extraction approaches. The approaches include regular expressions, XPath expressions, and the implementation of an automatic web extraction algorithm. We used these approaches on 3 different web sites: RTV SLO, Overstock, and Mimovrste.

II. Extracted data items

1) RTV SLO:

- Author
- Title
- Published time
- Subtitle
- Lead
- Content

2) Overstock:

For all items we extracted:

- Title
- Price
- List price
- Saving
- Saving percent
- Content

3) Mimovrste:

We selected two web pages from the Mimovrste web site and saved it for offline viewing using chromes built-in saving feature. both pages include a list of items (graphic cards and processors). Extracted data items are listed below. The listed price is an optional data item that is not always present. For all items we extracted also presented in figure 1:

- Title
- Price
- List price
- Description
- Availability

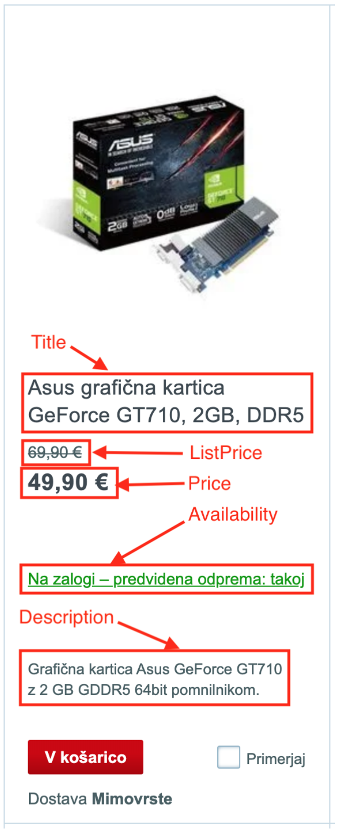
Title
Asus grafična kartica GeForce GT710, 2GB, DDR5
69,90 €
ListPrice
49,90 €
Price
Availability
Na zalogi – predvidena odprema: takoj
Description
Grafična kartica Asus GeForce GT710 z 2 GB GDDR5 64bit pomnilnikom.
V košarico
Primerjaj
Dostava Mimovrste

Figure 1: Extracted data items from Mimovrste.

III. Extraction approaches

A. Regular expressions

1) RTV SLO:

- Author:
  ```
  <div class=\"author-timestamp\">\s+<strong>(.*)
      <\/strong>
  ```
- Title:
  ```
  <h1>(.*)<\/h1>\s*<div class=\"subtitle\">
  ```
- Published time:
  ```
  <div class=\"author-timestamp\">\s+<strong>.*
      <\/strong>\|\s+(.*:[0-9]{2})[\s\t\S]*
  ```
- Subtitle:
  ```
  <div class=\"subtitle\">(.*)<\/div>
  ```
- Lead:
  ```
  <p class=\"lead\">(.*)<\/p>
  ```
- Content:
  ```
  (?<=<article class=\"article\">).*?((<p.*?>)(.*?)
      (<\/p.*?>).*?)*(?=<\/article>)
  ```

Content regular expression captures all text inside p tags in an article tag. We then iterate through all of the returned results and just concatenate them to get the final content.

2) Overstock:

For extracting Overstock items and its data items we used a single regular expression. Data items correspond to the match groups in the following order: title, listPrice, price, saving, savingPercent, content.

```
<td valign=\"top\">\s*<a.*>\s*<b>(.*)<\/b>[\s\S|.]*?
<td align=\"left\" nowrap=\"nowrap\">\s*<s>(.*)<\/s>\s*<\/td>
[\s\S|.]*?<span class=\"bigred\">\s*<b>(.*)<\/b>
[\s\S|.]*?<span class=\"littleorange\">
([$€]\s*[0-9\.,]+) \((.*)\)<\/span>[\s\S|.]*?
<span class=\"normal\">([\s\S|.]*?)<br>
```

3) Mimovrste:

