Iskanje in ekstrakcija podatkov s spleta - 2. naloga

Anei Makovec, 63150186

April 2020

1 Uvod

Za 2. nalogo sem implementiral program za ekstrakcijo besedila iz spletnih strani. Podprte so 3 metode ekstrakcije: z regularnimi izrazi, z jezikom XPath ter z implementacijo avtomatske spletne ekstrakcije, pri čemer algoritem proizvede le ovojnico, ki bi jo lahko uporabili pri ekstrakciji dejanskega besedila.

2 Izbrani spletni strani

Za tretji primer spletnih strani sem izbral dva primera z domene bookdepository.com. Obe spletni strani vsebujeta seznam knjig, ki so naprodaj. Vsaka knjiga je obravnavana kot podatkovni zapis, vsebuje pa več atributov, ki jih prikazuje slika 1.

3 Implementacija

3.1 Regularni izrazi

Pri ekstrakciji besedila z uporabo regularnih izrazov sem uporabil sledeče izraze za posamezne atribute podatkovnih zapisov pripadajočih spletnih strani:

1. **overstock.com**

   - *Title*:

     ```
     <a\s*href="\S*"><b>([^V][\w-'.,\s()]*)<\Wb><\Wa>
     ```

   - *List price:*

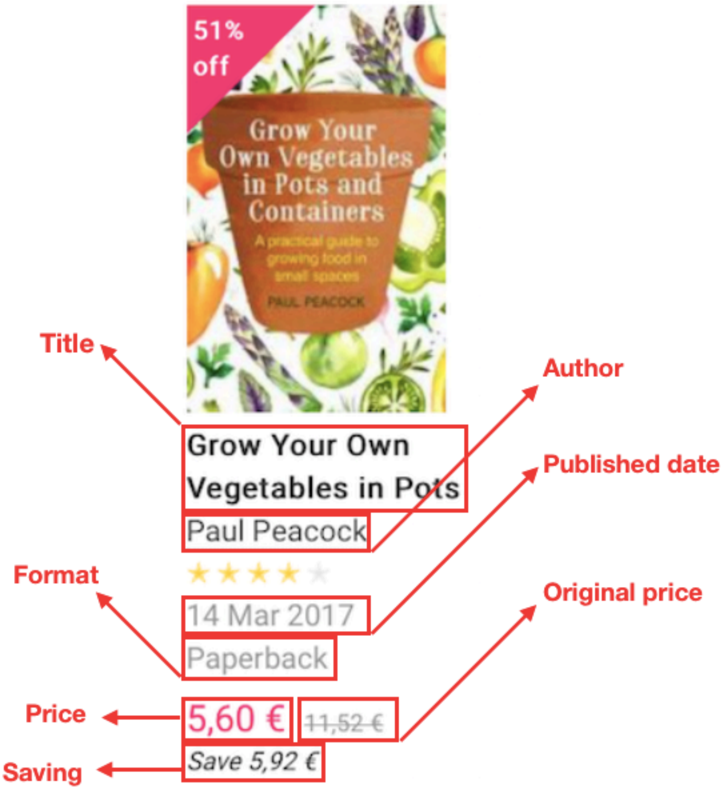
     ```
     List Price:<\/b><\/td><td align="\D*" nowrap="\D*"><s>(\$\d*,?\d*.\d*)
     ```

   - *Price:*

     ```
     Price:<\/b><\/td><td align="\D*" nowrap="\D*"><span class="bigred"><b>(\
         $\d*,?\d*.?\d*)
     ```

   - *Saving, Saving percentage:*

     ```
     You Save:<\/b><\/td><td align="\D*" nowrap="\D*"><span class="
         littleorange">(\$\d*,?\d*.?\d* \(\d*\%\))
     ```



Slika 1: Podatkovni zapis s svojimi atributi.

