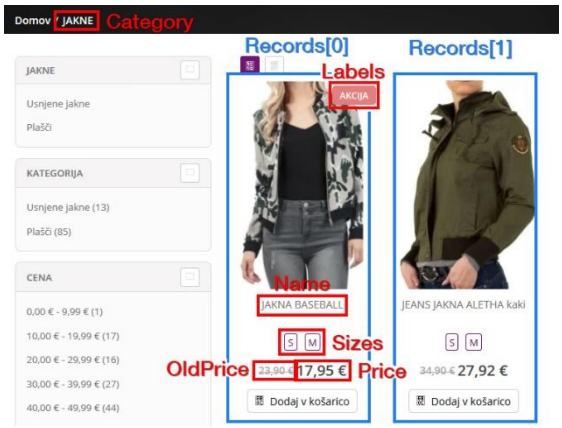
Web Information Extraction and Retrieval Programming Assignment 2

Rok Cej, Metodija Bucevski

Selected web pages

We selected two pages from the site manazara.si. Each page has a category (jakne, puloverji, ...) and a list of records. Each record has a name, a list of available sizes (S, M, L, ...), a price, an optional old price and a list of optional labels (akcija, nov, ...).



RoadRunner implementation

Our RoadRunner implementation was based on the method proposed in the original paper (source). The main difference is that our method builds a tree instead of parsing a list of tokens. There are 4 types of tokens, STRING, TAG, ITERATOR and OPTIONAL. Data items are represented as STRING tokens with their content set to #TEXT.

Pseudocode

```
// Preprocessing
-Parse and clean HTML
-Remove unnecessary HTML tags
-Generate tree structure

// Tree matching
-Input: wrapper_node, sample_node
-Check if nodes (roots) match
-Iterate over children:
   -If children are strings, match
   contents
   -If children are tags, recursively
   match their children
   -If children are patterns
   (iterator/optional), match their
   siblings
-If found match → generalize wrapper,
continue
-Else: // Found a mismatch
   -Attempt iterator discovery:
      -Generate iterator candidates
      -Test iterator candidates
      -If found iterator pattern →
      generalize wrapper, continue
   -Attempt optional discovery:
      -If found optional pattern →
      generalize wrapper, continue
-Return wrapper

// Displaying results
-Convert wrapper tree to UFRE
-Only show ancestors of data items
(#TEXT nodes)
```

Heuristics

HTML tag removal

We remove the following tags when parsing HTML:
, <hr>, <script>, <input>, and <option>. They do not contain any data records and are therefore not useful for data extraction.

Pattern assumptions

We assume that data records will never be nested relative to each other. In other words, they will appear as siblings in the tree. This allows us to represent a pattern (optional, iterator) as a single node.

Candidate limit

As suggested by the original authors, we limited the number of iterator candidates to $k = 4$. This is due to the fact that candidates tend to be close to the mismatch point. As a result, unnecessary computation is avoided on large websites.

Regex implementation

Overstock.com

Data item	Regex
Title	<a href=(\")http:(\/\/)www.overstock.com(\/)cgi-bin(\/)d2.cgi(\?)PAGE=PROFRAME(\&)amp;PROD_ID=(\d{5,7})(\")>(.*)
ListPrice	<tbody><tr><td align=\"right\" nowrap=\"nowrap\">(.*)<\/b><\/td><td align=\"left\" nowrap=\"nowrap\"><s>(\$(\d{1,5})(\,)?(\d{1,5})?\.(\d{1,5}))<\/s><\/td><\/tr>
Price	<tr><td align=\"right\" nowrap=\"nowrap\">Price:<\/b><\/td><td align=\"left\" nowrap=\"nowrap\"><span class=\"bigred\">(\$(\d{1,5})(\,)?(\d{1,5})?\.(\d{1,5}))<\/b><\/span><\/td><\/tr>
Saving	<tr><td align=\"right\" nowrap=\"nowrap\">You Save:<\/b><\/td><td align=\"left\" nowrap=\"nowrap\"><span class=\"littleorange\">(\$(\d{1,5})(\,)?(\d{1,5})?\.(\d{1,5})) (\()[0-9][0-9](\%)(\))<\/span><\/td><\/tr>
SavingPercent	<tr><td align=\"right\" nowrap=\"nowrap\">You Save:<\/b><\/td><td align=\"left\" nowrap=\"nowrap\"><span class=\"littleorange\">(\$(\d{1,5})(\,)?(\d{1,5})?\.(\d{1,5})) ((\()([0-9][0-9](\%))(\)))<\/span><\/td><\/tr>
Content	(?:<td valign=\"top\"><span class=\"normal\">)((.*?)[\s\S]*?) (.*?)\"><span class=\"tiny\">(.*)<\/b>

