Data Extractor

Timotej Kovač

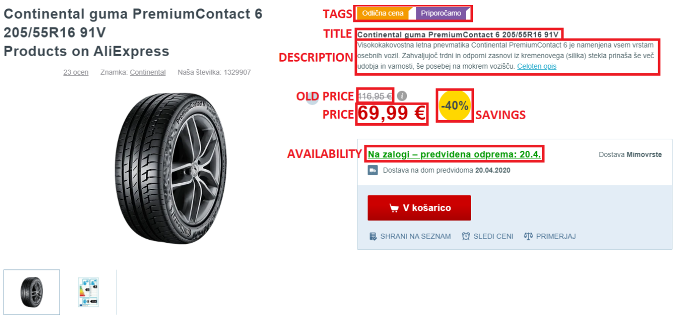
I. Introduction

In this paper we provide and discuss results gathered from our data extraction implementations using regular expressions, XPath and automatic web extraction. The later was heavily based on the RoadRunner approach for automatic web extraction [1]. We also provide most of the expressions used in our first two approaches (regular expressions and XPath).

II. Selection of optional web pages

For our first website we have chosen a detailed description page of a product available on web store Mimovrste [2]. Here we have chosen some interesting fields that we might be useful as shown in figure 1. Fields marked with an asterisk are optional. These were:

- **TAGS***, which further describe the item as having a discount, being a recommended product, etc.";
- **TITLE**;
- **DESCRIPTION**;
- **OLD PRICE***, which states the price before the now discounted price;
- **PRICE**, which states the current price;
- **SAVINGS***, which represents the percentage saved;
- **AVAILABILITY**, which states when the product will be available for shipment.



Continental guma PremiumContact 6 205/55R16 91V
Products on AliExpress

23 ocen Znamka: Continental Naša številka: 1329907

Odlična cena Priporočamo

Continental guma PremiumContact 6 205/55R16 91V

Visokokakovostna letna pnevmatika Continental PremiumContact 6 je namenjena vsem vrstam osebnih vozil. Zahvaljujoč trdni in odporni zasnovi iz kremenovega (silika) stekla prinaša še več udobja in varnosti, še posebej na mokrem vozišču. Celoten opis

~~116,95 €~~
69,99 €
-40%

Na zalogi – predvidena odprema: 20.4.
Dostava Mimovrste
Dostava na dom predvidoma 20.04.2020

V košarico

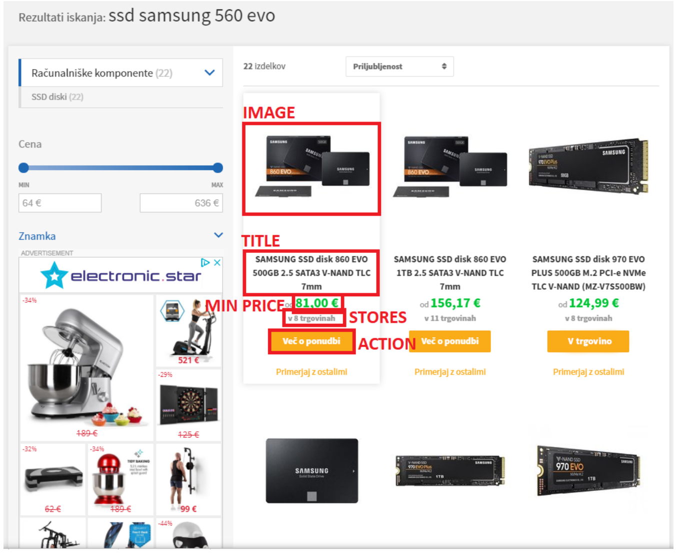
SHRANI NA SEZNAM SLEDI CENI PRIMERJAJ

Figure 1. Web store mimovrste.si with tagged fields that we used for extraction of web content in a detailed web page example.

For our second website we have chosen a page containing multiple items in a grid pattern on a web site Ceneje [3]. Here we have chosen the fields listed bellow:

- **IMAGE**;
- **TITLE**;
- **MIN PRICE**, which states the minimal price in all of the stores that provide the product;
- **NUMBER OF STORES**, which states the number of stores that provide the product;
- **ACTION**, which states what the button does either takes the user to a particular web store or to a list of web stores still on the same ceneje.si domain.

Here none of the fields are optional but some do vary as they may contain some other words in front of the fields or are ads which have a slightly different structure. This too can be seen in figure 2.



