Regularni izrazi in Xpath - Iskanje in ekstrakcija podatkov iz spleta

Robert Košir, Aljoša Omejc

May 1, 2020

1 Uvod

V naslednjih nekaj poglavjih bomo na kratko opisali projekt pri predmetu Iskanje in ekstrakcija podatkov s spleta. Sam projekt se ukvarja z zajemom podatkov iz različnih strani z uporabo regularnih izrazov, xpathi in automatičnim algoritmom. Ob zagonu programa podate arugemnt, ki določi, katero metodo bo program uporabil.

2 Definicija spletne strani slovenskenovice.si

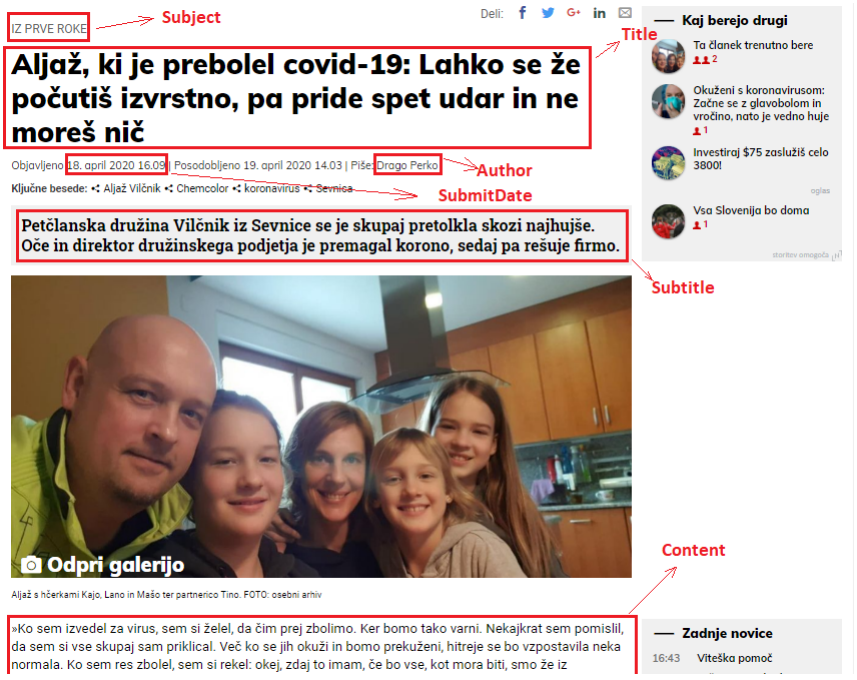

Figure 1: Definicija objektov za spletno stran slovenske novice

3 Regularni izrazi

V naslednjih podpoglavjih bom v obliki tabele podal regularne izraze za vsako stran posebaj.

3.1 rtvslo.si

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

Sam content nam je v tem primeru vrnil cel html niz, kar pa smo dalje procesirali na naslednji način:

- Birsanje ¡Script¿ tagov in njene vsebine
  `re.sub(r"<script([\S\s]*?)>([\S\s]*?)<\/script>", "", content)`
- Brisanje vseh HTML tagov
  `re.sub(r"<[^>]*>", "", content)`
- Brisanje vseh `\s`
  `re.sub(r"\s+", " ", content)`

3.2 overstock.com

Title	`<a.*<b>(.*)<\/b>\n*<\/a> `
ListPrice	`<td align=\"left\" nowrap=\"nowrap\"><s>(.*)<\/s><\/td>`
Price	`<span class=\"bigred\"><b>(.*)<\/b><\/span>`
Saving	`<span class=\"littleorange\">(.*) \(.*\)<\/span>`
SavingPercent	`<span class=\"littleorange\">\$.* (.*)<\/span>`
Content	`<td valign=\"top\"><span class=\"normal\">([\s\S]*?)` ` <a href.*><span class=\"tiny\"><b>(.*)<\/b>`

3.3 slovenskenovice.si

Title	`<h1 class=\"itemTitle\">\s*([\s\S]*)<\/h1>`
Subject	`<span class=\"itemSuperscript\">(.*)<\/span>`
SubmitDate	`<span class=\"itemDatePublished\">\s*Objavljeno (.*)<\/span>`
Author	`<span class=\"itemAuthor\">\s*Piše:\s+<span>(.*)<\/span>`
Subtitle	`<h2 class=\"itemSubtitle\">\s*<span>(.*)<\/span>`
Content	`<div class=\"itemFullText\" .*>([\s\S]*)<div class=\"itemInfoboxText\"`

4 Xpath izrazi

V naslednjih podpoglavjih bom v obliki tabele podal xpath izraze. ki so iz vseh strani prebrale vnaprej določene objekte.

