

IEPS Poročilo 1 - Ekipa "Vipavska Burja"

Matjaž Rupnik, Blaž Ličen

Marec 2021

1 Implementacija

Pajka smo sprogramirali tako, da uporablja podatkovno bazo (v nadaljevanju PB) kot repozitorij za shranjevanje podatkov o spletnih straneh, hkrati pa jo uporablja tudi kot "frontier". To deluje na način, da so strani ki so v frontierju v tabeli "page", označene kot "FRONTIER" v stolpcu "page_type_code". Ob zapisu v frontier se novi povezavi samodejno določi naslednji "id", kar skrbi, da se povezave iz frontierja jemljejo v pravilnem vrstnem redu oz. po načinu BFS.

Ob zagonu pajka se preveri ali je frontier (v PB) prazen. Če je prazen to pomeni, da je to prvi zagon pajka, zato se sproži funkcija "initFrontier", ki napolne frontier s podanimi URL naslovi. Če pa je v frontierju več kot 1 zapis tipa "FRONTIER" pomeni, da je pajek na znova zagnan, zato se zažene funkcija "initFrontierProcessing", ki spremeni vse zapise tipa "PROCESSING", v tip "FRONTIER". To je varovalka, če se pajek ustavi sredi procesiranja strani in mu ne uspe zapisati vseh podatkov o strani v bazo, se bo pajek ob ponovnem zagonu ponovno obiskal stran.

Nato se v ukazni vrstici uporabnika vpraša, s koliko procesi naj pajek deluje. Zatem se zažene izbrano število procesov.

1.1 Delovanje funkcije process

Vsak proces predstavlja zagon funkcije "process". Ta funkcija je ubistvu velika zanka, ki na začetku dobi naslednji URL iz baze (to stori tako, da pokliče funkcijo "getNextUrl"), nato preveri ali je domena te strani že v tabeli "site", če ni se na domeni poskusi dostopati do robots.txt. Če pravila obstajajo se shranijo v PB. Če pa je domena že v PB, iz baze naloži pravila za robots, če obstajajo. Nato se preveri ali sploh lahko gremo na ta URL. Če ne smemo iti na url se v PB zapiše "NOTALOWED", če pa lahko gremo na ta URL se sproži funkcija "fetchPageContent", ki pridobi HTML vsebino iz URL povezave.

Nato se pokliče funkcija "getHrefUrls", ki iz vsebine pridobi vse "<a href=" povezave in jih zapiše v frontier.

Nato se pokliče funkcija "getJsUrls", ki iz vsebine pridobi vse "onclick" povezave in jih zapiše v frontier.

Nato se pokliče funkcija "getImgUrls", ki iz vsebine pridobi vse "

