Crawler

V tem poročilu je predstavljena implementacija spletnega pajka pa1, ki je orodje, razvito v Pythonu za ekstrakcijo in analizo podatkov s spletnih strani. Opisane so ključne odločitve, konfiguracije in izzivi, s katerimi smo se srečali med razvojem, ter rešitve, ki smo jih uporabili.

Implementacija

Pajek je implementiran v Pythonu z uporabo modulov kot so requests za HTTP komunikacijo, BeautifulSoup iz bs4 za razčlenjevanje HTML in psycopg2 za interakcije z bazo podatkov PostgreSQL. Arhitektura sledi modularnemu pristopu, ki ločuje skrbi na posvečene komponente, kot so izvlečenje povezav, prenos vsebine in shranjevanje podatkov.

Razvojni izzivi

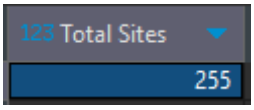
Pomemben izziv je bilo ravnanje z napakami zaradi premajhnega pomnilnika pri delu z velikimi SQL skriptami. Uporaba metod toka za obdelavo datotek je omogočila ravnanje z velikimi podatkovnimi izpisi brez izčrpavanja sistemskih virov.

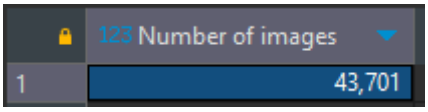
Za skladnost z najboljšimi praksami spletnega strganja smo implementirali logiko omejevanja hitrosti, da ne preobremenimo strežnikov. Integracija RobotFileParser in lastnih funkcij zakasnitve zagotavlja upoštevanje specifikacij robots.txt in vljudno brskanje.

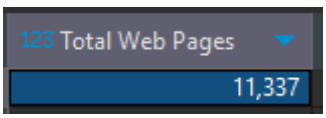
Crawler je sprva deloval tako, da ni hranil podatkov o obiskanih straneh (frontier pages) v podatkovno bazo, kar je zahtevalo, da se je moral začeti izvajati vedno znova od začetka. Z nadgradnjo smo to spremenili in omogočili shranjevanje teh strani v bazo, kar nam je posledično omogočilo, da smo lahko crawler začasno ustavili in ponovno zagnali, ne da bi izgubili dosedanji napredek.

V procesu nadgradnje baze podatkov smo uvedli nov stolpec content_hash v tabelo page, ki nam omogoča, da za vsako spletno stran ustvarimo hash vrednost HTML vsebine. S tem mehanizmom lahko učinkovito zaznamo in obvladujemo dvojnike, saj pred shranjevanjem nove strani v bazo primerjamo hash z že obstoječimi vrednostmi. Če se hash že pojavi v bazi, novo stran označimo kot dvojnika in se tako izognemo nepotrebnemu podvajanju podatkov. Ta pristop nam je omogočil bistveno zmanjšanje redundance v shranjenih podatkih in izboljšanje celotne učinkovitosti zbiranja podatkov.

Vsak thread, preden izvede operacijo nad podatkovno bazo, mora pridobiti ključavnico, kar zagotovi, da ima v danem trenutku ekskluziven dostop do baze. Po zaključku operacije se ključavnica sprosti, kar omogoča naslednjemu threadu, da varno izvede svojo operacijo. Ta mehanizem preprečuje tekmovanje stanj (race conditions) in ohranja integriteto naših podatkov. Zaradi teh izboljšav lahko naš crawler deluje v vzporednem okolju brez tveganja za korupcijo podatkov ali zaklepanja podatkovne baze, kar bi lahko privedlo do zastojev in zmanjšanja učinkovitosti.







123 Total Duplicates
208

ABC Binary Document Type	123 Number of Documents
DOC	15
PDF	95

123 Total Images
43,701

123 Average Images per Web Page
4.5417792559