4.1 rtvslo.si

Author	`//*[@id="main-container"]/div[3]/div/div[1]/div[1]/div`
PublishedTime	`//*[@id="main-container"]/div[3]/div/div[1]/div[2]/text()[1]`
Title	`//*[@id="main-container"]/div[3]/div/header/h1`
SubTitle	`//*[@id="main-container"]/div[3]/div/header/div[2]`
Lead	`//*[@id="main-container"]/div[3]/div/header/p`
Content	`//*[@id="main-container"]/div[3]/div/div[2]` `//*[not(self::script)]/text()`

4.2 overstock.com

Title	`/html/body/table[2]/tbody/tr[1]/td[5]/table/tbody/tr[2]/td` `/table/tbody/tr/td/table/tbody/tr/td/a/b`

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`

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`

Saving(Percent)	`/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`

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`

Tu smo z uporabo enega xpatha pridobili Saving in SavingPercent, ki pa smo ju kasneje razcepili.

4.3 slovenskenovice.si

Title	`//*[@id="ocmContainer"]/div[1]/div/div[1]/div[1]/div/div[3]/h1`
Subject	`//*[@id="ocmContainer"]/div[1]/div/div[1]/div[1]/div/div[2]/span`
SubmitDate	`//*[@id="ocmContainer"]/div[1]/div/div[1]/div[1]/div/div[4]/span /span[1]`
Author	`//*[@id="ocmContainer"]/div[1]/div/div[1]/div[1]/div/div[4]/span /span[5]/span`
Subtitle	`//*[@id="ocmContainer"]/div[1]/div/div[1]/div[1]/div/div[6]/h2/span`
Content	`//*[@id="ocmContainer"]/div[1]/div/div[2]/div[1]/div[1]/div[1] //text()[not(parent::script)]`

5 Automatska ekstrakcija

5.1 Psevdokoda

Algorithm 1: Get Page Wrapper

Result: wrapper
page1,page2 = removeNewLinesOnContent();
diffLines = difflib.compare(page1, page2);
pairs = GetPairs(diffLines) ;
pairsAttr = ShortenAttributes(pairs);
result = replaceDynamicText(pairsAttr, page1);
soup = BeautifulSoup(result);
RemoveUselessTags(soup);
RemoveUselessAttributs(soup);

5.2 Description

5.2.1 Difflib.compare()

s primerjanjem dobimo skupaj združeni datoteki, kjer se vrstice z začnejo z [" ", "-", "+", "?"]

```
- "-" line unikatna vrstica daototke 1
- "+" line unikatna vrstica daototke 2
- " " line enak v obeh
- "?" šum
```

5.2.2 GetPairs()

Pare pridobimo tako, da gledamo zaporedne vrstice izbiramo po metrikah: - enaka začetna oznaka - dovolj podobni atributi

5.2.3 ShortenAttributes()

Skrajšamo vse atribute tako, da jih primerjamo in vzamemo začetni del, kjer sta enaka.

```
 Vhod:
    page1: <h4 class="block-title blue">
    page2: <h4 class="block-title green">
Izhod:
    <h4 class="(block-title .*)"
```

5.2.4 ReplaceDynamicText()

Za vsak par pogledava, če ima različno vsebino in v primeru, da ima potem delava podobno, kot pri roadrunnerju, da zamenjava vsebino z oznako "#text"

```
Vhod:
    -<div class="subtitle">Test novega  modela</div>
    +<div class="subtitle">Test nove generacije</div>
Izhod:
    <div class="subtitle">#text</div>
```

5.2.5 RemoveUselessTags()

Odstraninila sva oznake [iframe, img, br, footer, nav, script], ker se nama je zdelo, da večinoma ne držijo nobenih uporabnih informacij.

Obdrži vse oznake, ki same ali pa njihovi nasledniki vsebujejo #text. Na ta način dobimo drevo, kjer listi vsebujejo koristne informacije.