Rezultati iskanja: ssd samsung 560 evo

Računalniške komponente (22)
SSD diski (22)

Cena
MIN MAX
64 € 636 €

Znamka

ADVERTISEMENT

22 izdelkov Priljubljenost

SAMSUNG SSD disk 860 EVO 500GB 2.5 SATA3 V-NAND TLC 7mm
od 81,00 €
v 8 trgovinah
Več o ponudbi
Primerjaj z ostalimi

SAMSUNG SSD disk 860 EVO 1TB 2.5 SATA3 V-NAND TLC 7mm
od 156,17 €
v 11 trgovinah
Več o ponudbi
Primerjaj z ostalimi

SAMSUNG SSD disk 970 EVO PLUS 500GB M.2 PCI-e NVMe TLC V-NAND (MZ-V7S500BW)
od 124,99 €
v 8 trgovinah
V trgovino
Primerjaj z ostalimi

Figure 2. Web site ceneje.si with tagged fields that we used for extraction of web content in a list web page example.

III. Regular Expressions Implementation

Bellow is the list of all regular expressions that were used on the target web pages.

A. rtvslo.si

```
Title:  <h1>(.*?)</h1>
SubTitle: <div class=\"subtitle\">(.*?)</div>
Lead: <p class=\"lead\">(.*?)</p>
Content: <div class=\"article-body\">(.*?)</div>[ ]
*<div class=\"article-column\">
Author:  <div class=\"author-name\">(.*?)</div>
PublishedTime: <div class=\"publish-meta\">(.*?)<br>
```

B. overstock.com

```
Title(s): <td valign=\"top\">\W*<a.*?PROD_ID=([0-9]+)\"
.*?<b>(.*?)</b></a>
Content(s): <td valign=\"top\">\W*<a.*?PROD_ID=([0-9]+)
\".*?<span class=\"normal\">(.*?)<br>
ListPrice(s): <td valign=\"top\">\W*<a.*?PROD_ID=
([0-9]+)\".*?<s>(.*?)</s>
Price(s): <td valign=\"top\">\W*<a.*?PROD_ID=([0-9]+)
\".*?<span class=\"bigred\"><b>(.*?)</b>
Saving(s):
SavingPercent(s): <td valign=\"top\">\W*<a.*?PROD_ID=
([0-9]+)\".*?<span class=\"littleorange\">(.*?) \((
[0-9]{0,2}\%)\)</span>
```

C. mimovrste.si

```
Title: <h3.*?>(.*?)</h3>
Description: <p.*?itemprop=\"description\".*?>(.*?)<a
OldPrice: <del.*?class=\"rrp-price\".*?>(.*?)</del>
Price: <b class=\"pro-price.*?>(.*?)</b>
```

```
Availability: <a data-sel=\"availability-detail\".*?>
(.*?)</a>
Tags: <em class=\"label.*?>(.*?)</em>
Savings: <div class=\"label--round-sale.*?>(.*?)</div>
```

D. *ceneje.si*

```
Image(s): <div class=\"innerProductBox\">.*?<img.*?
alt=\"(.*?)\".*?src=\"(.*?)\"
Titles(s): <div class=\"innerProductBox\">.*?<img.*?
alt=\"(.*?)\".*?<h3>\W*<.*?>(.*?)</.*?>
MinPrice(s): <div class=\"innerProductBox\">.*?<img.*?
alt=\"(.*?)\".*?<b>(.*?)</b>
Store(s): <div class=\"innerProductBox\">.*?<img.*?
alt=\"(.*?)\".*?class=\"qtySellers\">\W*<b>(.*?)</b>
Action(s): <div class=\"innerProductBox\">.*?<img.*?
alt=\"(.*?)\".*?<div class=\"rBox\">\W*<.*?>(.*?)</.*?>
```

IV. XPath Implementation

Bellow are the lists of XPath expressions for every of the targeted pages.

A. *rtvslo.si*

```
Title: //*[@id=\"main-container\"]/div[3]/div/header/
h1/text()
SubTitle: //*[@id=\"main-container\"]/div[3]/div/
header/div[2]/text()
Lead: //*[@id=\"main-container\"]/div[3]/div/header/
p/text()
Content: string(//*[@id=\"main-container\"]/div[3]/div/
div[2])
Author: //*[@id=\"main-container\"]/div[3]/div/div[1]/
div[1]/div/text()
PublishedTime: //*[@id=\"main-container\"]/div[3]/div/
div[1]/div[2]/text()[1]
```

B. *overstock.si*

```
Title(s):  //table[@cellpadding=’2’]/tbody/
tr[" + str(i) + "]/td[2]/a/b/text()
Content(s): //table[@cellpadding=’2’]/tbody/
tr[" + str(i) + "]/td[2]/table/tbody/tr/td[2]/span/
text()
ListPrice(s): //table[@cellpadding=’2’]/tbody/
tr[" + str(i) + "]/td[2]/table/tbody/tr/td[1]/table/
tbody/tr[1]/td[2]/s/text()
Price(s): //table[@cellpadding=’2’]/tbody/
tr[" + str(i) + "]/td[2]/table/tbody/tr/td[1]/table/
tbody/tr[2]/td[2]/span/b/text()
Saving(s):
SavingPercent(s): //table[@cellpadding=’2’]/tbody/
tr[" + str(i) + "]/td[2]/table/tbody/tr/td[1]/table/
tbody/tr[3]/td[2]/span/text()
```