- *Content:*

```
<td valign="top"><span class="normal">([\s\S]*?)\s*<br><a href="[\S]*"><span class="tiny"><b>Click here to purchase[.]<\/b><\/span><\/a><\/span><br>
```

2. **rtvslo.si**

- *Title*:

```
<h1>([\s\S]*)<\/h1>
```

- *Subtitle*:

```
<div class="subtitle">([\s\D]*)<\/div>
```

- *Author, Published time*:

```
<div class="author-timestamp">\s*<strong>([\s\D]*)<\/strong>\|\s*(\d*\.\D*\d*\D*\d*\:\d*)\s*<\/div>
```

- *Lead*:

  ```
  <p class="lead">([a-zA-Z0žščŽŠČ-9.,\s]*)<\/p>
  ```

- *Content*:

  ```
  <p(?: class="Body")?>(?:<strong>)?((?!Test|Preizkus|žOgroena|Obuditev|
      Ekipa|Razvoj)[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)(?:<br>)?(?:<\/strong>)
      ?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ-
      -9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-
      zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)
      ?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ-
      -9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-
      zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*(?:<
      sub>)?\d?(?:<\/sub>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-
      Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br
      >)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ-
      -9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-
      zA-Z0žščŽŠČ--9",.:\/=-]*)?((?:<br>)?[\sa-zA-Z0žščŽŠČ--9",.:\/=-]*)?
  ```

3. **bookdepository.com**

- *Title*:

  ```
  <h3 class="title">\s*<a href="[\S]*">\s*([\sa-zA-Z0-9,.&\/;:()+*#?!'-]*)
      <br>\s*<\/a>\s*<\/h3>
  ```

- *Author*:

  ```
  <p class="author">\s*<a href="[\S]*" itemprop="author">([\sa-zA-Z'.,-]*)
      <\/a>\s*<\/p>
  ```

- *Published date*:

  ```
  <p class="published" itemprop="datePublished">([ \da-zA-Z]*)<\/p>
  ```

- *Format*:

  ```
  <p class="format">([a-zA-Z]*)<\/p>
  ```

- *Original price*:

  ```
  <span class="rrp">([ \d€,]*)<\/span>
  ```

- *Price*:

  ```
  <p class="price">\s*([ \d€,]*)\s*.*\s*<\/p>
  ```

- *Saving*:

  ```
  <p class="price-save">\s*Save ([ \d€,]*)<\/p>
  ```

Vsak atribut je pridobljen z enim samim regularnim izrazom. Nekateri atributi so hkrati pridobljeni z enim izrazom, kasneje pa sprocesirani z metodami manipulacije nizov tako, da se izlušči posamezen atribut. Na spletnih straneh, kjer je prisoten seznam podatkovnih zapisov, so atributi istega tipa vseh podatkovnih zapisov pridobljeni hkrati.

3.2 XPath

Pri ekstrakciji besedila z uporabo jezika XPath sem uporabil sledeče izraze za posamezne atribute podatkovnih zapisov pripadajočih spletnih strani:

1. **overstock.com**
   - *Title*:

     ```
     //td[@valign="top"]/a/b/text()
     ```

   - *List price:*

     ```
     //tr[td[1]/b[text()="List Price:"]]/td[2]/s/text()
     ```

   - *Price:*

     ```
     //tr[td[1]/b[text()="Price:"]]/td[2]/span/b/text()
     ```

   - *Saving, Saving percentage:*

     ```
     //tr[td[1]/b[text()="You Save:"]]/td[2]/span/text()
     ```

   - *Content:*

     ```
     //td[@valign="top"]/span[@class="normal"]/text()
     ```

2. **rtvslo.si**
   - *Title*:

     ```
     //h1/text()
     ```

   - *Subtitle*:

     ```
     //div[@class="subtitle"]/text()
     ```

   - *Author*:

     ```
     //div[@class="author-timestamp"]/strong/text()
     ```

   - *Published time*:

     ```
     //div[@class="author-timestamp"]/text()
     ```

   - *Lead*:

     ```
     //p[@class="lead"]/text()
     ```

   - *Content*:

     ```
     //p[@class="Body"]/text() | //p/strong/text() | //p[contains(text(),
         "-") or contains(text(), ":") or contains(text(), ",") and not(@class
         )]/text() | //p/sub/text()
     ```

3. **bookdepository.com**

- *Title*:

  ```
  //h3[@class="title"]/a/text()
  ```

- *Author*:

  ```
  //p[@class="author"]/a/text()
  ```

- *Published date*:

  ```
  //p[@class="published"]/text()
  ```

- *Format*:

  ```
  //p[@class="format"]/text()
  ```

- *Original price*:

  ```
  //span[@class="rrp"]/text()
  ```

- *Price*:

  ```
  //p[@class="price"]/text()
  ```

- *Saving*:

  ```
  //p[@class="price-save"]/text()
  ```

Vsak atribut je pridobljen z enim samim XPath izrazom. Nekateri atributi so hkrati pridobljeni z enim izrazom, kasneje pa sprocesirani z metodami manipulacije nizov tako, da se izlušči posamezen atribut. Na spletnih straneh, kjer je prisoten seznam podatkovnih zapisov, so atributi istega tipa vseh podatkovnih zapisov pridobljeni hkrati.