Rtvslo.si

Data item	Regex
Author	<div class=\"author-timestamp\">\s*\n?[]+(.*)<\/strong>|(.*)
PublishedTime	<div class=\"author-timestamp\">\s*\n?[]+(.*)<\/strong>|(.*)
Title	<h1>(.*)<\/h1>
SubTitle	<div class=\"subtitle\">(.*)<\/div>
Lead	<p class=\"lead\">(.*)<\/p>
Content	(<div style=\"position:absolute\; left: -1000px\; top: -1000px\;\"><img (.*?)<p>)*?(<p class=\"Body\">(.*?)<iframe (.*?)<\/p>)?<p>(.*?)(<\/p>)

Manazara.si

Data item	Regex
Category	
Name	
Sizes	
Price	
OldPrice	
Labels	

Xpath implementation

Overstock.com

Data item	XPath
Title	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/a/b/text()
ListPrice	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s/text()
Price	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b/text()
Saving	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text()
SavingPercent	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text()
Content	/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td/table/tbody/tr/td/table/tbody/tr/td[2]/table/tbody/tr/td[2]/span/text()

Rtvslo.si

Data item	XPath
Author	//div[@class='article-meta']/div[@class='author']/div[@class='author-name']/text()
PublishedTime	//div[@class='article-meta']/div[@class='publish-meta']/text()
Title	//header[@class='article-header']/h1/text()
SubTitle	//header[@class='article-header']/div[@class='subtitle']/text()
Lead	//header[@class='article-header']/p[@class='lead']/text()
Content	//div[@class='article-body']/article[@class='article']/p/text() | //div[@class='article-body']/article[@class='article']/p/strong/text()

Manazara.si

Data item	XPath
Category	/html/body/div/div/div[4]/div[2]/div/div/div/ul/li/strong/text()
Name	//div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='details-area']/h2[@class='product-name']/a/text()
Sizes	//div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='details-area']/div[@class='items-size-wrapper']/span[@class='size']/text()
Price	//div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='details-area']/div[@class='price-box']/span[@class='regular-price']/span[@class='price']/text() | //div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='details-area']/div[@class='price-box']/p[@class='special-price']/span[@class='price']/text()
OldPrice	//div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='details-area']/div[@class='price-box']/p[@class='old-price']/span[@class='price']/text()
Labels	//div[@class='category-products']/ul[contains(@class, 'columns4')]/li[contains(@class, 'item')]/div[@class='item-area']/div[@class='product-image-area']/div[@class='product-label']/span/text()

Overstock.com wrapper

```
<body>
 <table>
  <tbody>
   <tr>
    <td>
     <table>
      <tbody>
       <tr>
        <td>
         <table>
          <tbody>
           (
            <tr>
             <td>
              (
               <a>
                (
                 #TEXT
                )?
               </a>
              )?
              (
               <span>
                <b>
                 #TEXT
                </b>
               </span>
              )?
             </td>
            </tr>
           )+
          </tbody>
         </table>
        </td>
       </tr>
      </tbody>
     </table>
    </td>
    <td>
     <b>
      <a>
       #TEXT
      </a>
     </b>
     <table>
      <tbody>
       <tr>
        <td>
         <table>
          <tbody>
           <tr>
            <td>
             <table>
              <tbody>
               (
                <tr>
                 <td>
                  <span>
                   <b>
                    #TEXT
                   </b>
                  </span>
                 </td>
                 <td>
                  <b>
                   #TEXT
                  </b>
                  #TEXT
                  <a>
                   #TEXT
                  </a>
                  (
                   <a>
                    #TEXT
                   </a>
                  )+
                 </td>
                </tr>
               )+
              </tbody>
             </table>
            </td>
           </tr>
          </tbody>
         </table>
        </td>
       </tr>
      </tbody>
     </table>
    </td>
   </tr>
   <tr>
    <td>
     <table>
      <tbody>
       <tr>
        <td>
         <table>
          <tbody>
           (
            <tr>
             (
              <td>
               (
                <a>
                 <b>
                  #TEXT
                 </b>
                </a>
                <table>
                 <tbody>
                  <tr>
                   <td>
                    <table>
                     <tbody>
                      <tr>
                       <td>
                        <s>
                         #TEXT
                        </s>
                       </td>
                      </tr>
                      <tr>
                       <td>
                        <span>
                         <b>
                          #TEXT
                         </b>
                        </span>
                       </td>
                      </tr>
                      <tr>
                       <td>
                        <span>
                         #TEXT
                        </span>
                       </td>
                      </tr>
                     </tbody>
                    </table>
                   </td>
                  </tr>
                 </tbody>
                </table>
               </td>
              </tr>
             )+
            </tbody>
           </table>
          </td>
         </tr>
        </tbody>
       </table>
      </td>
     </tr>
     <tr>
      <td>
       <table>
        <tbody>
         <tr>
          <td>
           <table>
            <tbody>
             (
              <tr>
               <td>
                <span>
                 <b>
                  #TEXT
                 </b>
                </span>
               </td>
               <td>
                <b>
                 #TEXT
                </b>
                #TEXT
                <a>
                 #TEXT
                </a>
                (
                 <a>
                  #TEXT
                 </a>
                )+
               </td>
              </tr>
             )+
            </tbody>
           </table>
          </td>
         </tr>
        </tbody>
       </table>
      </td>
     </tr>
     <tr>
      <td>
       <table>
        <tbody>
         <tr>
          <td>
           <table>
            <tbody>
             (
              <tr>
               <td>
                <span>
                 <b>
                  #TEXT
                 </b>
                </span>
               </td>
               <td>
                <b>
                 #TEXT
                </b>
                #TEXT
                <a>
                 #TEXT
                </a>
                (
                 <a>
                  #TEXT
                 </a>
                )+
               </td>
              </tr>
             )+
            </tbody>
           </table>
          </td>
         </tr>
        </tbody>
       </table>
      </td>
     </tr>
    </tbody>
   </table>
  </td>
 </tr>
</tbody>
</table>
</body>
```