5.2.6 RemoveUselessAttributs()

Obdržimo samo atributa class in id, ki pomagata navigirati po drevesu. V primeru, da bi imela več prostora bi obdržala tudi nekatere druge kot so title, src.

5.3 Opcijske značke

Za implementacijo opcijskih značk nama je zmanjkalo časa. Vendar bi jo implemenitrala tako, da vzameva vse različne vrstice iz datoteke1 in datoteke2, ki nimajo svojega para.

5.4 Izhod

5.4.1 Rtv1

<html><body class="first-page _www hide-submenu article user-logged-out"><div class=""
id="main-bar"> <div class="col-12 d-flex
align-items-center"><ul id="main-menu"><li class="green"><a>#text </d
iv> <div class="section-heading
blue"><h3 class="section-title animated-circles-onhover"><a>#text</h3> <div
class="edit-btn-container" id="article-edit-btn">#text</di
v> <header
class="article-header"> <h3 class="section-title
animated-circles-onhover"><a>#text</h3> <h1>#text</h1> #text
#text
<div class="swiper-container swiper-header-gallery mobile-show-only
swiper-container-horizontal swiper-container-wp8-horizontal"> <figure cla
ss="photoswipe swiper-slide mobile-show-only"><figcaption>#text</figcaption></figure> <p
class="lead">#text</p></header> <figure class="c-figure-emphasis"><ul
class="emphasis-list">
#text
</figure> <div class="ar
ticle-body"><article class="article"> <div class="swiper-container swiper-footer-gallery
swiper-container-horizontal swiper-container-wp8-horizontal"> <figure
class="photoswipe swiper-slide"><figcaption>#text</figcapti
on></figure><figure class="photoswipe swiper-slide"><figcaption>#text</figcaption></figure><figure
class="photoswipe swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="p
hotoswipe swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe swiper-slide"><fig
caption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></
figure><figure class="photoswipe swiper-slide"><figcaption>#text</figcaption></figure><figure
class="photoswipe swiper-slide"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide"><figcaption>#text</figcaption></figure><figure class="photos
wipe swiper-slide swiper-slide-prev"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide swiper-slide-active"><figcaption>#text</figcaption></figure><figure class="photoswipe
swiper-slide swiper-slide-next"><figcaption>#text</figcaption></
figure><figure class="photoswipe swiper-slide"><figcaption>#text</figcaption></figure> <div
class="gallery-bottom-thumbs"><figure class="photoswipe
gallery-bottom-thumb"><figcaption>#text</figcaption></figure><figure class="photoswipe
gallery-bottom-thum
more-photos"><figcaption>#text</figcaption></figure> </article> <div clas
s="article-column"> <h4 class="(block-title
.*)">#text</h4> <a
class="list-title">#text <div class="ar
ticle-container"> <a
class="list-title">#text <div

more-photos"><figcaption>#text</figcaption></figure> </article> <div clas
s="article-column"> <h4 class="(block-title
.*)">#text</h4> <a
class="list-title">#text <div class="ar
ticle-container"> <a
class="list-title">#text <div
class="article-tags">#text#text <d
iv class="share">#text <div class="news-block
blue"> <h3><span
class="news-cat"><a>#text</h3> <div class="col-lg-3 col-md-6"
> <h3><a>#text</h3> <div
class="col-lg-3 col-md-6"> <h3><span
class="news-cat"><a>#text</h3> <div
id="article-comments-anchor"><div class="section

-heading blue"><h3 class="section-title animated-circles-onhover"><a
class="btn-show-comments">#text</h3> <div class="edit-btn-container"
id="article-comments-edit-btn">#text <div class="news-block
article-comments article-comm
ents-toggle hide-comments"><p
class="hidden-comments-notice">#text</p> </body></html>