3.3 Algoritem za avtomatsko spletno ekstrakcijo

Algoritem sem implementiral po vzoru programa Webstemmer [1], njegovo delovanje pa opisuje procedura 1. Algoritem kot vhod prejme dve spletni strani istega tipa. Vsako najprej prebere ter iz HTML formata izlušči vse elemente, ki jih nato organizira v drevesno strukturo (vrstici 2 in 3). Sprehodi se po obeh drevesih ter ustvari seznam elementov po vrstnem redu pojavitve (vrstici 5 in 6). Sedaj primerja dobljena seznama ter poskuša najti sorodne elemente ter tako oblikuje vzorec sorodnih elementov (vrstici 8 in 9) za vsako spletno stran posebej. Nato za vsak element v njem izračuna *Diff Score*, ki nam pove za koliko se njegovo besedilo razlikuje med spletnima stranema (vrstica 11), *Main Score*, ki nam pove kakšno informacijsko vrednost ima njegovo besedilo (vrstica 12), naposled pa še *Layout Pattern score*, ki nam pove kakšno informacijsko vrednost ima celoten vzorec ter koliko je uporaben (vrstica 13). Na koncu pa še izpiše zgenerirano ovojnico na standardni izhod.

Pri izračunu posameznih vrednosti sem uporabil enačbe, ki jih prikazuje slika 2. *Diff Score* moramo izračunati, da ugotovimo, če se besedilo določenega elementa med spletnima stranema razlikuje. Tako lahko elemente katerih ta vrednost je 0 ali manjša od neke predefinirane vrednosti k kasneje v ekstrakciji izločimo, saj to pomeni, da gre najverjetneje za statično besedilo, ki pa nas ne zanima. *Main Score* je pomemben, ker lahko z njegovo pomočjo identificiramo elemente, ki vsebujejo glavno besedilo, ki je najverjetneje tisto, kar iščemo. *Layout Pattern Score* pa je koristen, če imamo več različnih ovojnic, ki smo jih pridobili iz večjega števila

Algorithm 1 Generiranje ovojnice.

```
procedure AutoWeb($Text_1$, $Text_2$)
    $tree_1 \leftarrow read(Text_1)$
    $tree_2 \leftarrow read(Text_2)$

    $layoutblocks_1 \leftarrow parsePage(tree_1)$
    $layoutblocks_2 \leftarrow parsePage(tree_2)$

    $layoutpattern_1 \leftarrow genLayoutPattern(layoutblocks_1)$
    $layoutpattern_2 \leftarrow genLayoutPattern(layoutblocks_2)$

    $diffscores \leftarrow calcDiffscores(layoutpattern_1, layoutpattern_2)$
    $mainscores \leftarrow calcMainscores(layoutpattern_1, layoutpattern_2, diffscores)$
    $layoutpatternscore \leftarrow calcLayoutPatternScore(mainscores)$

    $printWrapper()$
```

spletnih strani grupiranih po sorodnosti. Takrat nam ta vrednost pove, v kolikšni meri je določena ovojnica uporabna oz. koliko besedila, ki ga iščemo bomo lahko izluščili z uporabo določene ovojnice.

Nasjednje podpoglavje prikazuje pridobljene ovojnice za vsak tip spletnih strani. Ker je izpis ovojnic precej dolg, sem podal le dele ovojnic, ki so najbolj zanimive. Ovojnica za *overstock.com* vključuje le nekaj atributov njenih podatkovnih zapisov, kot npr. *Title*, *Price* ali *List price*, vendar pa ne od vseh. Ovojnica za *rtvslo.si* je boljša, saj vključuje več atributov: *SubTitle*, *Author*, *PublishedTime* ter *Content.* Ovojnica za *bookdepository.com* pa spet vključuje le posamezne atribute, kot npr. *Author*, *Title*, *Price*, *OriginalPrice*, a bolj konsistentno, kot ovojnica za *overstock.com.* Na splošno mislim, da je ta implementacija algoritma za generacijo ovojnice primernejša za spletne strani, ki ne vsebujejo seznamov objektov temveč le besedilo, kot npr. članki in razne objave.