For extracting Overstock items and its data items we used a single regular expression. Data items correspond to the match groups in the following order: title, - , listPrice, - , price, availability, description.

```
class=\"lay-block con-no-decoration\">(.*)</a>[\s\S|.]*?
<(del|span) class=\"lst-product-item-price--retail\">(.*)
<\/(del|span)>[\s\S|.]*?
<span class=\"lst-product-item-price-value\">
[\s\S]*?\t\t\t\t\t\t    (.*)\s*<\/span>[\s\S|.]*?
```

```
<p class=\"lst-product-item-availability
con-text--availability text-collapse\">(.*?) - [\s\S|.]*?
<div class=\"lst-product-item-description \"><p>(.*?)<\/p>
```

B. XPath expressions

XPath is an XML query language. It enables us to get and compute data from XML documents.

1) RTV SLO:
Used "XPath"s for gathering data start with '//*[@id="main-container"]/div[3]/div/' and then continue with:

- Author - 'div[1]/div[1]/div/text()'
- Title - 'header/h1/text()'
- Published time - 'div[1]/div[2]/text()[1]'
- Subtitle - 'header/div[2]/text()'
- Lead - 'header/p//text()'
- Content - 'div[2]/article/p/text()'

The only other thing we had to do, used a regular expression ('[^\S]') to clean the "Published time".

2) Overstock:
Used "XPath"s for gathering data start with '/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/' and then continue with:

- Titles - 'a/b/text()'
- Prices - 'table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s/text()'
- List prices - 'table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b/text()'
- Savings and Saving percents - 'table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text()'
- Contents - 'table/tbody/tr/td[2]/span/text()'

We zip all of the above data and then iterate over it. We also had to split the "Saving" and "Saving percent". We did that with the help of regular expression ('[\S]+').

3) Mimovrste:
Used "XPath"s for gathering data start with '/html/body/div[3]/div/div[2]/main/section/section/div/article/' and then continue with:

- Titles - 'div/div/div[1]/h3/a/text()'
- Prices - 'div/div/div[2]/span/text()'
- Descriptions - 'div/div/div[6]/div[1]/p//text()'
- Availabilitys - 'div/div/div[3]/div/div/p/text()'
- Article IDs - '@id'

We zip all of the above data and then iterate over it. For each item, we also need to get the list price. We get it one by one because there is a possibility that one item doesn't have it. We get the list price with the help of the next XPath.

```
xpath = '//*[@id="' + articleId + '"]
    /div/div/div[2]/del/text()'

listPrice = tree.xpath('concat(
    substring(
        ' + xpath + ',
        1,
        number(not(not(' + xpath + '))) * 100),

    substring(
        'None',
        1,
        number(not(' + xpath + ')) * 4)
    )')
```

That XPath checks if the list price occurs at a specific article. If it does it returns the list price if not it returns "None". It is possible to use "if then else" in XPath, but python library doesn't support it, that's why we had to do it like that.
With the help of regular expressions we also had to clean price ('(\t)\s*(\d*(.\d)*,*\d* €)') and availability ('(.*?)(–)').

C. Automatic web extraction

For automatic web extraction, we implemented Webstemmer like extractor with the help of Webstemmer description[1].
The procedure of our program is the following:

- Firstly we parsed the HTML with BeautifulSoup to clean and fix the HTML. We removed "script", "img", "iframe", "style", "a" and "svg" tags.
  (We did that because those tags usually not provide much useful information and can be different from page to page which can cause false positives in the wrapper.)
- Then we extracted the paths to text areas, so we were able to find the paths to the found blocks.
- With the help of "difflib" we found all the differences between two files.
- We compared the differences in blocks and removed all of the blocks that had a similarity of more than 0.6 based on "difflib.SequenceMatcher" method.
  (Removing of similar blocks is necessary because only different blocks will show us what is important on the page.)
- For each block, we also calculated the "main score" which tells us how important the block is.
  (With it we can find the main text.)
- , In the end, we combined all of the data for each block and got the path to the block. We also added the fist 30 characters of text just to show what the algorithm found.

