Marko Marinković, Tit Arnšek in Damijan Randl

Poročilo

2. del projekta pri predmetu Iskanje in ekstrakcija podatkov s spleta

# Povzetek

V poročilu so predstavljeni rezultati treh postopkov ekstrakcije podatkov s spleta. To so:

* Ekstrakcija s pomočjo **regularnih izrazov**
* Ekstrakcija s pomočjo **xpath-a**
* Avomatizirana ekstrakcija podatkov (**Roadrunner**)

Rezultati so pridobljeni na podlagi treh domen:

* ***Overstock.com***
* ***Rtvslo.si***
* ***Imdb.com***

# Opis strani imdb.com

Za našo tretjo domeno smo si izbrali domeno imdb.com. Podatke pa smo pridobivali na podlagi strani, ki so kot zadnje navedene v virih.

## Prikaz podatkov na strani



RAITING

CONTENT

RUNTIME

GENRE

YEAR

TITLE

# Implementacija z regularnimi izrazi

Overstock:

* ***Title***: </tbody></table></td><td valign=\"top\">\s\*<a.\*>\s\*<b>(.\*)<\/b>
* ***ListPrice***: [\s\S|.]\*?<td align=\"left\" nowrap=\"nowrap\">\s\*<s>(.\*)<\/s>
* ***Price***: [\s\S|.]\*?<span class=\"bigred\">\s\*<b>(.\*)<\/b>
* ***Saving***: [\s\S|.]\*?<span class=\"littleorange\">([$€]\s\*[0-9\.,]+).\*
* ***SavingPercent***: \((.\*?)\)
* ***Content***: [\s\S|.]\*?<span class=\"normal\">([\s\S|.]\*?)<br>

Rtvslo:

* ***Title***: <h1>(.\*)<\/h1>\s\*<div class=\"subtitle\">
* ***SubTitle***: <div class=\"subtitle\">(.\*)<\/div>
* ***Lead***: <div class=\"author-timestamp\">\s+<strong>(.\*)<\/strong>
* ***Author***: <p class=\"lead\">(.+)<\/p>
* ***PublishedTime***: <\/strong>\|\s+(.\*?)\s\*<\/div>\s\*<div class=\"place-source\">
* ***Content***: <article(?:\s\*<p>(.\*?)<\/p>\s\*)+<\/article>(?=[^<]\*(?:<|$))

Imdb:

* ***Title***: </tbody></table></td><td valign=\"top\">\s\*<a.\*>\s\*<b>(.\*)<\/b>
* ***Year***: [\s\S|.]\*?<td align=\"left\" nowrap=\"nowrap\">\s\*<s>(.\*)<\/s>
* ***Runtime***: [\s\S|.]\*?<span class=\"bigred\">\s\*<b>(.\*)<\/b>
* ***Genre***: [\s\S|.]\*?<span class=\"littleorange\">([$€]\s\*[0-9\.,]+).
* ***Rating***: \((.\*?)\)
* ***Content***: [\s\S|.]\*?<span class=\"normal\">([\s\S|.]\*?)<br>

# Implementacija z xpath-om

Overstock:

* ***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:

* ***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***: //\*[@id="main-container"]/div[3]/div/div[2]/article/p/text()
* ***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]

Imdb:

* ***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()

# Avtomatizirana ekstrakcija (Roadrunner-like implementacija)

## Pseudo koda

FUNCTION roadrunner(html\_1, html\_2, regex)

#html\_1 in html\_2 sta objekta ki predstavljata drevesno strukturo celotnega html-ja

#Preverimo html zančke

IF html značke zadovoljijo heuristikam AND IF trenutno html drevo vsebuje besedilo:

regex-u dodamo ime značke #npr. '<html>'

IF html\_1.text = html\_2.text

#teksta se povsem ujemata

regex-u dodamo vsebino enga izmed tekstov

ELSE

#teksta se razlikujeta

regex-u dodamo niz '#text'

#poiščemo otroke html\_1 in otroke html\_2

children\_1 🡨 html\_1.children

children\_2 🡨 html\_2.children

WHILE children\_1 IS NOT EMPTY

child\_1 🡨 children\_1[0]

WHILE children\_2 IS NOT EMPTY

child\_2 🡨 children\_2[0]

IF znacki se ujemata:

trenutni\_izraz = roadrunner(child\_1, child\_2, regex)

IF zadnji izraz v regex-u PODOBEN trenutni\_izraz

zadnji izraz = '(zadnji izraz) +' #dodamo '() +'

ELSE

regex += trenutni\_izraz

ELSE

#znački se ne ujemata

preverimo ali se ujema s katerim od naslednjih otrok

IF se ujema

Ponovimo postopek na trenutnih otrocih

ELSE

#dodamo samo znacko

regex += '(<child\_2.name ...></child\_2.name>)?'