$$DiffScore(E) = (\sum_{s1,s2inE} |s1| + |s2| - 2 * |MCS(s1,s2)|)/(\sum_{s1,s2inE} |s1| + |s2|)$$

$$MainScore(E) = DiffScore(E) * (\sum_{sinE} |s|)/|E|$$

$$LayoutPatternScore(P) = log(|P|) * (\sum_{EinP} MainScore(E))$$

Slika 2: Enačbe uporabljene pri izračunih vrednosti [1]. P - vzorec sorodnih elementov, E - element v vzorcu, s1, s2 - besedilo elementa v posamezni spletni strani, MCS - največje skupno zaporedje črk (*ang. maximum common sequence*).

3.3.1 Ovojnica za overstock.com

```
# 2630.3480627195454 overstock.com/jewelry01.html (pattern 1)
#       (Pages which belong to this cluster)
#       overstock.com/jewelry01.html
#       overstock.com/jewelry02.html
(2630.3480627195454,                              (Overall score of the cluster)
 'overstock.com/jewelry01.html', (Cluster ID)
 [                                                       (List of layout blocks)
  # (diffscore, mainscore, block feature)
  (0.0, 0.0, title),
  ...
  ...
  (0.6666666666666666, 8.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border=2:
        cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/span:class=littleorange/b),
  ...
  ...
  (0.28, 7.000000000000001, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border=2:
        cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=187),
  ...
  ...
  (0.7692307692307693, 20.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border
        =2:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=1136),
  ...
  ...
  (0.7391304347826086, 17.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border
        =2:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=188),
  ...
  ...
  (0.42857142857142855, 12.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border
        =2:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=1031),
  ...
  ...
  (0.7647058823529411, 13.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border
        =2:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=185),
  ...
  ...
  (0.8620689655172413, 25.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#d2dbfb:border
        =2:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td/table/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&
        PRO_SSUB_CAT=602),
  ...
  ...
  (0.6, 9.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/span:class=littleorange/b),
  ...
  ...
  (0.7209302325581395, 30.999999999999996, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#
        eeeeee:border=0:cellpadding=3:cellspacing=0:width=100%/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STORELIST&STO_ID=7),
  ...
  ...
  (0.7727272727272727, 34.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border
        =0:cellpadding=3:cellspacing=0:width=100%/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STORELIST&STO_ID=3),
  ...
  ...
  (0.673469387755102, 33.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border=0:
        cellpadding=3:cellspacing=0:width=100%/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STORELIST&STO_ID=2),
  ...
  ...
  (0.75, 24.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border=0:cellpadding
        =3:cellspacing=0:width=100%/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STORELIST&STO_ID=1),
  ...
  ...
  (0.8518518518518519, 23.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border
        =0:cellpadding=3:cellspacing=0:width=100%/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STORELIST&STO_ID=4),
  ...
  ...
  (0.8125, 26.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border=0:cellpadding
        =3:cellspacing=0:width=160/tbody/tr/td/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=STATICPAGE&PAGE_ID=5),
  (1.0, 3.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border=0:cellpadding=3:
        cellspacing=0:width=160/tbody/tr/td),
  (0.7692307692307693, 20.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#eeeeee:width=160:valign=top/table:bgcolor=#eeeeee:border
        =0:cellpadding=3:cellspacing=0:width=160/tbody/tr/td/a:href=http://www.overstockb2b.com/),
  ...
  ...
  (0.6666666666666666, 8.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#ffffff:valign=top:align=left/table:width=100%/tbody/tr/td:
        align=center/table:border=1:cellpadding=0:cellspacing=0/tbody/tr/td/table:width=100%:border=0:cellpadding=0:cellspacing=5:bgcolor=#d2dbfb/tbody/tr/td:
        nowrap=nowrap/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&PRO_SSUB_CAT=999&CAT_SORTBY=3),
  (0.7777777777777778, 7.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#ffffff:valign=top:align=left/table:width=100%/tbody/tr/td:
        align=center/table:border=1:cellpadding=0:cellspacing=0/tbody/tr/td/table:width=100%:border=0:cellpadding=0:cellspacing=5:bgcolor=#d2dbfb/tbody/tr/td:
        nowrap=nowrap/a:href=http://www.overstock.com/cgi-bin/d2.cgi?PAGE=CATLIST&PRO_SUB_CAT=307&PRO_SSUB_CAT=999&CAT_SORTBY=1),
  ...
  ...
  (0.38461538461538464, 10.0, table:border=0:cellpadding=0:cellspacing=0:width=770/tbody/tr/td:bgcolor=#ffffff:valign=top:align=left/table:width=100%/tbody/tr/td
        /table:border=0:cellpadding=0:cellspacing=0:width=100%/tbody/tr/td:valign=top/table:border=0:cellpadding=2:cellspacing=0:width=100%/tbody/tr:bgcolor=#
        dddddd/td:valign=top/table/tbody/tr/td:valign=top/table/tbody/tr/td:align=right:nowrap=nowrap/b),
  ...
  ...
 ]
)
```