IV. Conclusion

We have successfully implemented all of the three extraction approaches. The easiest to implement was the XPath approach because it's easy to find the XPath of the text block, but there were some problems where some data was missing and not formatted correctly where we used the Regular expression to fix that. With the Regular expressions, we had the most flexibility and even though it was a bit harder to get to the correct Regular expression we were able to better extract data even where some of it were missing. With the Webstemmer like implementation, it was the easiest to get out all of the data but the problem was you don't know exactly what you are getting. s

References

[1] "Webstemmer - How it works?" http://www.unixuser.org/~euske/python/webstemmer/howitworks.html#extract, [Online; accessed 1-May-2020].

Appendix A
Wrapper - RTV SLO

```
[
    {
        "ScoreDiff": 0.6176470588235294,
        "ScoreMain": 0.5352941176470588,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div",
        "File1": "| 28. december 2018 ob",
        "File2": "| 25. januar 2019 ob 15"
    },
    {
        "ScoreDiff": 0.37398373983739835,
        "ScoreMain": 1.4022764227642275,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/header/h1",
        "File1": "Audi A6 50 TDI quattro:",
        "File2": "Volvo XC 40 D4 AWD momen"
    },
    {
        "ScoreDiff": 0.3442622950819672,
        "ScoreMain": 1.5475409836065575,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div/figure/ul/li",
        "File1": "Zaradi videza se mu na",
        "File2": "XC40 se pohvali z robu"
    },
    {
        "ScoreDiff": 0.39080459770114945,
        "ScoreMain": 0.6579310344827587,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/header/div",
        "File1": "Test nove generacije",
        "File2": "Ekipa Avtomobilnosti na"
    },
    {
        "ScoreDiff": 0.3,
        "ScoreMain": 0.7,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div/article/p/p/br/
            br/br/br/p/br/br/br/br/br/br/br/br/br/
            p/br/br/div/div/div/div/p",
        "File1": "Test novega modela",
        "File2": "Razvoj avtomobilske var"
    },
    {
        "ScoreDiff": 0.25806451612903225,
        "ScoreMain": 1.4838709677419353,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/article/p/p",
        "File1": "- na testu Audi A6 50 T",
        "File2": "Mere:"
    },
    {
        "ScoreDiff": 0.68,
        "ScoreMain": 0.33279999999999993,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/header/div",
        "File1": "Test nove generacije",
        "File2": "Test novega modela"
    },
    {
        "ScoreDiff": 0.009174311926605505,
        "ScoreMain": 7.015045871559633,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/header/p",
        "File1": "To je novi audi A6. V ra",
        "File2": "XC 40 je najmanjši Volvo"
    },
    {
        "ScoreDiff": 0.48,
        "ScoreMain": 0.45760000000000006,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div/article/p/p/br/
            br/br/br/p/br/br/br/br/br/br/br/br/br/p/br/
            br/br/br/br/br/div/div/div/figure/figcaption",
        "File1": "MMC RTV SLO",
        "File2": "Foto: David Šavli"
    },
    {
        "ScoreDiff": 0.4583333333333333,
        "ScoreMain": 0.45500000000000007,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div/article/p/p/br/br/br/
            br/p/br/br/br/br/br/br/br/br/p/br/br/br/br/br/
            br/div/div/div/figure/figcaption",
        "File1": "MMC RTV SLO",
        "File2": "Foto: David Šavli"
    },
    {
        "ScoreDiff": 0.411214953271028,
        "ScoreMain": 3.4856074766355136,
        "BlockFeature": null,
        "File1": "Audi A6 je avtomobil z",
        "File2": "XC40 se v mestu odličn"
    },
    {
        "ScoreDiff": 0.40186915887850466,
        "ScoreMain": 3.5648598130841127,
        "BlockFeature": null,
        "File1": "Audi A6 je avtomobil",
        "File2": "XC40 se v mestu odlično"
    },
    {
        "ScoreDiff": 0.52,
        "ScoreMain": 0.5184,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/header/div",
        "File1": "Test nove generacije",
        "File2": "Obuditev legende"
    },
    {
        "ScoreDiff": 0.5806451612903226,
        "ScoreMain": 0.5367741935483871,
        "BlockFeature": "html/body/div/div/div/div/form/
            div/div/div/div/div/div/div",
        "File1": "28. december 2018 ob 08:",
        "File2": "25. januar 2019 ob 15:23"
    },...
```