REMOVE child\_2

REMOVE child\_1

regex += '</ html\_1.name>'

RETURN regex

Heuristike, ki smo jih določili so:

* Značka ne sme biti ena izmed naslednjih: ***head***, ***script***, ***iframe***, ***footer***, ***nav***, ***style***, ***map***, ***input***. Zaradi implementacije z beautifulsoup pa smo dodali še, da ne sme biti ***none***.
* Če neka značka sploh ni vsebovala besedila (tudi v sinovih), smo jo izpustili.
* Če je celotno besedilo prvega html-ja v nekem podrevesu identično drugemu html-ju, v izraz **prepišemo kar besedilo, brez html značk**. To naredimo zato, ker obstaja velika verjetnost, da gre v tem primeru za vsebine, ki jih vsebujejo vse strani te domene (npr. glava, noga, meni, ...) in nas zato ne zanimajo preveč. S tem zmanjšamo dolžino output-a.
* Izraza sta **podobna**, če je **jaccardova razdalja manjša od 0.25**. To nam je prišlo prav pri straneh na domeni imdb.com, saj se html za posamezne filme nekoliko razlikuje in bi v tem primeru bilo potrebno zapisati ves html strani. Strmeli pa smo k temu, da uporabimo oznako za naštevanje '(...) +' saj s tem močno zmanjšamo dolžino output-a. Tako smo se odločili da, če sta si dva izraza podobna po jaccardovi razdalji, v končni output dodamo daljšega izmed njih.

## Outputs

### Overstock.com

<html><body><table> Search: All StoresHome & GardenElectronics & ComputersBooks, Movies, CDs, GamesJewelry, Watches & GiftsSports, Travel & ToysWorldstockApparel, Shoes & Access. </table><table><tbody><tr><td><table><tbody><tr><td><table><tbody><tr><td><span><b>#text</b></span></td></tr>(<tr><td><a>#text</a></td></tr>) +</tbody></table></td></tr></tbody></table><span>Stores</span><table> Apparel, Shoes & Access.Books, Movies, CDs, GamesElectronics & ComputersHome & GardenJewelry, Watches & GiftsSports, Travel & ToysWorldstock</table><span>New Stock</span><table> Ralph Lauren $29.95 Ben Sherman 53% off Pre-order Harry Potter DVD HP 2GHz System $499 New Items within 7 Days </table><span>Customer Service</span><table> Shopping Cart & Checkout Track Your Order Your Account Help & FAQ Best Price Guarantee </table><span>About Us</span><table> About Us Privacy & Security Terms & Conditions Become An Affiliate Business Purchases Have Products to Sell? Investor Relations </table></td><td><b>>>> View All<a>Jewelry, Watches & Gifts</a><a>Jewelry</a><a>#text</a></b><table><tbody><tr><td><table><tbody><tr><td><table><tbody><tr> List Sorted By: Top Sellers | Discount | Newest First | Price | Quantity | Markdowns</tr></tbody></table></td></tr></tbody></table></td></tr><tr><td><table><tbody><tr><td><table><tbody>(<tr><td> More Info... </td><td><a><b>#text</b></a><table><tbody><tr><td><table><tbody><tr><td>List Price:</td><td><s>#text</s></td></tr><tr><td>Price:</td><td><span><b>#text</b></span></td></tr><tr><td>You Save:</td><td><span>#text</span></td></tr></tbody></table></td><td><span>#text<a>Click here to purchase.</a></span></td></tr></tbody></table></td></tr>) +</tbody></table></td></tr></tbody></table></td></tr><tr><td><table><tbody><tr><td><table><tbody><tr> List Sorted By: Top Sellers | Discount | Newest First | Price | Quantity | Markdowns</tr></tbody></table></td></tr></tbody></table></td></tr></tbody></table></td></tr><tr> © 2003Overstock.com \* $2.95 flat rate shipping to the lower 48 states only. Some items excluded due to size and/or weight. </tr></tbody></table></body></html>

### Komentar:

Uporaben del tega izhoda je označen z rdečo barvo. Ta del predstavlja posamezne izdelke na spletni strani, kjer se lepo opazi, da se spreminjo ključni podatki.

