IEPS PA1 skupina G

Za implementacijo našega spletnega pajka smo izbrali programski jezik Python.

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 ali pa jo označimo kot duplikat.

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 …)