Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Jakob Dekleva, Matjaž Bevc, Urban Poljšak

Poročilo za tretjo domačo nalogo

Seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTOR: prof. dr. Marko Bajec, doc. dr. Slavko Žitnik

1. Uvod

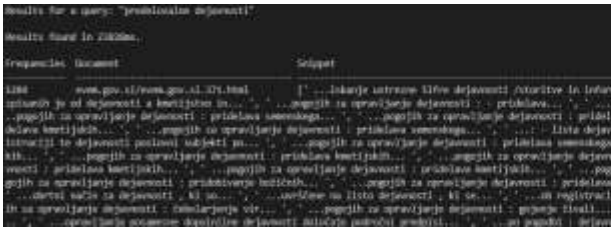
V poročilu je opisan razvoj tretje domače naloge pri predmetu Iskanje in ekstrakcija podatkov s spleta. Cilj naloge je implementacija preprostega iskalnika z uporabo obratnega indeksa, ki išče besede v izbranih dokumentih. Za primerjavo hitrosti je bilo potrebno implementirati še iskanje brez obratnega indeksa.

2. Preprocessing

Pred iskanjem je bilo treba dokumente še obdelati, torej za vsakega izluščiti vse uporabne besede, zabeležiti število njihovih pojavitev v dokumentu ter indekse teh pojavitev. Vse te rezultate smo nato shranili v podatkovno bazo dveh tabel, katerih prva hrani samo besede, druga pa poleg teh še pot do dokumenta, frekvenco ter indekse. Sistem je nato pripravljen na iskanje.

3 Poizvedbe – obratni indeks

Program za iskano besedo pridobi podatke iz baze in rezultate (dokumente) razvrsti po frekvenci besede. Če je besed več, to stori za vsako besedo in rezultate razvršča po vsoti frekvenc iskanih besed.

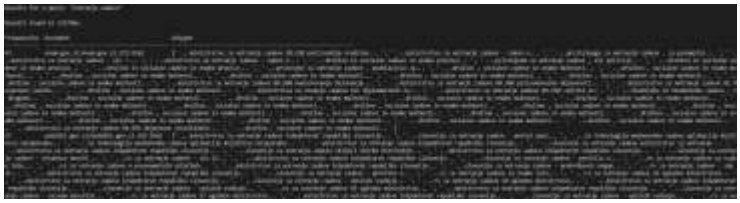


Slika 1: odsek rezultatov poizvedbe »pridelovalne dejavnosti«

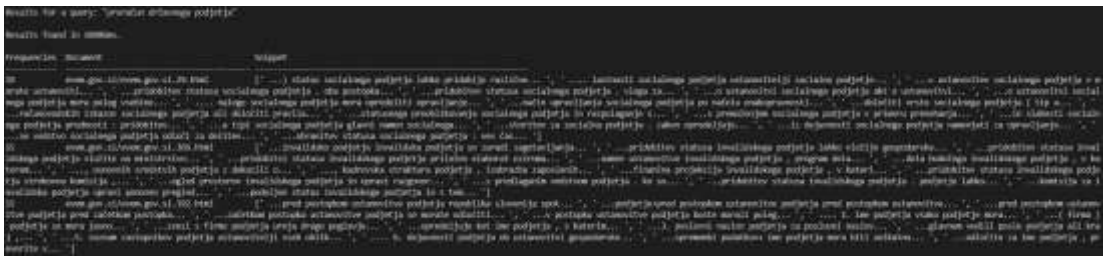
Slika 2: odsek rezultatov poizvedbe »trgovina«

Slika 3: odsek rezultatov poizvedbe »social services«

Slika 4: odsek rezultatov poizvedbe »vrednote«



Slika 5: odsek rezultatov poizvedbe »notranje zadeve«



Slika 6: odsek rezultatov poizvedbe »proračun državnega podjetja«

4. Poizvedbe – naivno iskanje

Za primerjavo hitrosti iskanja z obratnim indeksom smo implementirali še iskanje brez le-tega. V tem primeru je algoritem sorazmerno podoben, vendar pa ne uporablja podatkovne baze, ampak prečesava dokumente v iskanju iskane besede.

5. Rezultati

V spodnji tabeli je prikazana razlika med hitrostjo izvajanja run-basic-search in run-sqlite-search

Besede	run-sqlite-search.py v milisekundah	run-basic-search.py v milisekundah	Razlika v milisekundah	Razlika v sekundah
proračun državnega podjetja	17059	146650	129591	129,591
notranje zadeve	45215	153888	108673	108,673
pridelovalne dejavnosti	26134	145640	119506	119,506
social services	2435	135374	132939	132,939
trgovina	25026	140834	115808	115,808
vrednote	203	143151	142948	142,948

6. Težave

Med delom nismo imeli veliko težav, največja pa je bila neujemanje indeksov, ker smo pri shranjevanju in iskanju uporabljali malenkost drugačne postopke.

7. Zaključek

Z implementacijo smo sorazmerno zadovoljni. Z malo več časa bi verjetno še malo pohitrili iskanje besed (kot tudi vse ostale postopke). Naloga ni bila zelo zahtevna, a je vseeno terjala nekaj razmišljanja.