Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Aljaž Srša 63120233 in Gregor Sušnik 63120102

Zajemanje vsebine spletnih strani

Poročilo druge seminarske naloge pri predmetu Iskanje in ekstrakcija podatkov s spleta

Asistent: asist. prof. dr. Slavko Žitnik

**Povzetek**

V seminarski nalogi implementirava zajemanje podatkov na tri različne načine . Pri vsakem načinu opiševa uporabljene izraze, ki zajamejo zahtevano vsebino na spletni strani. Implementirane metode preizkusiva na določenih parih spletnih strani – overstock, rtvslo.si in newegg.com. Dobljene rezultate primerjava v poglavju rezultati.

1. **Uvod**

Ročno zajemanje podatkov iz spletnih strani je precej zahtevno in ne praktično. Še posebej, če imamo opravka z veliko količino podatkov. Zajemanje večje količine podatkov poskušamo avtomatizirati, saj s tem prihranimo veliko časa in dela, ki bi ga opravljali ročno.

V drugi seminarski nalogi smo se lotili procesa zajemanja podatkov. Ideja je bila implementirati na tri različne načine za pridobivanje podatkov, in sicer: uporaba regularnih izrazov, uporaba XPath izrazov in uporaba RoadRunner pristopa. Vsak način smo na koncu preizkusili na treh različnih tipih strani. Vsak tip strani je vseboval dve strani. Nad vsemi tremi pari stranmi – skupaj torej šest strani, so se preizkusili vsi trije načini za pridobivanje podatkov in med posameznima stranema v paru se je primerjalo pridobljene podatke.

1. **Metode pridobivanja podatkov**

Seminarsko nalogo sva izdelala v jeziku Python. Program se začne izvajati v glavni datoteki main.py, v kateri se prebere HTML datoteka s pomočjo funkcije read(). Prebrana datoteka se s pomočjo knjižnice BeautifulSoup olepšala v smislu optimizacije, zaključka značk, itp. Tako pripravljen HTML je vhod za vse tri funkcije pridobivanja podatkov. Vsak način je implementiran kot funkcija v ločeni datoteki. Kateri tip spletne strani pa ureja spremenljivka pageType. Tip je lahko nastavljen na 0 (Overstock), 1 (rtvslo.si) ali 2 (newegg.com)

* 1. **Uporaba regularnih izrazov**

Funkcija sprejme obdelan HTML string in tip spletne strani. Na podlagi podatka o tipu spletne strani se izvede koda. Za zajem podatkov poskrbijo spodnji regularni izrazi. Pridobljeni podatki se shranijo v slovar, ta pa se s pomočjo funkcije json\_dump() izvozi v berljivo obliko.

**Overstock**

*regex = r'<b>([0-9].+)<\/b>.\*\n.\*\n.\*<s>([\$0-9,.]+).\*\n.\*<b>([\$0-9.]+).\*\n.\*>([\$0-9.,]+)\s.([0-9]+.).\*\n.\*\n.\*<span class=\"normal\">(.\*|.\*\n.\*\n.\*|.\*\n.\*\n.\*\n.\*)<br/><a'*

Razlaga regularnega izraza:

* <b>([0-9].+)<\/b> poišče naziv izdelka, ki pa se vedno prične z neko številko
* .\*\n.\*\n.\* Izpusti nekaj vsebine, ki je med nazivom izdelka in cenami
* ([\$0-9,.]+) Zajame vsebino 'listPrice'
* .\*\n.\* Izpusti nekaj vsebine
* <b>([\$0-9.]+) Zajame vsebino 'price'
* .\*\n.\* Izpusti nekaj vsebine
* >([\$0-9.,]+)\s.([0-9]+.).\*\n.\*\n.\* Zajame 'saving' in 'savingPercent'
* <span class=\"normal\">(.\*|.\*\n.\*\n.\*|.\*\n.\*\n.\*\n.\*)<br/><a' Zajame vsebino opisa izdela. Pri čemer ima opis lahko več vrstic opisa.

**Rtvslo.si**

*regex = r'<h1>(.\*)<\/h1>[\s\S]+<div class=\"subtitle\">(.\*)<\/div>[\s\S]+<p class=\"lead\">(.\*)<\/p>[\s\S]+<div class=\"author-name\">(.\*)<\/div>[\s\S]+\"publish-meta\">[\n\W]+(.\*)<br\/>[\s\S]+<\/div>[\n]\*<\/figure>[\n]\*<p([\s\S]\*.\*)<div class=\"gallery\">'*

Razlaga regularnega izraza:

* <h1>(.\*)</h1> poišče naslov članka
* >[\s\S]+ izpusti nekaj nezanimive vsebine
* <div class=\"subtitle\">(.\*)</div> poišče podnaslov
* [\s\S]+ ponovno izpusti nekaj vsebine
* <p class=\"lead\">(.\*)</p> poišče nagovorno besedilo
* [\s\S]+ izpusti nekaj vsebine
* <div class=\"author-name\">(.\*)</div> poišče avtorja članka
* [\s\S]+ izpusti nekaj vsebine
* \"publish-meta\">[\n\W]+(.\*)<br/>' poišče datum objave članka
* *<\/div>[\n]\*<\/figure>[\n]\*<p([\s\S]\*.\*)<div class=\"gallery\">* zajame celotno vsebino članka

Vsebina članka vsebuje poleg besedila tudi preostale značke in znake. Te s pomočjo funkcije lxml.html.fromstring(content).text\_content() odstraniva oz. funkcija poskrbi, da se iz podanega konteksta izlušči le tekst.