Appendix B
Wrapper - Overstock

```
[
    {
        "ScoreDiff": 0.8831168831168831,
        "ScoreMain": 0.15718763994626067,
        "BlockFeature": "html/body/input/table/tbody/
            tr/td/table/tbody/tr/tr/table/tbody/tr/td/
            br/br/br/br/br/br/br/br/br/td/br/br/table/
            tbody/tr/tr/td/table/tbody/tr/td/table/
            tbody/tr/td/br/tr/td/br/tr/td/br/tr/td/br/
            tr/td/br/tr/td/br/tr/td/br/tr/td/br/tr/td/
            br/tr/td/br/table/tbody/tr/td/table/tbody/
            tr/td/span",
        "File1": "$185",
        "File2": "$64."
    },
    {
        "ScoreDiff": 0.5504587155963303,
        "ScoreMain": 0.8680797216070864,
        "BlockFeature": null,
        "File1": "Show you",
        "File2": "Carved o"
    },
    {
        "ScoreDiff": 0.8461538461538461,
        "ScoreMain": 0.20689655172413793,
        "BlockFeature": "html/body/input/table/tbody/
            tr/td/table/tbody/tr/tr/table/tbody/tr/td/
            br/br/br/br/br/br/br/br/br/td/br/br/table/
            tbody/tr/tr/td/table/tbody/tr/td/table/
            tbody/tr/td/br/tr/td/br/tr/td/br/tr/td/
            br/tr/td/br/tr/td/br/tr/td/br/tr/td/br/
            table/tbody/tr/td/table/tbody/tr/td/span",
        "File1": "$700",
        "File2": "$275"
    },
    {
        "ScoreDiff": 0.9552238805970149,
        "ScoreMain": 0.05249613998970666,
        "BlockFeature": "html/body/input/table/tbody/
        tr/td/table/tbody/tr/tr/table/tbody/tr/td/br/
        br/br/br/br/br/br/br/br/td/br/br/table/tbody/
        tr/tr/td/table/tbody/tr/td/table/tbody/tr/td/
        br/tr/td/br/tr/td/br/tr/td/br/tr/td/br/tr/td/
        br/table/tbody/tr/td/table/tbody/tr/td/span/b",
        "File1": "$14",
        "File2": "$49"
    },
    {
        "ScoreDiff": 0.8717948717948718,
        "ScoreMain": 0.1724137931034483,
        "BlockFeature": "html/body/input/table/tbody/
            tr/td/table/tbody/tr/tr/table/tbody/tr/td/br/
            br/br/br/br/br/br/br/br/td/br/br/table/tbody/
            tr/tr/td/table/tbody/tr/td/table/tbody/tr/td/
            br/tr/td/br/tr/td/br/tr/td/br/tr/td/br/table/
            tbody/tr/td/table/tbody/tr/td/span",
        "File1": "$100",
        "File2": "$101"
    },
    {
        "ScoreDiff": 0.5420560747663551,
        "ScoreMain": 0.8685143409603611,
        "BlockFeature": null,
        "File1": "This lad",
        "File2": "Hoops of"
    },
    {
        "ScoreDiff": 0.7941176470588235,
        "ScoreMain": 0.24847870182555787,
        "BlockFeature": "html/body/input/table/tbody/
            tr/td/table/tbody/tr/tr/table/tbody/tr/
            td/br/br/br/br/br/br/br/br/br/td/br/br/
            table/tbody/tr/tr/td/table/tbody/tr/td/
            table/tbody/tr/td/br/tr/td/br/tr/td/br/
            tr/td/br/tr/td/br/tr/td/br/tr/td/br/tr/
            td/br/table/tbody/tr/td/table/tbody/tr/td/s",
        "File1": "$1,0",
        "File2": "$375"
    },
    {
        "ScoreDiff": 0.8955223880597015,
        "ScoreMain": 0.12249099330931547,
        "BlockFeature": "html/body/input/table/tbody
        /tr/td/table/tbody/tr/tr/table/tbody/tr/td/
        br/br/br/br/br/br/br/br/br/td/br/br/table/
        tbody/tr/tr/td/table/tbody/tr/td/table/tbody/
        tr/td/br/tr/td/br/tr/td/br/tr/td/br/tr/td/br/
        tr/td/br/table/tbody/tr/td/table/tbody/tr/td/span/b",
        "File1": "$14",
        "File2": "$28"
    },
    {
        "ScoreDiff": 0.5471698113207547,
        "ScoreMain": 0.811971372804164,
        "BlockFeature": null,
        "File1": "Over a c",
        "File2": "Carved o"
    },
    {
        "ScoreDiff": 0.5663716814159292,
        "ScoreMain": 0.8822093378089717,
        "BlockFeature": null,
        "File1": "Designed",
        "File2": "Green ja"
    },...
```

