Poročilo prve naloge pri predmetu Iskanje in ekstracija podatkov s spleta

Sara Bertoncelj Čadež, Gaja Ana Boc, Aljaž Gobec

April 3, 2020

Povzetek

V poročilu je predstavljena implementacija spletnega pajka v okviru prve seminarske naloge pri predemtu Iskanje in ekstrakcija podatkov s spleta.

1 Uvod

Namen naloge je bila imelementacija spletnega pajka. Spletni pajek se premika po spletnih straneh domene *gov.si. Pajek s pomočjo večnitnega delovanja vzporedno pridobiva spletne strani in njhove informacije zapiše v podatkovno bazo. Za strategijo premikanja uporablja iskanje v širino.

2 Implementacija

2.1 Zajem strani

Za zajemanje strani smo uporabljali knjižnjico Selenium, ki avtomatsko zažene JavaScript kodo. Knjižnjica s pomočjo ChromeDriverja simulira Google Chrome brskalnik. Čas med zahtevami za prenos strani na določen strežnik smo omejili na 5 sekund.

2.1.1 Upoštevanje robots.txt in sitemap.xml

Za vsako obiskano domeno smo upoštevali datoteko robots.txt, kjer so zapisane strani, ki jih pajek ne sme obiskati. Pri upoštevanju smo uporabli knjižnjico urllib.robotparser. Če domena ni imela datoteke robots.txt, smo obiskali vse njene strani ter podstrani. V bazo smo poleg robots.txt zapisali tudi vsebino datoteke sitemap.xml.

2.1.2 Upoštevanje HTTP glav

V HTTP glavi strani smo vsem naslovom povezav z relativno potjo npr. (/podrocja) poiskali atribut $< basehref >$ in naslov spremenili v primeren naslov z absolutnjo potjo. Če ni bilo omenjenega atributa, smo naslovu dodali naslov domene. Poleg tega smo v HTTP glavi razbrali ustrezne statusne kode, ki smo jih zapisali v bazo in povezavo primerno obravnavali, na primer: če je bila povezava na stran uspešna (200), smo stran ustrezno razčlenili.

2.2 Razčlenjevanje strani

2.2.1 Ekstrakcija povezav in vsebine

S pomočjo knjižnjice Beautiful Soap smo na straneh izluščili vse povezave, ki smo jih kasneje dodali v frontier. Poleg HTML vsebine, smo na vsaki stran izluščili naslov strani, tip strani, tip statusne kode, čas dostopa in povezave med posameznimi stranmi. Prav tako smo poiskali vse slike in izluščili ID strani na kateri je slika, ime slike, tip slike in čas dostopa do slike.

2.2.2 Normalizacija naslovov

Po tem, ko smo vse naslove z relativno potjo spremenili v naslove z absolutnjo potjo, smo naslove normalizirani. Odstranili smo jim nepotrebne podatke, kot so številka vrat, številke morebitnega fragmenta, kodirani znaki, nepotrebne končnice, privzeto ime datoteke in spremenili velike črke v male ter dodali morebitne potrebne znake (npr. "/"). Rezultat je bila ustrezna povezava, ki jo lahko dodamo v frontier.

2.3 Frontier

Za stretegijo premikanja pajka smo uporabili premikanje v širino, ki uporablja FIFO vrsto, kar pomeni, da tiste povezave, ki jih prvo dodanmo, tudi prvo obiščemo. Implementirali smo jo s pomočjo modula queue, ki je omogočal učinkovito izmenjavo podatkov med več spletnimi delavci.

2.4 Iskanje duplikatov

V seznam zgodovine smo hranili vse povezave, ki jih je pajek že obiskal. V frontier nismo dodajali povezav, ki smo jih že obiskali oz. shranili v frontier. Poiskali smo vse vse kanonične povezave in morebitna podvajanja. Poleg tega smo vsako stran s pomočjoo zgoščevalne funkcije primerjali z že obiskanimi stranmi. V primeru, da se je "hash" ujemal, smo stran označili kot duplikat.

3 Statistika in vizualizacija podatkovne baze

V tabeli 1 so predstavljene glavne značilnosti podatkov v podatkovni bazi.

število/domena	gov.si	evem.gov.si	e-uprava.gov.si	e-prostor.gov.si
vseh strani	9797	9090	264	117
duplikatov	1076	559	0	2
HTML strani	8377	8392	115	87
datotek	14302	4817	1	300
slik	140847	67565	115	453

število	spletišč	vseh strani	HTML strani	duplikatov	datotek	slik
*gov.si	71	52191	33058	2874	31282	449534

Table 1: Statistika

Na sliki je izrisana vizualizacija povezav. Za preglednejšo vizualiazacijo smo izbrali le tiste strani, ki imajo vsaj 5 izhodnih povezav. Na abscistni osi so posamezna spletišča oz. domene. Krogi ponazarjajo posamezne strani, velikost kroga pa ponazarja število izhodnih povezav iz te strani. Iz vizualizacije lahko primerjamo število izhodnih povezav imajo strani posameznega spletišča.

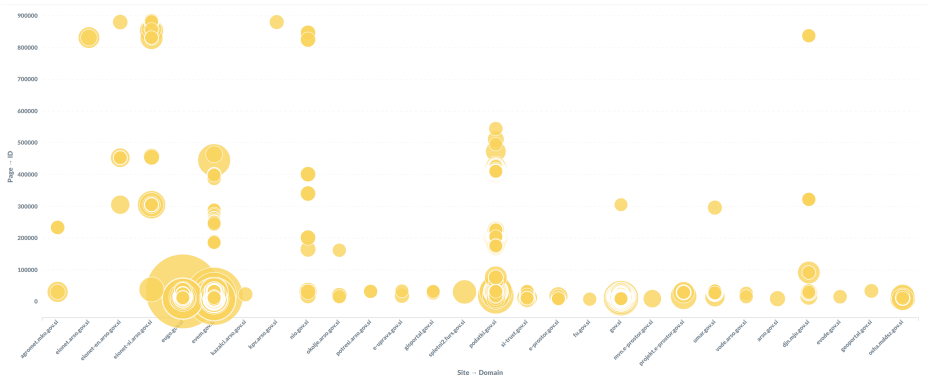

Figure 1: Vizualizacija

4 Zaključek in izboljšave

Pri implementaciji smo se soočali z nekaj izzivi. Poleg manjših hroščev, ki smo jih uspešno rešili, je glavni izziv bil omejitev virov. Spletenga pajka smo bili primorani zaganjati na osebnem prenosnem računalniku, ki ga, še posebej v času izolacije, potrebujemo za ostale obveznosti, zato smo morali omejiti število spletnih delavcev in posledično podcenili hitrost delovanja spletenga pajka. Daljše delovanje spletnega pajka bi se odražalo v višjem številu obiskanih straneh in višjem številu pridobljenih podatkov. Kljub temu, smo se spoznali z delovanjem in uspešno implementirali preprostega spletnega pajka.