3.3.2 Ovojnica za rtvslo.si

```
# 19330.038271789537 rtvslo.si/Audi A6 50 TDI quattro_ nemir v premijskem razredu - RTVSLO.si.html (pattern 1)
#       (Pages which belong to this cluster)
#       rtvslo.si/Audi A6 50 TDI quattro_ nemir v premijskem razredu - RTVSLO.si.html
#       rtvslo.si/Volvo XC 40 D4 AWD momentum_ suvereno med˘najboljse v razredu - RTVSLO.si.html
(19330.038271789537,                        (Overall score of the cluster)
 'rtvslo.si/Audi A6 50 TDI quattro_ nemir v premijskem razredu - RTVSLO.si.html', (Cluster ID)
 [                                                 (List of layout blocks)
  # (diffscore, mainscore, block feature)
  (0.4666666666666667, 12.6, title),
  ...
  ...
  (1.0, 7.333333333333333, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div:
        class=row/header:class=article-header/div:class=subtitle),
  (0.0, 0.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div:class=row/
        header:class=article-header/div:class=article-meta-mobile),
  (0.2727272727272727, 1.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div:
        class=row/header:class=article-header/div:class=article-meta-mobile/div:class=author-timestamp),
  ...
  ...
  (0.5555555555555556, 1.6666666666666667, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old
        article-type-1/div:class=row/div:class=article-meta/div:class=author),
  (0.8297872340425532, 13.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div
        :class=row/div:class=article-meta/div:class=publish-meta),
  ...
  ...
  (0.9885714285714285, 519.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/
        div:class=row/div:class=article-body/article:class=article/p),
  (0.9769230769230769, 254.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/
        div:class=row/div:class=article-body/article:class=article/p),
  (0.979933110367893, 293.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div
        :class=row/div:class=article-body/article:class=article/p),
  (0.0, 0.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div:class=row/div:
        class=article-body/article:class=article/p),
  (0.7777777777777778, 7.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div:
        class=row/div:class=article-body/article:class=article/p),
  (0.5675675675675675, 42.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div
        :class=row/div:class=article-body/article:class=article/p),
  (0.8666666666666667, 39.0, div:id=main-container:class= container article-container:data-id=475392/div:class=news-container blue article-old article-type-1/div
        :class=row/div:class=article-body/article:class=article/p),
  ...
  ...
 ]
)
```

