Urban Poljšak, Matjaž Bevc

Web Crawler

Seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

1. Uvod

Naloga opisuje zgradbo in implementacijo spletnega pajka, ki je uporabljen za prebiranje vladnih strani. Najprej je opisana arhitektura programa, nato pa delovanje posameznih komponent. Zatem je nanizana statistika rezultatov skupaj z vizualizacijo, na koncu pa še težave pri razvoju ter zaključek.

2. Struktura programa

V osnovi je program zgrajen iz dveh delov; strežnika in odjemalca. Strežnik pošilja strani odjemalcem in shranjuje prejete podatke v bazo. Celoten projekt je spisan v jeziku Python.

2.1. Odjemalec

Odjemalec je ravno tako zgrajen iz dveh delov – pajka, ki razčlenjuje strani, ter krmilnika, ki opravlja komunikacijo med pajkom in strežnikom. Pajek teče na lokalnem Flask strežniku, krmilnik pa je skripta, ki na isti napravi pošilja ustrezne HTTP zahteve tako pajku, kot strežniku (frontier). Nastavljen je tako, da čas med dvema zahtevkoma posamezni domeni ni manjši od časa, definiranega v robots.txt datoteki oziroma 5 sekund.

2.2. Strežnik

Strežnik (frontier) skrbi za določanje naslednjih strani, ki bodo poslane v razčlenjevanje ter shranjevanje pridobljenih podatkov v bazo. Odjemalci za dostop do strežnika potrebujejo avtorizacijo, prav tako frontier do povezave z podatkovno bazo.

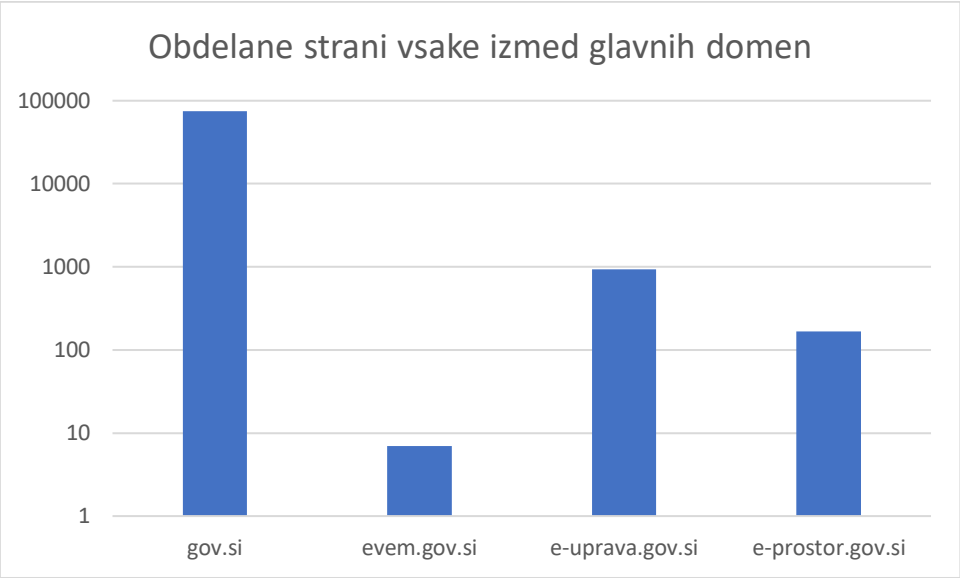
Na zahtevo pošlje po en url pajku – prvi url tipa FRONTIER. Poleg tega pošlje tudi informacijo o tem, ali je bila domena te strani že obiskana, da pajek ne izgublja časa z ponovnim pridobivanjem teh podatkov.

3. Statistika

Število obdelanih strani vsake izmed glavnih domen.

	gov.si	evem.gov.si	e-uprava.gov.si	e-prostor.gov.si
vse	74573	7	930	166
HTML	69437	7	870	158
DUPLICATE	5136	0	60	8
BINARY	0	0	0	0

4.Vizualizacija



5. Težave

Pri prvih poskusih zagona smo naleteli na mnogo majhnih težav, npr. nedostopnost elementov, ki jih je nabral Selenium, branje JSON objektov, podvajanje branja domenskih podatkov ter vračanja raznoraznih napak. Najprej je tudi samo strganje strani potekalo počasi, a smo zadevo pohitrili. Za največjo težavo se je pa izkazalo pomanjkanje časa.

Odpravljali smo jih s vztrajnim popravljanjem, ponovnim poizkušanjem in pozornim branjem sporočil napak.

6. Zaključek

Sam razvoj pajka se je izkazal za dolgotrajnejše delo, kot je bilo sprva predpostavljeno. Na koncu pa je le opravil svoje delo, saj je uspel razčleniti solidno število strani.