C. *mimovrste.si*

```
Title: //*[@id=\"content\"]/div/article/div[1]/
section[2]/h3/text()
Description: //*[@id=\"content\"]/div/article/div[1]/
section[2]/p[2]/text()
OldPrice: //*[@id=\"content\"]/div/article/div[1]/
section[2]/div[3]/div[1]/div[1]/div/del/text()
Price: //*[@class=\"price-wrapper\"]/div[1]/
div[1]/b/text()
Availability: //*[@class=\"delivery-wrapper\"]/a/
text()
Tags: //*[@id=\"content\"]/div/article/div[1]/
section[2]/p[1]/em[" + str(i) + "]/text()
Savings: //*[@id=\"content\"]/div/article/div[1]/
section[2]/div[3]/div[1]/div[2]/text()
```

D. *ceneje.si*

```
Image(s): //*[@id=\"productGrid\"]/div[" + str(i) + "]
/div/div[1]/a/img/@src
Title(s): //*[@id=\"productGrid\"]/div[" + str(i) + "]
/div/div[2]/h3/a/text()
MinPrice(s): //*[@id=\"productGrid\"]/div[" + str(i) + "]
/div/div[2]/p/a[1]/b/text()
Store(s): //*[@id=\"productGrid\"]/div[" + str(i) + "]
/div/div[2]/p/a[2]/b/text()
Action(s): //*[@id=\"productGrid\"]/div[" + str(i) + "]
/div/div[3]/a/text()
```

V. Automatic Web Extraction Implementation

In our approach we heavily relied on observations of the RoadRunner approach [1].

We first generated DOM structures from both input HTML pages. Then we cleaned them by removing any <head>, <style> and <script> tags along with any comments, links, navigation bar and footer. We achieved that by using a Cleaner class which was provided by the lxml library. After that the auto_ex function was called to produce a union–free regular expression. When checking for node mismatches we mainly relied on the type of tag the node represented, the id of the node, if it didn't contain more than 3 numbers which generally described an ad node, and sequential attributes of the tag to determine if the two nodes in separate trees are the same. If the node didn't match we then looked at the next node in the second tree to see if the match was there. That way we determined with some accuracy which of the tags was optional. We then proceeded recursively though all of the children of the node processing them and then adding any text that was followed by the current node. The pseudo code of the function is described bellow.

A. Pseudo code

```
def auto_ex(tree1, tree2):
    for child in tree1:

        # Check if there is any node on count position in tree2
        target = get_target(node_2, count)
        if target is None:
            wrapper += "(<" + str(child.tag).upper() + "... >)?"
            continue

        target = node_2.getchildren()[count]

        # Handle tag mismatches
        if mismatch(child, target):
            if contains(node_2, count, child):
                wrapper += "(<" + target.tag.upper() + "... >)?"
                count += 1
                target = get_target(node_2, count)
            else:
                wrapper += "(<" + str(child.tag).upper() + "... >)?"
                continue
        # Add starting tag
        new_tag = "<" + str(child.tag).upper() + ">"

        # Add text inside of the tag
        if child.text is not None:
            new_tag += get_smt(child.text, target.text)

        # Handle children of this node
        new_tag += auto_ex(child, target)

        # Add a closing tag
        new_tag += "</" + str(child.tag).upper() + ">"

        # Handle text after the current node
        if child.tail is not None:
            tail = child.tail
            if len(tail) > 0:
                new_tag += get_smt(tail, target.tail)

        # Replace multiple occurrences with a special tag
        if new_tag == prev_tag:
            wrapper = re.sub(new_tag + "$", "(" + new_tag + ")+", wrapper)
            continue

        wrapper += new_tag

        prev_tag = new_tag
        count += 1

        # If this is the last node in tree1 but there are still nodes in tree2 this one must be optional
        if node_exists(tree2, level, count):
            wrapper += "(<" + node_that_exists_in_tree2 + "... >)?"

    return wrapper
```