3.3.3 Ovojnica za bookdepository.com

```
(8371.951572252485,                        (Overall score of the cluster)
 'bookdepository.com/Book_Depository_1.html', (Cluster ID)
 [                                                 (List of layout blocks)
  # (diffscore, mainscore, block feature)
  (0.0, 0.0, title),
  ...
  ...
  (0.3898305084745763, 11.5, div:class=page-slide/div:class=content-wrap/div:class=main-content/h1),
  ...
  ...
  (1.0, 45.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=format),
  (0.8297872340425532, 39.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
        div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
        /div:class=item-info/div:class=price-wrap/p:class=price),
  ...
  ...
  (1.0, 46.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=format),
  (0.8333333333333334, 40.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
        div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
        /div:class=item-info/div:class=price-wrap/p:class=price),
  ...
  ...
  (1.0, 46.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=format),
  (0.875, 42.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=
        block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class
        =item-info/div:class=price-wrap/p:class=price),
  ...
  ...
  (1.0, 45.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=format),
  (0.8723404255319149, 41.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
        div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
        /div:class=item-info/div:class=price-wrap/p:class=price),
  ...
  ...
  (1.0, 45.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=format),
  (0.8723404255319149, 41.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
        div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
```

```
       /div:class=item-info/div:class=price-wrap/p:class=price),
...
...
(1.0, 45.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=format),
(0.8723404255319149, 41.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/div:class=price-wrap/p:class=price),
...
...
(1.0, 45.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=format),
(0.8723404255319149, 41.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/div:class=price-wrap/p:class=price),
...
...
(1.0, 52.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/h3:class=title/a:href=https://www.bookdepository.com/Overcoming-Social-Anxiety-Shyness-2nd-Edition-Gillian-Butler/9781472120434),
...
...
(1.0, 15.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=author/a:href=https://www.bookdepository.com/author/William-Glasser:itemprop=author),
...
...
(0.8518518518518519, 23.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-actions/div:class=btn-wrap/a:rel=nofollow:href=https://www.bookdepository.com/basket/addisbn/isbn13/9781905862481:class=btn btn-sm btn-
       primary add-to-basket:data-ref=grid-view:data-isbn=9781905862481:data-currency=EUR:data-price=8.37:data-show-related=false),
(1.0, 15.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       img/a:href=https://www.bookdepository.com/Orange-Blossom-Honey-John-Gregory-Smith/9780857834157:itemprop=url),
(1.0, 13.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       img/a:href=https://www.bookdepository.com/Orange-Blossom-Honey-John-Gregory-Smith/9780857834157:itemprop=url),
...
...
(0.8181818181818182, 27.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/h3:class=title/a:href=https://www.bookdepository.com/Gentle-Sleep-Book-Sarah-Ockwell-Smith/9780349405209),
(0.7894736842105263, 15.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/h3:class=title/a:href=https://www.bookdepository.com/ToddlerCalm-Sarah-Ockwell-Smith/9780349401058),
(0.7692307692307693, 20.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/p:class=author/a:href=https://www.bookdepository.com/author/Sarah-Ockwell-Smith:itemprop=author),
...
...
(1.0, 29.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=author),
...
...
(1.0, 6.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block /
       div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/h3:class=title/a:href=https://www.bookdepository.com/Glow15-Naomi-Whittel/9781912023639),
(1.0, 6.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block /
       div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/h3:class=title/a:href=https://www.bookdepository.com/Glow15-Naomi-Whittel/9781912023639),
(1.0, 13.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=author/a:href=https://www.bookdepository.com/author/Naomi-Whittel:itemprop=author),
...
...
(1.0, 13.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=author/a:href=https://www.bookdepository.com/author/Matthew-Biggs:itemprop=author),
...
...
(1.0, 24.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/h3:class=title/a:href=https://www.bookdepository.com/Claridges-Cookbook-Martyn-Nail/9781784723293),
(1.0, 24.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/h3:class=title/a:href=https://www.bookdepository.com/Claridges-Cookbook-Martyn-Nail/9781784723293),
(1.0, 11.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
       /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
       info/p:class=author/a:href=https://www.bookdepository.com/author/Martyn-Nail:itemprop=author),
...
...
(0.65, 26.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item
       -info/h3:class=title/a:href=https://www.bookdepository.com/Flowers-Colouring-Book-Arcturus-Publishing/9781782121800),
(0.7368421052631579, 28.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/h3:class=title/a:href=https://www.bookdepository.com/Flowers-Colouring-Book-Arcturus-Publishing/9781782121800),
(0.7333333333333333, 22.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/
       div:class=block /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book
       /div:class=item-info/p:class=author/a:href=https://www.bookdepository.com/author/Arcturus-Publishing:itemprop=author),
...
```

```
  ...
  (1.0, 12.0, div:class=page-slide/div:class=content-wrap/div:class=main-content/div:class=content-block/div:class=block-wrap search  no-heading/div:class=block
        /div:class=tab-wrap module type-book grid tab--2 tab-active/div:class=tab/div:class=book-item:itemscope=:itemtype=http://schema.org/Book/div:class=item-
        info/p:class=author/a:href=https://www.bookdepository.com/author/Atticus-Lish:itemprop=author),
  ...
  ...
 ]
)
```

Literatura

[1] Yusuke Shinyama. Webstemmer - how it works? http://www.unixuser.org/~euske/python/webstemmer/howitworks.html. Dostopano: 22.4.2020.