Search Engine Projekat

By Teodor Vidaković SV33/2021

U projektu su korištene sledeće strukture:

- Trie stablo
- Usmjereni graf

Prilikom učitavanja unesenog direktorijuma:

- Kreira se jedno Trie stablo koje sadrži sve riječi, svaki End čvor sadrži riječnik koji za ključ ima naziv HTML fajla u kom se pojavljuje ta riječ a za vrijednost broj ponavljanja te riječi u tom fajlu.
- Kreira se usmereni graf gdje svaki čvor predstavlja jedan HTML fajl koji sadrži rječnik ulaznih I rječnik izlaznih grana iz čvora. (S1------>S2 => stranica S1 sadrži link koji ukazuje na stranicu S2)
- U globalnom rečniku se čuva parsirani tekst i linkovi stranice gdje je ključ odgovarajuća apsolutna putanja do date stranice. Koristi se kasnije za prikaz teksta prve stranice da se ne bi morala opet parsirati čitava stranica.

U projektu su samostalno implementirani sledeći algoritmi:

- Quick Sort (korišteno uglavnom za sortiranje rječnika)
- KMP algoritam za traženje fraze (***dodatni zadatak***)

U projektu kao dodatni zadatak je implementirano traženje fraze.

Prvo se traži presjek stranica u kojoj se nalaze sve riječi faze, a onda pomoću KMP algoritma se određuje da li postoji fraza u datoj stranici. Ako je povratni index “–1” znači da ne postoji, u suprotnom se ta stranica dodaje u novu listu stranica koje sadrže frazu. Za prvu stranicu se vraća dio teksta stranice koji sadrži traženu frazu