IEPS PA1 skupina G

NAKONC BO TREBA PRETVORT V PDF, AMPAK JE TKO LAŽI UREJAT

Za implementacijo našega spletnega pajka smo izbrali programski jezik Python. Na začetku nastavimo lokacijo gonilnika geckodriver.exe, časovno omejitev 5 sekund, poljubno število delavcev in naslove predpisanih začetnih spletnih strani. Nastavimo, da brskalnik teče v načinu brez prikazanega uporabniškega vmesnika, uporabniškega agenta brskalnika nastavimo na "fri-wier-Skupina\_G", lokacijo izvršljive datoteke Firefoxa in omogočimo brskalniku, da sprejema neveljavne ali samopodpisane SSL certifikate.

Vzpostavi se povezava s podatkovno bazo postgreSQL, implementirana pa je tudi možnost, da se pred vpisovanjem pridobljenih podatkov v bazo le-ta izprazni.

Za zagotavljanje načina iskanja v širino smo za frontier uporabili vrsto, v kateri hranimo naslove spletnih strani, ki jih je treba obiskati. Niti po vrsti jemljejo strani iz frontierja, preverijo robots.txt in glede na odgovor izvedejo ustrezno dejanje. Če je stran označena kot prepovedana za obisk našega pajka, je le-ta ne obišče, sicer pa pošlje zahtevo za pridobitev strani in podatke ustrezno obdela.

V primeru, da je vsebina strani html, se stran prenese in njeno vsebino in podatke o njej shrani v podatkovno bazo. V primeru, da gre za sliko, se le-to shrani v tabelo 'images'. Vsebinskim datotekam z drugimi končnicami kodo tipa strani ('page\_type\_code') na binary, njihove vsebine pa ne shranjujemo v podatkovno bazo.

Spletni pajek na obdelanih straneh poišče nove hiperpovezave, ki jih shrani v frontier na konec vrste.

Za preprečevanje shranjevanja večih strani z isto vsebino smo v tabeli 'page' dodali stolpec 'hash\_page', v katerega z uporabo knjižnice hashlib in algoritma sha265 shranimo vrednost zgoščevalne funkcije. Preverimo, če novo izračunani hash že obstaja v tabeli in glede na to ustrezno ukrepamo – stran smatramo kot novo in shranimo v podatkovno bazo (s 'page\_type\_code' 'binary') ali pa jo označimo kot duplikat ('page\_type\_code' nastavimo na 'duplicate').

V primeru, da delovanje niti naleti na prazen frontier, počaka in večkrat pogleda še enkrat, s čimer smo povečali robustnost delovanja sistema. Poleg tega nadziramo tudi morebitne napake. (tuki mogoče dopišemo še, da če ne rata, večkrat proba …)

Aja, P.S. Smo že na 43650 (da ne bom spet spammala po Discordu) 😊