B. Results

C. rtvslo.si

```
<DIV><DIV>(<DIV></DIV>)+</DIV></DIV><DIV><DIV>...#TOP BAR</DIV></DIV><DIV><DIV><DIV><DIV><DIV>
(<DIV></DIV>)+</DIV><DIV><A>RTVSLO.si</A></DIV><UL>
  <LI><A>Slovenija</A></LI><LI><A>Svet</A></LI><LI><A>Šport</A></LI><LI><A>Kultura</A></LI>
  <LI><A>Življenjski slog</A></LI><LI><A>Svet zabave</A></LI></UL><DIV>
<DIV>Iskanje</DIV><DIV></DIV></DIV><DIV><DIV><SPAN></SPAN></DIV>(<DIV>
<A></A></DIV>)+(<DIV... >)?<DIV><DIV>(<SPAN></SPAN>)+</DIV><SPAN>Kazalo</SPAN></DIV></DIV></DIV></DIV>
</DIV><DIV><DIV><DIV><DIV><DIV></DIV><DIV><DIV><DIV><H3>Predlogi</H3><DIV><DIV>Odbojka</DIV><DIV>Hokej
</DIV><DIV>Tenis</DIV></DIV></DIV><DIV><H3>Rezultati iskanja</H3><DIV></DIV></DIV></DIV><P>Zaradi
  testiranja je dodanih umetnih 2 sekunde delaya.
  </P></DIV></DIV></DIV></DIV></DIV><DIV><DIV><DIV><DIV><A><IMG></IMG></A></DIV><DIV>
<H5><SPAN></SPAN><SPAN><A></A></SPAN><A></A></H5><P></P></DIV></DIV></DIV></DIV></DIV>
          (<DIV... >)?<DIV><DIV>(<SPAN></SPAN>)+</DIV></DIV><DIV></DIV><DIV><DIV><DIV><DIV>
<UL>... #NAV ITEMS</DIV></DIV></DIV></DIV><DIV>
</DIV><DIV><DIV><H3><SVG><G>(<CIRCLE></CIRCLE>)+</G>(<CIRCLE... >)?(<CIRCLE... >)?</SVG><A>#text</A></H3>
</DIV><DIV><A>Uredi</A></DIV><DIV><DIV><HEADER><DIV><H3><SVG><G>
(<CIRCLE></CIRCLE>)+</G>(<CIRCLE... >)?(<CIRCLE... >)?</SVG><A>#text</A></H3><A></A></DIV><H1>
          #text</H1><DIV>#text</DIV><DIV><DIV><STRONG>Miha Merljak</STRONG>#text</DIV><DIV>
Ljubljana  -  MMC RTV SLO </DIV></DIV>(<DIV... >)?(<P... >)?</HEADER><DIV>
<DIV><DIV>Miha Merljak</DIV></DIV><DIV>#text<BR></BR>Ljubljana  -  MMC RTV SLO</DIV><DIV>
(<DIV></DIV>)+(<A... >)?</DIV><FIGURE><H5>Poudarki</H5><UL>
(<LI>#text</LI>)+</UL></FIGURE></DIV><DIV>#text<DIV>#text(<FIGURE... >)?</DIV><ARTICLE>#text<FIGURE>
    #text(<DIV... >)?(<DIV... >)?</FIGURE>(<P... >)?(<P... >)?(<P... >)?(<P... >)?(<P... >)?(<P... >)?
    (<P... >)?(<P... >)?(<P... >)?(<P... >)?(<P... >)?(<P... >)?(<P... >)?(<DIV... >)?</ARTICLE></DIV>
    <DIV><DIV><DIV><DIV></DIV></DIV><DIV><H4>Zadnje iz sekcije</H4>(<DIV><DIV><DIV><A><IMG></IMG></A>
    </DIV><DIV><A>#text</A></DIV></DIV></DIV>)+</DIV></DIV></DIV><DIV><DIV>(<A>#text</A>)+</DIV><DIV>
<A>Prijavi napako</A>(<DIV></DIV>)+</DIV><FIGURE><DIV>Oglas</DIV><DIV>
</DIV></FIGURE></DIV></DIV></DIV><DIV>#text(<DIV... >)?(<A... >)?</DIV><DIV>
(<DIV... >)?</DIV>(<DIV... >)?<DIV>#text<DIV>#text(<H3... >)?</DIV>(<DIV... >)?(<DIV... >)?</DIV>
</DIV><DIV><DIV><DIV><DIV><H5>Prijava</H5></DIV><DIV>
            <DIV><LABEL>Uporabniško ime:</LABEL></DIV><DIV><LABEL>Geslo:</LABEL></DIV><DIV>
              <A>Registracija</A></DIV><A>Pozabljeno geslo?</A><DIV><DIV><SPAN>Temni način <SPAN>BETA
              </SPAN></SPAN><LABEL><SPAN></SPAN></LABEL></DIV></DIV></DIV></DIV></DIV></DIV><DIV>
    <DIV><DIV><DIV><H5>Prijavljen</H5></DIV><DIV><H2></H2><A>Odjava</A><A>Uporabniški račun</A><DIV><DIV>
    <SPAN>Temni način <SPAN>BETA</SPAN></SPAN><LABEL><SPAN></SPAN></LABEL></DIV></DIV></DIV></DIV></DIV>
  </DIV><DIV><DIV><DIV><DIV><H5>Časovno obdobje po meri</H5></DIV><DIV><DIV>
            <LABEL>Od</LABEL><DIV></DIV></DIV><DIV><LABEL>Do</LABEL><DIV></DIV></DIV></DIV><DIV>
      </DIV></DIV></DIV></DIV><DIV><DIV></DIV><DIV><DIV>(<DIV></DIV>)+</DIV><DIV>
      <DIV><DIV></DIV><DIV><DIV><DIV><DIV></DIV></DIV></DIV></DIV></DIV>(<DIV>
        <DIV></DIV></DIV>)+</DIV></DIV></DIV>(<IMG></IMG>)+
```