### Rtvslo.si

<html><body><div> RTV SLO Radio Televizija O RTV 4D Spored V živo Arhiv Prijava Uporabniško ime: Geslo: Prijavi Registracija Pozabljeno geslo? Temni način BETA Odjava Uporabniški račun Temni način BETA </div><div> RTVSLO.si Slovenija Svet Šport Kultura Življenjski slog Svet zabave Iskanje Kazalo Predlogi Odbojka Hokej Tenis Rezultati iskanja Zaradi testiranja je dodanih umetnih 2 sekunde delaya. </div>(<div ... />)?<div> Slovenija Gospodarstvo Lokalne novice Črna kronika Zdravje Okolje Znanost in tehnologija Slovenci v sosednjih državah Svet Evropska unija Evropa S. in J. Amerika Bližnji vzhod Afrika Azija z Oceanijo Šport Nogomet Košarka Rokomet Odbojka Hokej Tenis Atletika Zimski športi Motošporti F1 Kolesarstvo Preostali športi Rekreacija Športni SOS Kultura Film in TV Glasba Knjige Oder Vizualna umetnost Arhitektura in oblikovanje Dediščina Recenzije Drugo Jezikovni spletovalec 1. svetovna vojna Življenjski slog Ture avanture Kulinarika Lepota bivanja Avtomobilnost Moda 196x ljubezen Svet zabave Evrovizija Glasba Film in TV Svet znanih Družabno Zanimivosti Posebna izdaja MMC Podrobno Klepet Podkast Številke 1968 Kolumne MMC Teden MMC Analiza Praznično leto RTV Slovenija Info Vreme Snežne razmere Stanje na cestah TV Slovenija Info Kultura Filmi/serije Razvedrilo Otroci Izobraževanje Dokumentarci Duhovnost Radio Slovenija Prvi Val 202 ARS Radio Koper Radio Capodistria Radio Maribor Radio Si MMR RTV 4D V živo Arhiv Spored Oddaje Tematski portali Otroški portal Moja generacija Moj splet Teletext O RTV Slovenija Kdo smo RTV-prispevek Za medije Kontakti in informacije </div>(<div></div>) +<div><div>(<div ... />)?</div></div><div><div><div><div><h5>#text</h5><button> × </button></div><div>(<h2 ... />)?(<a ... />)?(<a ... />)?(<div ... />)?</div></div></div></div>(<link ... />)?(<img ... />)?(<img ... />)?</body></html>

### Komentar:

Algoritem ima največje težave pri straneh iz domene Rtvslo.si, saj se značke med sabo močno razlikujejo, in struktura ni konsistentna. Tako iz izhoda ne pridobimo zelo uporabnega regularnega izraza.