Rtvslo.si wrapper

```
<body>
 (
  <div>
   (
    (
     <div>
      (
       <div>
        (
         <header>
          <div>
           <h3>
            <a>
             #TEXT
            </a>
           </h3>
          </div>
          <h1>
           #TEXT
          </h1>
          (
           <div>
            (
             <div>
              (
               #TEXT
              )?
             </div>
            )?
            (
             #TEXT
            )?
           </div>
          )+
          <p>
           #TEXT
          </p>
         </header>
        )?
        (
         (
          <div>
           (
            <div>
             #TEXT
            </div>
           )?
           (
            (
             (
              <div>
               (
                (
                 #TEXT
                )?
               )?
              </div>
             )?
             (
              <figure>
               (
                <ul>
                 <li>
                  #TEXT
                 </li>
                 <li>
                  #TEXT
                 </li>
                 <li>
                  #TEXT
                 </li>
                </ul>
               )?
              </figure>
             )?
            )+
            (
             <article>
              (
               <p>
                (
                 #TEXT
                )?
                (
                 <strong>
                  #TEXT
                 </strong>
                )?
                (
                 #TEXT
                 #TEXT
                 #TEXT
                 #TEXT
                 #TEXT
                )?
               </p>
              )+
             </article>
            )?
            (
             <figure>
              <ul>
               <li>
                #TEXT
               </li>
               <li>
                #TEXT
               </li>
               <li>
                #TEXT
               </li>
              </ul>
             </figure>
            )?
           )?
           (
            <div>
             (
              <div>
               <div>
                <div>
                 <a>
                  #TEXT
                 </a>
                </div>
               </div>
               <div>
                <div>
                 <a>
                  #TEXT
                 </a>
                </div>
               </div>
               <div>
                <div>
                 <a>
                  #TEXT
                 </a>
                </div>
               </div>
               <div>
                <div>
                 <a>
                  #TEXT
                 </a>
                </div>
               </div>
               <div>
                <div>
                 <a>
                  #TEXT
                 </a>
                </div>
               </div>
              </div>
             )?
             (
              (
               <a>
                #TEXT
               </a>
              )+
             )?
            </div>
           )?
          </div>
         )+
         )?
        </div>
       )?
       (
        <h3>
         <a>
          #TEXT
         </a>
        </h3>
       )?
      </div>
     )+
    )?
    (
     <div>
      (
       <div>
        (
         (
          <div>
           (
            <a>
             #TEXT
            </a>
           )?
          </div>
         )+
         (
          <ul>
           (
            <li>
             <a>
              #TEXT
             </a>
            </li>
           )+
          </ul>
         )?
        )?
       </div>
      )+
     )?
    </div>
   )?
  </div>
 )+
</body>
```

Manazara.si wrapper

```
<body>
 <div>
  <div>
   <div>
    <div>
     #TEXT
    </div>
    <div>
     <div>
      <div>
       <div>
        <ul>
         <li>
          <strong>
           #TEXT
          </strong>
         </li>
        </ul>
       </div>
      </div>
     </div>
    </div>
   </div>
   <div>
    <div>
     <div>
      <div>
       (
        <div>
         (
          <h1>
           #TEXT
          </h1>
         )?
        </div>
       )+
      </div>
      <div>
       <div>
        <div>
         <strong>
          <span>
           #TEXT
          </span>
         </strong>
        </div>
        <div>
         <ul>
          (
           <li>
            <a>
             #TEXT
            </a>
           </li>
          )+
         </ul>
        </div>
       </div>
       <div>
        <div>
         <dl>
          <dd>
           <ol>
            (
             <li>
              <a>
               #TEXT
              </a>
              #TEXT
             </li>
            )+
           </ol>
          </dd>
          <dd>
           <ol>
            (
             <li>
              <a>
               <span>
                #TEXT
               </span>
               #TEXT
               (
                <span>
                 #TEXT
                </span>
               )?
              </a>
              #TEXT
             </li>
            )+
           </ol>
          </dd>
          <dd>
           <ol>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
           </ol>
          </dd>
          <dd>
           <ol>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
            <li>
             <a>
              #TEXT
             </a>
             #TEXT
            </li>
           </ol>
          </dd>
         </dl>
        </div>
       </div>
       <p>
        #TEXT
        <a>
         #TEXT
        </a>
       </p>
      </div>
     </div>
    </div>
   </div>
  </div>
 </div>
</body>
```