D. overstock.com

```
<TABLE><TBODY><TR><TD></TD><TD> <TABLE><TBODY><TR><TD><TABLE><TBODY><TR>
<TD><A><IMG></IMG></A></TD><TD><IMG></IMG></TD><TD><A><IMG></IMG></A></TD></TR></TBODY></TABLE><MAP>
(<AREA></AREA>)+</MAP></TD></TR><TR><TD>
(<A><IMG></IMG></A>)+</TD></TR></TBODY></TABLE></TD></TR><TR><TD><IMG></IMG></TD></TR><TR><TD>
<TABLE><TBODY><TR><TD><TABLE><TBODY><TR> <TD>
<SPAN>Search:</SPAN></TD><TD> </TD><TD></TD><TD>  </TD></TR></TBODY></TABLE></TD><TD>
<IMG></IMG><BR></BR><A><IMG></IMG></A></TD><TD><IMG></IMG></TD></TR></TBODY></TABLE></TD></TR><TR><TD>
<IMG></IMG></TD></TR></TBODY></TABLE><TABLE><TBODY><TR>
<TD><IMG></IMG></TD><TD>
<BR></BR><TABLE><TBODY><TR><TD>
<TABLE><TBODY><TR><TD> </TD><TD><SPAN><B>#text</B></SPAN></TD></TR>(<TR><TD> </TD><TD><A>#text</A></TD>
</TR>)+<TR><TD> </TD><TD><A>(<B... >)?</A></TD></TR></TBODY></TABLE></TD></TR></TBODY></TABLE><BR></BR>
<SPAN><B>Stores</B></SPAN><BR></BR><TABLE>
<TBODY><TR><TD><A>Apparel, Shoes & Access.</A></TD></TR><TR><TD><A>Books, Movies, CDs, Games</A></TD></TR>
<TR><TD><A>Electronics & Computers</A></TD></TR><TR><TD><A>Home & Garden</A></TD></TR><TR><TD><A>
<B>Jewelry, Watches & Gifts</B></A></TD></TR><TR><TD><A>Sports, Travel & Toys</A></TD></TR><TR><TD>
<A>Worldstock</A></TD></TR></TBODY></TABLE><BR></BR><SPAN><B>New Stock</B></SPAN><BR></BR><TABLE>
<TBODY><TR><TD>
<A>Ralph Lauren $29.95</A></TD></TR><TR><TD>
```

```
<A>Ben Sherman 53\% off</A></TD></TR><TR><TD>
<A>Pre-order Harry Potter DVD</A></TD></TR><TR><TD>
<A>HP 2GHz System $499</A></TD></TR><TR><TD>
<A>New Items within 7 Days</A></TD></TR></TBODY></TABLE><BR></BR><SPAN><B>Customer Service</B></SPAN><BR>
</BR><TABLE><TBODY><TR><TD><A>Shopping Cart & Checkout</A></TD></TR><TR><TD><A>Track Your Order</A></TD>
</TR><TR><TD><A>Your Account</A></TD></TR><TR><TD><A>Help & FAQ</A></TD></TR><TR><TD><A>
Best Price Guarantee</A></TD></TR></TBODY></TABLE><BR></BR><SPAN><B>About Us</B></SPAN><BR></BR><TABLE>
<TBODY><TR><TD><A>About Us</A></TD></TR><TR><TD><A>Privacy & Security</A></TD></TR><TR><TD><A>
Terms & Conditions</A></TD></TR><TR><TD><A>Become An Affiliate</A></TD></TR><TR><TD><A>Business Purchases</A>
</TD></TR><TR><TD><A>Have Products to Sell?</A></TD></TR><TR><TD><A>Investor Relations</A></TD></TR></TBODY>
</TABLE><IMG></IMG><BR></BR></TD>(<TD><IMG></IMG></TD>)+<TD>#text(<BR... >)?(<B... >)?(<BR... >)?(<TABLE... >)?
</TD></TR><TR>
<TD><IMG></IMG></TD><TD>© 2003<BR></BR>Overstock.com</TD><TD><IMG></IMG></TD><TD>
<BR></BR><SPAN>* $2.95 flat rate shipping to the lower 48 states only.  Some items excluded due to size and/or
weight.</SPAN>
<IMG></IMG><BR></BR><IMG></IMG></TD></TR><TR><TD><IMG></IMG></TD></TR></TBODY></TABLE><MAP>
(<AREA></AREA>)+</MAP>
```