**newegg.com**

regex: < to be added>

* 1. **Uporaba XPath**

Funkcija sprejme obdelan HTML string in tip spletne strani. Na podlagi podatka o tipu spletne strani se izvede koda. Za zajem podatkov poskrbijo spodnji XPath izrazi. Pridobljeni podatki se shranijo v slovar, ta pa se s pomočjo funkcije json\_dump() izvozi v berljivo obliko.

**Overstock**

Najprej poiščemo del v HTML, kjer se nahajajo željeni podatki. Iščejo se vsi takšni <TR> nodi, ki imajo atribut bgcolor nastavljen na določeno vrednost pri tem pa mora omenjen node imeti še točno dva otroka tipa <TD>

*objects = tree.xpath('//tbody/tr[(contains(@bgcolor, "#ffffff") or contains(@bgcolor, "#dddddd")) and count(td[@valign="top"]) = 2]*

Nato se nad trenutno lokacijo pridobivajo podatki o vsebini, ki nas zanimakar v for zanki.

for obj in objects:

*title = obj.xpath('string(a/b/text())')*

*listPrice = obj.xpath('string(table/tbody/tr/td[1]/table/tbody/tr[1]/td[2]/s/text())')*

*price = obj.xpath('string(table/tbody/tr/td[1]/table/tbody/tr[2]/td[2]/span/b/text())')*

*saving = obj.xpath('substring-before(table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text(), " ")')*

*savingPercent = obj.xpath('substring(substring-after(table/tbody/tr/td[1]/table/tbody/tr[3]/td[2]/span/text(), " "),2 ,3)')*

*content = obj.xpath('string(table/tbody/tr/td[2]/span/text())')*

Tudi tukaj se izvede json dump, kise vrne v main kot rezultat funkcije.

**Rtvslo.si**

Najprej se poišče lokacijo v HTML kodi, kjer se nahajajo željeni podatki. Nato pa se od omenjene lokacije na podlagi relativne poti zajemajo zahtevani podatki.

rootObject = tree.xpath('//div[contains(@class, "news-container")]/div')[0]

author = rootObject.xpath('string(div[@class="article-meta"]/div[@class="author"]/div/text())')

publishTime = rootObject.xpath('string(div[@class="article-meta"]/div[@class="publish-

meta"]/text())')

title = rootObject.xpath('string(header/h1/text())')

subTitle = rootObject.xpath('string(header/div[@class="subtitle"]/text())')

lead = rootObject.xpath('string(header/p/text())')

Vsebina članka pa se zajeme s spodnjim XPath izrazom:

contentList = rootObject.xpath('div[@class="article-body"]/article[@class="article"]//p | '

'div[@class="article-body"]/article[@class="article"]//strong')

content = ""

for p in contentList:

if p.text is not None:

content += '\n' + p.text

if p.tail is not None:

content += '\n' + p.tail

**newegg.com**

objects = tree.xpath('//div[@class="item-container"]')

for obj in objects:

dataInfo = obj.xpath('div[@class="item-info"]')

item = {}

title = dataInfo[0].xpath('string(a[@class="item-title"]/text())')

rating = dataInfo[0].xpath('string(div[@class="item-branding"]/a[@class="item-rating"]/@title)')

promo = dataInfo[0].xpath('string(p[@class="item-promo"]/text())')

priceCurrent = dataInfo[0].xpath('string(div[@class="item-action"]/ul/li[@class="price-current"])')

offers = dataInfo[0].xpath('string(div[@class="item-action"]/ul/li[@class="price-current"]/a/text())')

priceWas = dataInfo[0].xpath('string(div[@class="item-action"]/ul/li[@class="price-was"]/text())')

priceSave = dataInfo[0].xpath('string(div[@class="item-action"]/ul/li[@class="price-save"]/span[@class="price-save-percent"]/text())')

* 1. **Uporaba RoadRunner algoritma**

1. **Rezultati**

**Overstock**

**Stran A**

Pri zajemanju podatkov z uporabo regularnih izrazov in XPath izrazov je končni rezultat enak. Obe metodi uspešno zajemata zahtevane podatke. se doda izpis v poročilo?

**Stran B**

Pri zajemanju podatkov z uporabo regularnih izrazov in XPath izrazov je končni rezultat enak. Obe metodi uspešno zajemata zahtevane podatke.

**RTVSLO.si**

**Stran A**

Pri zajemanju podatkov z uporabo regularnih izrazov in XPath izrazov je končni rezultat enak. Obe metodi uspešno zajemata zahtevane podatke.

**Stran B**

Pri zajemanju podatkov z uporabo regularnih izrazov in XPath izrazov je končni rezultat enak. Obe metodi uspešno zajemata zahtevane podatke.

**<CUSTOM\_SITE>**

**Stran A**

**<**INSERT SOME TEXT HERE>

**Stran B**

**<**INSERT SOME TEXT HERE>

**Literatura**