Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Mihael Švigelj, Lovro Vražič in Luka Zakšek

Extraction

Domača naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTOR: prof. dr. Marko Bajec

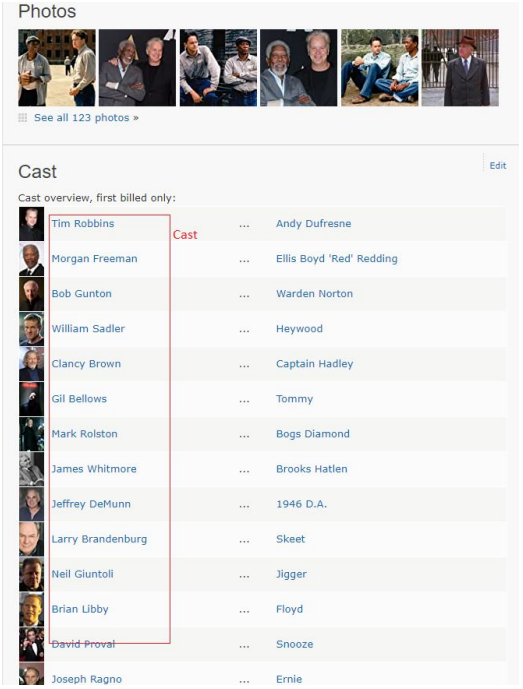
V nalogi smo po danih navodilih v Pythonu razvili dve različni metodi za ekstrakcijo podatkov s spletnih strani overstock.com, rtvslo.si in imdb.com

1. Uvod

Nalogo smo implementirali v jeziku Python. Pri implementaciji smo uporabili knjižnice: re, json, html2text in lxml.

2. Izbira tretje spletne strani

Kot tretjo spletno stran iz katere smo ekstrahirali podatke, smo si izbrali imdb.com. Specifično smo si izbrali strani filmov Kaznilnica odrešitve (*The Shawshank Redemption*) in Boter (The *Godfather*). Spodaj sta podani sliki, ki ponazarjata katere podatke smo ekstrahirali.



3. Struktura programa

Program se zažene s klicom *“python run-extraction.py A/B/C”,* kjer argument A/B/C definira način ekstrakcije. Preberejo se prenesene izbrane spletne strani nato pa se glede na izbrano metodo ekstrakcije kliče eno izmed treh ostalih pripravljenih skript, ki ekstrakcijo tudi izvedejo.
Po končani ekstrakciji se v terminalu izpišejo zbrani podatki.

3.1 Regex

Za vsakega izmed izbranih podatkov za ekstrakcijo je napisan regularni izraz, ki na podani spletni strani najde in vrne želeno vsebino. Podatki so potem z nekaj malega prečiščevanja vstavljeni v listo slovarjev, katera je kasneje izpisana v konzolo v json formatu.

Uporabljeni regularni izrazi:

Overstock:

Naslovi: `r"<td><a href=\"http://www\.overstock\.com/cgi-bin/d2\.cgi\?PAGE=PROFRAME[\w\W]*?\"><b>(.*?)</b></a><br>"`
Vsebina: `r"<span class=\"normal\">(.*?)<br>"`
Navedena cena: `r"<b>List Price:[\w\W]*?<s>(.*?)</s>"`
Cena: `r"<b>Price:[\w\W]*?<b>(.*?)</b>"`
Prihranek: `r"<b>You Save:[\w\W]*?class=\"littleorange\"\>(.*?)\)</span>"`
(pri prihranku se ekstraktira tako absolutna vrednost, kot tudi odstotki, vrednosti se kasneje loči)

Rtvslo:

Avtor: `r"<div class=\"author-name\">(.*)</div>"`
Čas objave: `r"<div class=\"publish-meta\">\n\t\t(.*)<br>"`
Naslov: `r"<h1>(.*)</h1>"`
Podnaslov: `r"<div class=\"subtitle\">(.*)</div>"`
Osnutek: `r"<p class=\"lead\">(.*)</p>"`
Vsebina: `r"<div class=\"article-body\">(.*?)<div class=\"gallery\">"`

