Seminarska naloga 3: Procesiranje in indeksiranje podatkov

Erik Rakušček, Jan Šmid, Qichao Chen
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

I. UVOD

V tem poročilu je opisan postopek procesiranja in indeksiranja podatkov. Opišemo tudi kako te podatke nato pridobimo s pomočjo (A) **obrnjenega indeksa** in (B) **brez obrnjenega indeksa**. Izvorna koda je dostopna na [1].

II. PROCESIRANJE PODATKOV Z INDEKSIRANJEM

Spletno stran smo najprej procesirali s knjižnico *BeautifulSoup*. Izbrisali smo vse elemente *script* in *style*, vse stop besede (z uporabo: [2]) in vse posebne znake (katere točno, lahko pogledate v spremenljivko *dissalowed* v datoteki *run-data-process*). Nato smo z uporabo knjižnice *nltk* tokenizirali besedilo in spremenili vse v male črke. Sprehodili smo se čez seznam besed (tokenov) in sproti gradili slovar (dictionary), ki smo ga na koncu dodali v bazo.

III. PODATKOVNA BAZA

Podatkovno bazo smo najprej implementirali po navodilih iz besedila naloge. Med implementacijo generiranja osnutkov pa smo ugotovili, da je najbolj učinkovito to, da v tabelo *Posting* dodamo stolpec *snippets*. V *snippets* shranimo niz osnutkov vseh pojavitev besede na posamezni spletni strani. Posamezen osnutek je sestavljen iz treh besed levo in treh besed desno od iskane besede, kjer ločil ne upoštevamo.

IV. ISKANJE Z UPORABO OBRNJENEGA INDEKSA

Iskanje z uporabo obrnjenega indeksa smo implementirali tako, da iskalni niz najprej razdelimo na besede. Nato se z zanko sprehodimo čez vse besede in z SQL poizvedbo iz podatkovne baze pridobimo vse vrstice za določeno besedo v tabeli *Posting*. Z dodatno zanko se nato sprehodimo čez vse pridobljene vrstice in seštevamo število pojavitev besede (*frequency*) v posameznih vrsticah. Prav tako združujemo osnutke (*snippets*) iz posameznih vrstic. Na koncu samo še razvrstimo strani po številu pojavitev besede in vrnemo rezultat.

V. ISKANJE BREZ OBRNJENEGA INDEKSA

Iskanje brez obrnjenega indeksa smo implementirali tako, da smo iskalni niz poiskali na vseh spletnih straneh. Datoteke smo naprej obdelali na podoben način kot pri iskanju z uporabo obrnjenega indeksa. Izbrisali smo elemente *script* in *style*, ohranili pa smo stop besede in posebne znake. Nato smo s knjižnico *nltk* razčlenili besedilo. Pri vsaki vhodni datoteki smo primerjali iskalni niz z vsakim nizom iz datoteke, iskalni niz in niz s katerim primerjamo smo pretvorili v male črke, da iskanje ni občutljivo na male in velike črke. Osnutke pa smo izpisali iz originalnega besedila. Sproti smo prištevali frekvenco pojavitev.

VI. REZULTATI

V bazo se je shranilo 35755 unikatnih besed. Najpogostejše besede so: *podatkov* (10502), *slovenije* (9856), *republike* (8583), *podatki* (5563), *dejavnosti* (5560).

Primerjali smo čase izvajanja iskanja z uporabo podatkovne baze in iskanja brez podatkovne baze. Rezultate smo pridobili z izvajanjem programa na računalniku s trdim diskom (brez SSD).

TABLE I
PRIMERJAVA ČASA ISKANJA

Iskalni niz	SQLite	brez SQLite
Sistem SPOT	0.28s	58.25s
predelovalne dejavnosti	0.40s	53.74s
trgovina	0.11s	54.34s
social services	0.18s	56.08s
ministrstvo za zdravje	0.47s	59.37s
republika slovenija	0.24s	56.61s
računalništvo in informatika	0.26s	56.59s

TABLE II
ŠTEVILO STRANI NA KATERIH SE POJAVI NIZ IN NAJVEČJE ŠTEVILO POJAVITEV NA POSAMEZNI STRANI

Iskalni niz	Št. strani	Največ
Sistem SPOT	1287	69
predelovalne dejavnosti	753	1288
trgovina	125	364
social services	4	5
ministrstvo za zdravje	1260	667
republika slovenija	994	126
računalništvo in informatika	13	4

VII. ZAKLJUČEK

Med izdelavo seminarske naloge smo se naučili indeksirati vsebino spletnih strani za hitro iskanje ključnih besed na le-teh. Z analizo metod iskanja smo ugotovili, da je iskanje z uporabo v naprej pripravljene podatkovne baze oziroma indeksa bistveno hitrejše od navadnega prečesavanja spletnih strani za vsako iskanje posebej.

REFERENCES

[1] Github repozitorij. Dosegljivo: https://github.com/van123helsing/pachong-11. [Dostopano: 22. 5. 2020].

[2] Slovenske stop besede. Dosegljivo: https://szitnik.github.io/wier-labs/data/pa3/stopwords.py. [Dostopano: 22. 5. 2020].