E. mimovrste.si

```
<DIV><DIV><DIV><DIV><A>Odgovorni</A></DIV><DIV><A>do kupcev in zaposlenih</A></DIV></DIV><DIV>
  <DIV><A>Brezkontaktna dostava</A></DIV><DIV><A>na dom</A></DIV></DIV><DIV>
  <DIV><A>Vse informacije</A></DIV><DIV><A>na enem mestu</A></DIV></DIV><DIV>
  <DIV><A>Spletna tržnica</A></DIV><DIV><A>Mimovrste Partner</A></DIV></DIV><DIV>
  <DIV><A>SKUPAJ V VSEM</A></DIV><DIV><A>več >></A></DIV></DIV><DIV>
  <DIV><A>Skupaj v vsem</A></DIV><DIV><A>več >></A></DIV></DIV></DIV></DIV><DIV></DIV><DIV>
<SPAN></SPAN><HEADER>
  <VIP-PROGRAM-NOTICE-PANEL></VIP-PROGRAM-NOTICE-PANEL><DIV><DIV>
  <DIV><A><DIV>(<SPAN></SPAN>)+</DIV><DIV>header_mobile_menu</DIV></A></DIV><DIV><A>
<IMG></IMG></A></DIV><DIV><UL><LI><A><SVG><PATH></PATH></SVG>Domača stran</A></LI><LI><A><SVG><PATH></PATH>
</SVG>Prevzemna mesta</A></LI><LI><A><SVG><PATH></PATH></SVG>Nakupovalni nasveti</A></LI><LI><A><SVG><PATH>
</PATH></SVG>Pooblaščeni serviserji</A></LI><LI><A><SVG><PATH></PATH></SVG>Kontakt</A></LI></UL><A><IMG>
</IMG><NOSCRIPT><IMG></IMG></NOSCRIPT></A></DIV></DIV><DIV>
 (<DIV... >)?<DIV>(<A... >)?(<DIV... >)?</DIV>(<DIV... >)?<DIV></DIV>(<DIV... >)?(<DIV... >)?</DIV><HR>
 </HR></DIV></HEADER><P>
<A>Preskoči na vsebino</A></P><DIV><DIV></DIV><DIV><DIV><DIV></DIV></DIV><MAIN><DIV> <ARTICLE><DIV>
    <SECTION><H1>#text<ELEMENT><DIV>Products on AliExpress</DIV></ELEMENT></H1><DIV><SPAN><SPAN>#text
    </SPAN><SPAN></SPAN>(<SPAN... >)?(<SPAN... >)?(<SPAN... >)?(<SPAN... >)?</SPAN>(<A... >)?
    (<SPAN... >)?</DIV><DIV><DIV><DIV><DIV><DIV><DIV><DIV><DIV><DIV><DIV>(<IMG></IMG>)+</DIV></DIV>
   </DIV>(<DIV... >)?</DIV></DIV></DIV></DIV>(<SPAN... >)?(<SPAN... >)?(<SPAN... >)?</DIV><DIV>(<DIV... >)?
    <SPAN>#text</SPAN>(<SPAN... >)?</DIV>(<DIV... >)?</DIV><DIV></DIV></DIV><NOSCRIPT>(<DIV... >)?
    </NOSCRIPT></SECTION><SECTION><P><EM>Odlična cena</EM><EM>Priporočamo</EM></P><H3>#text</H3><P>
     #text(<A... >)?</P><DIV><SPAN></SPAN><SPAN>#text</SPAN></DIV>(<DIV... >)?<P></P><DIV><DIV><DIV>
      (<DIV... >)?<B>#text</B>(<DIV... >)?</DIV>(<DIV... >)?</DIV><DIV></DIV></DIV><DIV><DIV><A>#text
      </A><DIV><SPAN>Dostava
<B>Mimovrste</B></SPAN><SPAN></SPAN></DIV></DIV><DIV><DIV><DIV><SVG><PATH></PATH></SVG></DIV>