Imdb:

Naslov: `r"<h1 class=\"\">(.*?)\ \;<span"`
Originalni naslov: `r"<div class=\"originalTitle\">(.*?)<span class=\"description\">"`

Leto: `r"<span id=\"titleYear\">\(<a href=\"[\w\W]*?\">(.*?)</a>\)</span>"`
Dolžina: `r"<time datetime=\"[\W\w]*?\">\s*(.*?)\s*</time>"`
Ocena: `r"<span itemprop=\"ratingValue\">(.*?)</span>"`
Opis: `r"<div class=\"summary_text\">\s*(.*?)\s*</div>"`
Režiser: `r"<h4 class=\"inline\">Director:</h4>\s<a href=\"[\w\W]*?\">(.*?)</a>"`
Igralci: `r"<td>\s<a href=\"https://www\.imdb\.com/name/[\w\W]*?\">\s(.*?)\s</a>"`

3.2 Xpath

Skripta deluje zelo podobno kot skripta za ekstrakcijo z uporabo regularnih izrazov, le da se pri ekstrakciji namesto regularnih izrazov uporablja html struktura spletne strani. Ponovno so za vse izbrane podatke strukture trdo vpisane.

Uporabljena Xpath drevesa:

Overstock:

Naslov:
`'/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()'`
Vsebina:
`'/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()'`
Navedena cena:
`'/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()'`
Cena:
`'/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()'`
Prihranek:
`'/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()'`

Rtvslo:

Avtor:
`'//*[@id="main-container"]/div[3]/div/div[1]/div[1]/div/text()'`

Čas objave:
`'//*[@id="main-container"]/div[3]/div/div[1]/div[2]/text()[1]'`
Naslov: `'//*[@id="main-container"]/div[3]/div/header/h1/text()'`
Podnaslov:
`'//*[@id="main-container"]/div[3]/div/header/div[2]/text()'`
Osnutek: `'//*[@id="main-container"]/div[3]/div/header/p/text()'`
Vsebina:
`'//*[@id="main-container"]/div[3]/div/div[2]/div/figure/figcaption/text()'`
in
`'//article[@class="article"]/p/text()'`

Imdb:

Naslov:
`'//*[@id="title-overview-widget"]/div[1]/div[2]/div/div[2]/div[2]/h1/text()'`
Originalni naslov:
`'//*[@id="title-overview-widget"]/div[1]/div[2]/div/div[2]/div[2]/div[1]/text()'`
Leto: `'//*[@id="titleYear"]/a/text()'`
Dolžina:
`'//*[@id="title-overview-widget"]/div[1]/div[2]/div/div[2]/div[2]/div[2]/time/text()'`
Ocena:
`'//*[@id="title-overview-widget"]/div[1]/div[2]/div/div[1]/div[1]/div[1]/strong/span/text()'`
Opis:
`'//*[@id="title-overview-widget"]/div[2]/div[1]/div[1]/text()'`
Režiser:
`'//*[@id="title-overview-widget"]/div[2]/div[1]/div[2]/a/text()'`
Igralci: `'//*[@id="titleCast"]/table/tbody/tr/td[2]/a/text()'`

4. Rezultati ekstrakcije

Rezultati ekstrakcije za oba implementirana načina ekstrakcije so podani v datoteki results.md znotraj folderja implementation-extraction. Dobljeni rezultati so v obeh implementiranih metodah enaki in taki, kot pričakovano.

5. Zaključek

Težav z implementacijo Regex ali Xpath načinov ekstrakcije nismo imeli. Težaven pa je bil road-runner pristop, katerega zaradi pomanjkanja časa nismo implementirali, saj se nam je že pri prebiranju članka, ki ga opiše zdelo, da bi nam težko uspelo algoritem implementirati in razhroščiti.