Appendix C
Wrapper - Mimovrste

```
[
    {
        "ScoreDiff": 0.6060606060606061,
        "ScoreMain": 0.20332355816226783,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/label/input/span/span",
        "File1": "GeForce GT 10",
        "File2": "AMD Radeon Ve"
    },
    {
        "ScoreDiff": 0.8095238095238095,
        "ScoreMain": 0.06451612903225806,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/section/section/div/
            article/article/article/div/div/div/del",
        "File1": "449,90 €",
        "File2": "699,99 €"
    },
    {
        "ScoreDiff": 0.5217391304347826,
        "ScoreMain": 0.2005610098176718,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/span/span/span/strong",
        "File1": "Grafične kartice",
        "File2": "Procesorji"
    },
    {
        "ScoreDiff": 0.7804878048780488,
        "ScoreMain": 0.07081038552321006,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/section/section/div/
            article/article/article/article/
            article/article/article/article/
            article/article/div/div/div/del",
        "File1": "57,99 €",
        "File2": "275,58 €"
    },
    {
        "ScoreDiff": 0.7142857142857143,
        "ScoreMain": 0.0967741935483871,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/section/section/div/
            article/article/div/div/div/del",
        "File1": "179,99 €",
        "File2": "538,81 €"
    },
    {
        "ScoreDiff": 0.01532567049808429,
        "ScoreMain": 4.383388950685947,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/section/section/div/
            article/article/article/article/
            div/div/div/div/p",
        "File1": "Tu je grafična k",
        "File2": "Procesor AMD Ryz"
    },
    {
        "ScoreDiff": 0.75,
        "ScoreMain": 0.10483870967741936,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/div/div/span",
        "File1": "1512.64",
        "File2": "899"
    },
    {
        "ScoreDiff": 0.8780487804878049,
        "ScoreMain": 0.04130605822187254,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/div/div/div/span",
        "File1": "154 artiklov",
        "File2": "35 artiklov"
    },
    {
        "ScoreDiff": 0.9090909090909091,
        "ScoreMain": 0.032258064516129045,
        "BlockFeature": "html/body/div/header/
            div/div/div/div/div/div/div/section/
            div/div/nav/aside/hr/hr/div/div/div/
            div/div/div/div/div/div/label/div/
            div/div/div/div/div/label/div/div/
            div/div/main/section/section/div/
            article/article/article/article/
            article/article/article/article/
            article/article/div/div/div/span/span",
        "File1": "Ocena: 4",
        "File2": "Ocena: 5"
    },...
```