<SPAN>Dostava na dom predvidoma
<B>#text</B></SPAN></DIV></DIV><DIV><SPAN><DIV><SPAN><DIV></DIV></SPAN><DIV><DIV></DIV></DIV></DIV></SPAN>
<DIV><A><SPAN><SVG><G><PATH></PATH></G></SVG></SPAN><SPAN>Shrani na seznam</SPAN></A><A><SVG><PATH></PATH>
</SVG><SPAN>Sledi ceni</SPAN></A><SPAN><A><SPAN><SVG><G><PATH></PATH></G></SVG></SPAN><SPAN>Primerjaj</SPAN>
</A></SPAN></DIV><DIV></DIV></DIV></DIV><P><SPAN>
Naša številka: <SPAN>#text</SPAN></SPAN></P><DIV><P><SPAN><SVG><G><PATH></PATH></G></SVG></SPAN></P><P><SPAN>
 <SVG><G><PATH></PATH></G></SVG></SPAN>Sledilec cene se namešča...</P><DIV><DIV><A>Prekliči sledenje</A><P>
  Aktivirali ste sledenje dobavljivosti artikla. Po e-pošti boste obveščeni o spremembah dobave.</P></DIV>
  <SPAN></SPAN><DIV><A>Prekliči sledenje</A><P>Aktivirali ste sledenje cene artikla na <SPAN>0</SPAN>.
   Po e-pošti boste obveščeni o morebitnih spremembah.</P></DIV></DIV></DIV></SECTION></DIV><DIV></DIV>
   (<SECTION... >)?<DIV><DIV><DIV><H2>Dodatki za artikel</H2></DIV></DIV></DIV><SPAN><DIV> <SPAN><DIV><DIV>
    <DIV><H2>Predstavitev</H2><SECTION><H3>Opis artikla</H3>(<H2... >)?(<P... >)?(<HR... >)?<DIV><DIV><DIV>
     <H2><SPAN>#text</SPAN></H2>(<P... >)?(<P... >)?(<P... >)?</DIV><DIV>(<DIV... >)?</DIV></DIV></DIV>
     <HR></HR>(<DIV... >)?(<HR... >)?<H3>#text</H3><TABLE><TBODY><TR><TD>#text</TD><TD><B>#text</B></TD>
     </TR>(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?</TBODY></TABLE><HR></HR></SECTION></DIV>
    </DIV><DIV><DIV><H2>Tehnične podrobnosti</H2><SECTION><H3>Parametri artikla \%1\% \%2\%</H3><TABLE>
<TBODY><TR><TH>#text</TH><TD>#text(<ABBR... >)?</TD></TR><TR><TH>#text</TH><TD>#text</TD></TR>
```

```
(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?(<TR... >)?</TBODY></TABLE><HR>
</HR></SECTION></DIV></DIV>(<DIV><SPAN></SPAN></DIV>)+(<DIV... >)?</DIV><DIV></DIV></SPAN></DIV>
</SPAN></ARTICLE></DIV></MAIN></DIV><DIV><A>X</A></DIV><A><DIV></DIV></A><DIV></DIV></DIV></DIV>
<DIV><DIV><DIV></DIV></DIV></DIV><DIV></DIV><DIV><DIV><DIV><DIV><DIV><DIV><DIV></DIV></DIV><DIV>
<DIV><A>(<DIV><DIV></DIV></DIV>)+</A></DIV></DIV><DIV><A><SVG><PATH></PATH></SVG></A><DIV><DIV>
<DIV><SVG><PATH></PATH></SVG></DIV></DIV></DIV></DIV></DIV></DIV></DIV></DIV></DIV>(<DIV... >)?
```