5.4.2 Overstock1

<html><body><table><tbody><tr><td><table><tbody><tr><td><table><tbody><tr><td><span
class="littlewhite">#text</td><td>#text</td><td>#text</td></tr></tbody></table></td><td>#text</td>
</tr></tbody></table></td></tr></tbody></table><table><tbody><tr><td><table>
<tbody><tr><td><table><tbody><tr><td>#text</td></tr><tr><td>#text</td></tr><tr><td>#text</td></tr><tr><
td>#text</td></tr><tr><td>#text</td></tr><tr><td>#text</td></tr></tbody></table></td></tr></tbody></table
>#text#text#text</td><td><table><tbody><tr><td><table><tb
ody><tr><td><table><tbody><tr><td>#text</td></tr></tbody></table></td></tr></tbody></table></td></tr
><tr><td><table><tbody><tr><td><table><tbody><tr><td><table><tbody><tr><td><a>#text</td></tr><
/tbody></table></td><td><a
>#text<table><tbody><tr><td><table><tbody><tr><td>#text</td></tr><tr><td>#tex
t</td></tr><tr><td>#text</td></tr></tbody></table></td></tr></tbody></table></td></tr><tr><td
><table><tbody><tr><td><a>#text</td></tr></tbody></table></
td><td><a>#text</td></tr><tr><td><table><tbody><tr><td><table><tbody><tr><td>#text</
b></td></tr><tr><td>#text</td></tr><tr><td>#text</td></tr></tbody></table></td></tr></tbo
dy></table></td></tr><tr><td><table><tbody><tr><td><span class="
normal">#text</td></tr></tbody></table></td></tr><tr><td><table><tbody><tr><td><a>#text
</td></tr></tbody></table></td></tr><tr><td><a>#text</td></tr><tr><td><table><tbody><tr><t
d><table><tbody><tr><td>#text</td></tr></tbody></table></td
></tr></tbody></table></td></tr><tr><td><table><tbody><tr><td><table><tbody><tr><td>#text</td
></tr><tr><td>#text</td></tr></tbody></table></td></tr></tbody></table></td></tr><tr><td><table>
<tbody><tr><td><table><tbody><tr><td>#text</td></tr></t
body></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
class="littleorange">#text</td></tr></tbody></table></td></tr></tbody></table></td></tr
></tb
ody></table></td></tr><tr><td>#text</td></tr></tbody></table></body></html>

