Seminarska naloga 1: Spletni pajek

Erik Rakušček, Jan Šmid, Qichao Chen
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

April 3, 2020

1 Uvod

V tem poročilu opišemo spletnega pajka, ki smo ga implementirali v sklopu predmeta Iskanje in ekstrakcija podatkov s spleta [1]. S pajkom smo preiskali strani na domeni gov.si [2] in njenih podomenah ter vizualizirali pridobljene podatke.

2 Implementacija

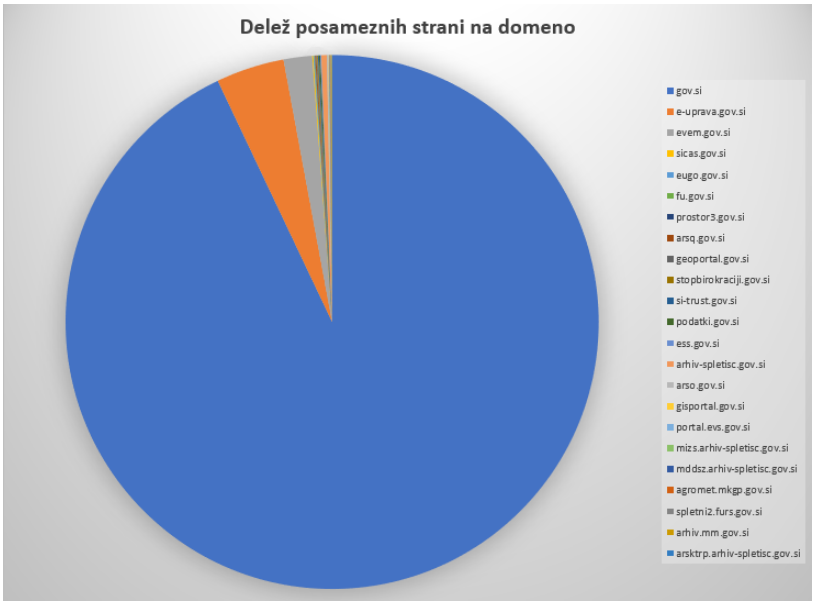
Spletnega pajka smo implementirali z uporabo Pythona. Uporabili smo več knjižnic, med drugimi tudi selenium, urllib in beautifulsoup. S pomočjo primera s predavanj smo implementirali več nitno delovanje pajka, tako da uporabnik ob zagonu navede število niti, ki jih želi dodeliti pajku. Vsaka nit iz frontier-ja vzame en url naslov in ga obdela.

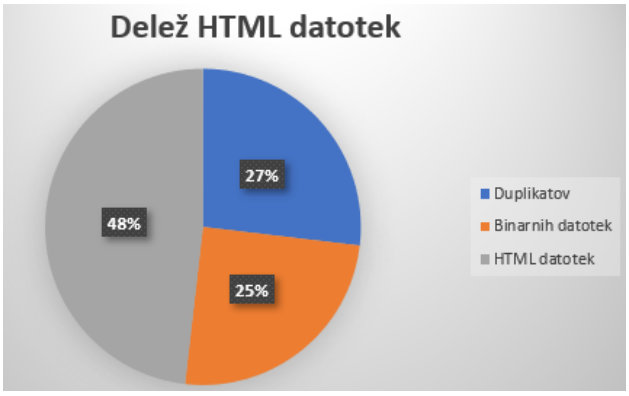
Frontier je implementiran kot seznam kateremu na konec dodajamo nove url-je, beremo oziroma odstranjujemo pa jih od začetka (FIFO vrsta). V glavni metodi imenovani *crawler* najprej preverimo ali je od zadnjega klica na isti IP naslov že minilo dovolj časa. Če da, potem najprej obdelamo *header*, nato izračunamo hash celotne strani, za ugotavljanje duplikatov (v tabelo *crawldb.page* smo tudi dodali polje *hash*). Za hitro preverjanje duplikatov si tudi hranimo seznam *history*. V tabelo *crawldb.link* dodamo vse duplikate in url-je strani katerim so identični.

Z uporabo *xpath* poiščemo vse slike in nove url-je na strani ter slike dodamo v bazo, url-je pa, v primeru da so veljavni, v frontier.

3 Statistika

	Število
Strani	60865
Domen	126
Duplikatov	16362
Slik	68832
Binarnih datotek	15177





4 Vizualizacija

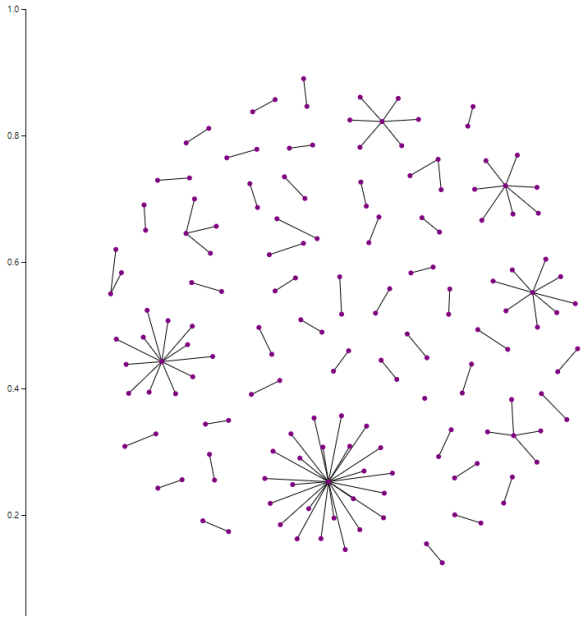


Figure 1: Slika prikazuje koliko strani kaže na isto vsebino

5 Zaključek

Naša implementacija spletnega pajka se je izkazala za učinkovito pri pridobivanju podatkov s spletnih strani. Kljub temu bi pajka lahko še izboljšali tako, da bi bolje izkoristili paralelizacijo in bi sočasno obiskali url-je v frontierju ki so na različnih IP naslovih (brez iskanja v širino). Prav tako bi lahko v bazi shranjevali še binarne datoteke (zip, pdf...) za kasnejšo analizo.

References

[1] Github repozitorij. Dosegljivo: https://github.com/van123helsing/pachong-11. [Dostopano: 3. 4. 2020].

[2] Republika Slovenija gov.si. Dosegljivo: https://www.gov.si/. [Dostopano: 29. 3. 2020].