F. ceneje.si

```
<DIV><P>Ceneje.si uporablja piškotke za zagotavljanje kvalitetne uporabniške izkušnje, nemoteno delovanje
  vseh funkcij spletne strani, prikazovanje oglasov in merjenje ter analizo podatkov o uporabi spletne
  strani.</P><P>Za nemoteno uporabo naše strani se moraš strinjati z uporabo piškotkov.</P><A>Strinjam se
  </A><A>Nastavitve</A></DIV><DIV>#text<HEADER>
</HEADER><DIV><DIV><DIV><DIV>
      <A><IMG></IMG></A></DIV><DIV><DIV><DIV><DIV>
        <IMG></IMG><H6>Kategorije</H6></DIV></DIV><DIV><DIV><DIV></DIV><IMG></IMG></DIV></DIV></DIV><DIV>
    <DIV>...# NAV_ITEMS </DIV><DIV><DIV>
            ...# NAV ITEMS </DIV>(<DIV></DIV>)+</DIV></DIV><DIV></DIV></DIV></DIV></DIV><DIV><DIV><DIV>
    </DIV></DIV><DIV>
  </DIV><DIV>#text<DIV>#text(<DIV... >)?<DIV>
    <A>Domov</A></DIV><DIV>#text<H1>(<SPAN... >)?</H1>(<DIV... >)?</DIV><DIV>#text(<DIV... >)?<DIV>
      (<DIV... >)?(<DIV... >)?(<DIV... >)?(<DIV... >)?(<DIV... >)?(<DIV... >)?</DIV>(<DIV... >)?</DIV>
      <DIV>#text(<DIV... >)?(<DIV... >)?</DIV></DIV></DIV></DIV><DIV><DIV>
       <DIV><DIV></DIV></DIV></DIV><DIV>
  </DIV><DIV><DIV><SPAN>Advertisement</SPAN></DIV><DIV></DIV></DIV></DIV><DIV>#text<DIV><DIV></DIV>
</DIV></DIV><DIV><DIV><P>
          <SPAN>Želiš med prvimi izvedeti za najboljše sezonske ponudbe?</SPAN><BR></BR><SPAN>Prijavi se
             na e-novice, ki jih tedensko pripravljamo zate.</SPAN></P><P>#text<SPAN>
            <SPAN>Vnesi pravilen e-naslov
</SPAN><SPAN></SPAN></SPAN></P><P>Hvala za prijavo. Kmalu lahko pričakuješ prve tedenske e-novice.</P>
<DIV><P>Dajem privolitev, da želim prejemati:</P><P>* Izberi vsaj eno izmed možnosti.</P><DIV>
            <TABLE>
              <TBODY><TR>
                <TD><LABEL></LABEL></TD><TD>
                  <P><B>Ceneje.si e-novičke</B>s predstavitvijo kategorij izdelkov, zanimivimi ponudbami
                    in nakupovalnimi nasveti ter pomembna obvestila uporabnikom</P></TD></TR><TR>
                <TD><LABEL></LABEL></TD><TD>
                  <P><B>Personalizirane in relevantne ponudbe</B>izdelkov iz kategorij, ki te zanimajo<BR>
                  </BR><I>(Da ti lahko ponudimo prilagojeno vsebino, bomo na ceneje.si spremljali tvoje
                    aktivnosti, analizirali interese in uporabili profiliranje ter avtomatizirano odločanje)
                  </I></P></TD></TR><TR>
                <TD><LABEL></LABEL></TD><TD>
                  <P><B>Spletne kataloge in letake trgovcev</B>- prelistaš jih lahko takoj, ko izidejo</P>
                </TD></TR></TBODY></TABLE></DIV><P><A>Nameni obdelave</A>, način hrambe in varovanja
                  osebnih podatkov ter tvoje pravice so opisani v<A>Pravilih o zasebnosti</A>.</P></DIV>
                </DIV></DIV><DIV><DIV>
    <SPAN><SPAN>® 2020 Ceneje d.o.o., del skupine Rockaway</SPAN></SPAN><UL>
(<LI><A><IMG></IMG></A></LI>)+</UL></DIV></DIV></DIV><DIV>
<DIV><DIV><UL><LI><DIV>
  <SPAN>0</SPAN>Seznam</DIV><DIV>
<DIV></DIV><DIV><DIV></DIV><UL>
    <LI><A>Primerjaj</A></LI><LI><A>Pošlji</A></LI></UL></DIV><DIV><DIV>
  </DIV><DIV>Označi izdelke in nato zgoraj izberi funkcijo</DIV></DIV><DIV>
  <DIV>#text<DIV>
          <P><B>Seznam je prazen.</B></P><P>
           Za dodajanje izdelkov v seznam uporabi "V seznam" ali ikona
           <IMG></IMG>, ki se nahaja na vsaki strani z izdelkom kot tudi na seznamih izdelkov v posamezni
            kategoriji.</P></DIV></DIV></DIV></DIV></LI><LI>
<DIV>Zadnji ogledani</DIV><DIV>
  <DIV></DIV><DIV><UL>
      <LI><A>Izdelki</A></LI></UL></DIV><DIV>
    <DIV></DIV></DIV></DIV></LI></UL></DIV></DIV></DIV>(<DIV>
  </DIV>)+(<DIV... >)?(<NOSCRIPT... >)?(<DIV... >)?(<DIV></DIV>)+
```

VI. Conclusion

We were successful in implementing all of the above listed approaches and have gathered good results. Our automatic web extraction approach could of course be improved with additional information gathered from other articles that tackle with the same problems or with testing the algorithm on more websites.

References

[1] V. Crescenzi, G. Mecca, P. Merialdo *et al.*, "Roadrunner: Towards automatic data extraction from large web sites," in *VLDB*, vol. 1, 2001, pp. 109–118.

[2] "mimovrste=) | računalništvo, prenosniki, GSM telefoni, avdio-video," Accessable: https://www.mimovrste.com/, 2020, [Accessed: 24. 04. 2020].

[3] "Ceneje.si - Prva misel pred nakupom," Accessable: https://www.ceneje.si/, 2020, [Accessed: 24. 04. 2020].