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

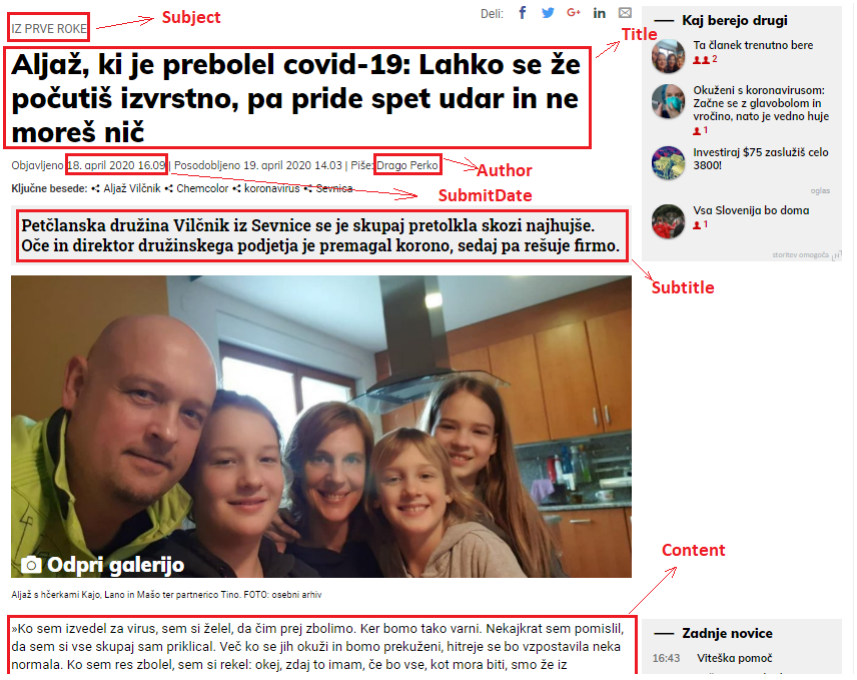

Figure 1: Definicija objektov za spletno stran slovenske novice

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.

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.