### Imdb.com

<html><body><section>MenuMoviesRelease CalendarTop 250 MoviesMost Popular MoviesBrowse Movies by GenreTop Box OfficeShowtimes & TicketsMovie NewsIndia Movie SpotlightTV ShowsWhat's on TV & StreamingTop 250 TV ShowsMost Popular TV ShowsBrowse TV Shows by GenreTV NewsIndia TV SpotlightWatchWhat to WatchLatest TrailersIMDb OriginalsIMDb PicksIMDb PodcastsAwards & EventsOscarsSXSWStar Wars CelebrationCannesSTARmeter AwardsAwards CentralFestival CentralAll EventsCelebsBorn TodayMost Popular CelebsMost Popular CelebsCelebrity NewsCommunityHelp CenterContributor ZonePollsFor Industry ProfessionalsAllAllTitlesTV EpisodesCelebsCompaniesKeywordsAdvanced SearchWatchlistSign InENFully supportedEnglish (United States)Partially supportedFranÃ§ais (Canada)FranÃ§ais (France)Deutsch (Deutschland)à¤¹à¤¿à¤‚à¤¦à¥€ (à¤­à¤¾à¤°à¤¤)Italiano (Italia)PortuguÃªs (Brasil)EspaÃ±ol (EspaÃ±a)EspaÃ±ol (MÃ©xico)</section><div><div><div> </div><div><div><div><div><h1>#text</h1><div><div> View Mode: Compact | Detailed </div><div><span>#text</span><span>|</span><a>Next »</a></div></div><div> Sort by: Popularity | A-Z | User Rating▼ | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating </div><div><div>(<div><div><h3><span>#text</span><a>#text</a><span>#text</span></h3><p><span>#text</span><span>|</span><span>#text</span><span>|</span><span>#text</span></p><div><div><strong>#text</strong></div><div><span> Rate this </span><div><div><span> 1 2 3 4 5 6 7 8 9 10 </span><span><span>#text</span>(<span>10</span>) +</span><span>X</span></div></div></div><div>Metascore<span>#text</span></div></div><p>#text</p><p>Director:Stars:,,,(<a>#text</a>) +<span>|</span>(<a>#text</a>) +</p><p><span>Votes:</span><span>#text</span><span>|</span><span>Gross:</span><span>#text</span></p></div></div>) +<div><div><h3><span>#text</span><a>#text</a><span>#text</span></h3><p><span>#text</span><span>|</span><span>#text</span></p><div><div><strong>#text</strong></div><div><span> Rate this </span><div><div><span> 1 2 3 4 5 6 7 8 9 10 </span><span><span>#text</span>(<span>10</span>) +</span><span>X</span></div></div></div></div><p>#text</p><p>#text<a>#text</a>(<a ... />)?<span>|</span>(<a ... />)?(<a ... />)?(<a ... />)?(<a ... />)?</p><p><span>Votes:</span><span>#text</span></p></div></div></div></div><div><span>#text</span><span>|</span><a>Next »</a></div></div></div><div> </div></div></div><div> Clear your history Recently Viewed </div></div></div><section>Get the IMDb AppSign in for more accessSign in for more accessGet the IMDb AppHelpSite IndexIMDbProBox Office MojoIMDb DeveloperPress RoomAdvertisingJobsConditions of UsePrivacy PolicyYour Ads Privacy ChoicesIMDb, an Amazon companyÂ© 1990-2023 by IMDb.com, Inc.</section><div> </div></body></html>

### Komentar:

Tudi v tem primeru je z rdečo barvo označen del, ki predstavlja določen film na spletni strani. Sprminjajoči se podatki pa so označeni s '#text'.

# Viri

* Parsel [Online]. Dosegljivo: <https://parsel.readthedocs.io/en/latest/usage.html> (zadnji dostop 5.5.2023)
* Beautiful Soup Documentation [Online]. Dosegljivo: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (zadnji dostop 5.5.2023)
* Python RegEx, W3Schools [Online]. Dosegljivo: <https://www.w3schools.com/python/python_regex.asp> (zadnji dostop 5.5.2023)
* V. Crescenzi, G. Mecca, P. Meraildo. ROADRUNNER: Towards Automatic Data Extraction from Large Web Sites [Online]. Dosegljivo: <https://vldb.org/conf/2001/P109.pdf> (zadnji dostop 5.5.2023)
* Feature Film, Rating Count at least 25,000, Action (Sorted by IMDb Raiting Descending) [Online]. Dosegljivo:
* <https://www.imdb.com/search/title/?genres=action&sort=user_rating,desc&title_type=feature&num_votes=25000,&pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=94365f40-17a1-4450-9ea8-01159990ef7f&pf_rd_r=VTPTT29EC5M36Q86ZTPK&pf_rd_s=right-6&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_gnr_1> (zadnji dostop 5.5.2023)
* Feature Film, Rating Count at least 25,000, Action (Sorted by IMDb Raiting Descending) [Online]. Dosegljivo:
* <https://www.imdb.com/search/title/?genres=sport&sort=user_rating,desc&title_type=feature&num_votes=25000,&pf_rd_m=A2FGELUUNOQJNL&pf_rd_p=94365f40-17a1-4450-9ea8-01159990ef7f&pf_rd_r=VTPTT29EC5M36Q86ZTPK&pf_rd_s=right-6&pf_rd_t=15506&pf_rd_i=top&ref_=chttp_gnr_18> (zadnji dostop 5.5.2023)