5.4.3 Slovenske novice

```
<html class="com_ocm view-item itemid-158 j38 mm-hover no-touch"><body><div
class="t3-wrapper"><header class="container topmenu_cont t3-header outter_cont" id="t3-header"><div
class="row"><div class="col-xs-12 visible-lg visible-md"><div class="topmenu"><div class=
"custom topmenubar24"><div id="bar24"><div class="bar24"><div class="bar24-title"><span><span
class="VigLink_is_cancer_of_the_internet">#text</span><span
class="VigLink_is_cancer_of_the_internet">#text</span></span></div></div></div></div></div></div></d
></header
><div class="t3-content" id="t3-content"><div class="itemView itemIsFeatured"
id="ocmContainer"><div><div class="container inner_cont"><div class="row col_grid"><div class="col-lg-
col-md-6 col-sm-12 col-xs-12 stickycalc1-0"><div class="itemHeader well"><div class
="col-xs-12 col_nest"><span class="itemSuperscript">#text</span></div><div class="col-xs-12
col_nest"><div class="itemTags"><a class="itemTag"><span
class="itemTagName">#text</span></a></div></div><div class="col-xs-12 col_nest"><h2
class="itemSubtitle"><span>#text
</span></h2></div></div></div><div class="col-lg-3 col-md-3 col-sm-12 col-xs-12 stickycalc1-1"><div
class="col-xs-12 col_nest col-sticky stickycalc1-2"><div><div id="livenetlife_links"><div><div
class="livenetlife_links"><div class="livenetlife_links_title">#text</
div></div></div></div></div></div></div><div class="row col_grid col_catad"><div class="col-lg-3
col-md-3 col-sm-12 col-xs-12 item_right_02 stickycalc2-1"><div class="col-xs-12 col_nest col-sticky
stickycalc2-2"><div><div id="outbrain768"><div class="OUTBRAIN
" id="outbrain_widget_0"><div class="ob-widget ob-strip-layout SB_1"><div class="ob-widget-section
ob-first"><ul class="ob-widget-items-container"><li class="ob-dynamic-rec-container ob-recIdx-0 ob-o"><a
class="ob-dynamic-rec-link"><span class="ob-unit ob-rec-text"
>#text</span></a></li></ul><ul class="ob-widget-items-container ob-multi-row ob-row-1"><li
class="ob-dynamic-rec-container ob-recIdx-1 ob-o"><a class="ob-dynamic-rec-link"><span class="ob-unit
ob-rec-text">#text</span></a></li></ul><ul class="ob-widget-items-contai
ner ob-multi-row ob-row-2"><li class="ob-dynamic-rec-container ob-recIdx-2 ob-o"><a
class="ob-dynamic-rec-link"><span class="ob-unit ob-rec-text">#text</span></a></li></ul><ul
class="ob-widget-items-container ob-multi-row ob-row-3"><li class="ob-dynamic-rec-contain
er ob-recIdx-3 ob-o"><a class="ob-dynamic-rec-link"><span class="ob-unit
ob-rec-text">#text</span></a></li></ul></div></div></div></div><div class="widget priporocamo"><div
class="well"><div class="well mod_107 card_small"><article><div class="card_article"><a><h3
class="h-title"><span
class="card_title">#text</span></h3></a></div></article></div></div></div></div></div></div></div></
    v></div><div class="container item_break_02_cont outter_cont"><div class="row col_grid"><div
    class="col-xs-12"><div><div class="custom"><div
    class="OUTBRAIN" id="outbrain_widget_1"><div class="ob-widget ob-strip-layout AR_1"><div
    class="ob-widget-section ob-first"><ul class="ob-widget-items-container"><li
    class="ob-dynamic-rec-container ob-recIdx-0 ob-o"><a class="ob-dynamic-rec-link"><span class="ob
    it ob-rec-text">#text</span></a></li><li class="ob-dynamic-rec-container ob-recIdx-1 ob-o"><a
    class="ob-dynamic-rec-link"><span class="ob-unit ob-rec-text">#text</span></a></li></ul><ul
    class="ob-widget-items-container ob-multi-row ob-row-1"><li class="ob-dynamic-r
    ec-container ob-recIdx-7 ob-o"><a class="ob-dynamic-rec-link"><span class="ob-unit
    ob-rec-text">#text</span><span class="ob-unit
    ob-rec-source">#text</span></a></li></ul></div></div></div><div class="OUTBRAIN"
    id="outbrain_widget_2"><div class="ob-widget ob-strip-l
```

```
v></div><div class="container item_break_02_cont outter_cont"><div class="row col_grid"><div
class="col-xs-12"><div><div class="custom"><div
class="OUTBRAIN" id="outbrain_widget_1"><div class="ob-widget ob-strip-layout AR_1"><div
class="ob-widget-section ob-first"><ul class="ob-widget-items-container"><li
class="ob-dynamic-rec-container ob-recIdx-0 ob-o"><a class="ob-dynamic-rec-link"><span class="ob
it ob-rec-text">#text</span></a></li><li class="ob-dynamic-rec-container ob-recIdx-1 ob-o"><a
class="ob-dynamic-rec-link"><span class="ob-unit ob-rec-text">#text</span></a></li></ul><ul
class="ob-widget-items-container ob-multi-row ob-row-1"><li class="ob-dynamic-r
ec-container ob-recIdx-7 ob-o"><a class="ob-dynamic-rec-link"><span class="ob-unit
ob-rec-text">#text</span><span class="ob-unit
ob-rec-source">#text</span></a></li></ul></div></div></div><div class="OUTBRAIN"
id="outbrain_widget_2"><div class="ob-widget ob-strip-l
ayout AR_2"><div class="ob-widget-section ob-first"><ul class="ob-widget-items-container"><li
class="ob-dynamic-rec-container ob-recIdx-2 ob-o"><a class="ob-dynamic-rec-link"><span class="ob
ob-rec-text">#text</span></a></li><li class="ob-dynamic-rec-container
 ob-recIdx-3 ob-o"><a class="ob-dynamic-rec-link"><span class="ob-unit
ob-rec-text">#text</span></a></li></ul><ul class="ob-widget-items-container ob-multi-row ob-row-1">
class="ob-dynamic-rec-container ob-recIdx-4 ob-o"><a class="ob-dynamic-rec-link"><span clas
s="ob-unit ob-rec-text">#text</span></a></li><li class="ob-dynamic-rec-container ob-recIdx-5 ob-o">
class="ob-dynamic-rec-link"><span class="ob-unit
ob-rec-text">#text</span></a></li></ul></div></div></div></div></div></div></div></div></div></div>
</body></
```

```
html>
```

6 Zaključek

Sama seminarska naloga je bila zelo zanimiva. Njaveč problem nama je povzročala 3 naloga, a vendar mislim, da sva uspešon opravila večino dela.