

# Programska naloga 3 - Indeksiranje in poizvedovanje

Kim Ana Badovinac, Jakob Petek, Jovan Prodanov  
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

## I. UVOD

Indekser je programska oprema, ki procesira vsebino spletne strani in indeksira vsako besedo na posamezni spletni strani. Vsaka beseda je shranjena v podatkovno bazo, kjer so povezane besede in spletne strani. Ko je indeks zgrajen, ga lahko uporabimo za poizvedovanje. Naša implementacija je sestavljena iz grajenja indeksa in algoritmov za poizvedovanje z uporabo indeksa in brez njega. Poizvedovanje z uporabo indeksa je bistveno hitrejšo kot poizvedovanje brez njega.

## II. IMPLEMENTACIJA

Komponente programske naloge smo implementirali v programskem jeziku Python. Pridobljene podatke shranjujemo v podatkovno bazo SQLite, ki jo kreiramo s programom *create-database.py*. Implementacija je sestavljena iz procesiranja podatkov z indeksiranjem (*run-indexing.py*), poizvedovanja podatkov z uporabo invertnega indeksa (*run-sqlite-search.py*) in poizvedovanja podatkov brez uporabe invertnega indeksa (*run-basic-search.py*).

### A. Procesiranje podatkov z indeksiranjem

Sprehodili smo se čez vse spletne strani in za vsako stran ekstrahirali tekstualni kontekst, preprocesirali podatke in shranili v indeks. Tekstovno vsebino smo iz datoteke pobrali s pomočjo knjižnice BeautifulSoup. Nato smo besedilo pretvorili v seznam besed (angl. tokens) s pomočjo knjižnice NLTK. Besede smo nato normalizirali v majhne črke. Odstranili smo tudi besede iz seznama neveljavnih besed (angl. stopwords), kamor smo dodali še nekaj simbolov, ki jih ne želimo shraniti v indeks: !, ×, −, −, −, ., (, ), [, ], , , ' , ; , /, ©, >, <, ", " ipd. Te simbole smo zasledili pri testiranju gradnje indeksa in jih ne potrebujemo v indeksu. Poleg simbolov odstranimo tudi besede, sestavljene samo iz števil in simbolov, torej ohranimo samo besede, ki vsebujejo črke. Isti postopek procesiranja je uporabljen tudi na podani poizvedbi uporabnika v postopku poizvedovanja.

Po zaključenem procesiranju besed, filtrirane besede shranimo v indeks. Vse unikatne besede shranimo v tabelo *IndexWord*. Poleg tega vstavimo v tabelo *Posting* tudi podatke (*beseda, ime dokumenta, frekvenca, seznam indeksov kjer se beseda pojavi*).

### B. Poizvedovanje podatkov z uporabo invertnega indeksa

Z uporabo zgrajenega invertnega indeksa lahko dobimo rezultate na poizvedbe (angl. query). S SQL poizvedbo pridobimo vse dokumente, ki vsebujejo besede iz poizvedbe. Poizvedba za vsak dokument vrne ime dokumenta, vsoto frekvenc

in seznam indeksov, kjer se besede pojavijo v besedilu. Prav tako nastavimo poizvedbo, da nam vrne seznam dokumentov urejen padajoče po frekvenci. Poleg teh rezultatov, formuliramo izseke iz besedila okoli indeksiranih besed. Vse rezultate izpišemo na standardni izhod.

### C. Poizvedovanje podatkov brez uporabe invertnega indeksa

Za primerjavo smo implementirali tudi poizvedovanje brez indeksa. Za implementacijo le-tega smo združili elemente indeksiranja in elemente poizvedovanja. Pri tem smo se sprehodili čez vse spletne strani (tako kot pri indeksiranju) in za vsako stran ekstrahirali tekstualni kontekst in procesirali podatke. Na tem mestu smo poiskali, ali obstaja beseda iz poizvedbe v seznamu besed iz datoteke. Shranili smo si indekse, kjer se beseda pojavi ter oblikovali izseke iz besed okoli indeksiranih besed. Najdene besede in rezultate za to besedo smo si shranili v tabelo. V primeru, da smo besedo poindeksirali v dokumentu, ki je bil že najden in shranjen v tabelo rezultatov, seštejemo njune frekvence in ne shranimo ponovno tega dokumenta v tabelo. Ob koncu sprehajanja čez vse datoteke, smo to tabelo uredili po frekvenci padajoče. Vse rezultate izpišemo na standardni izhod.

## III. STATISTIKA IN REZULTATI

### A. Podatkovna baza

Implementiran indeks je pregledal 1416 spletnih strani iz 4 domen, obdelal vse besede in filtrirane besede shranil v podatkovno bazo. Baza je po kreiranju indeksa sestavljena iz 37.448 indeksiranih besed in 356.400 povezav med besedami in dokumenti (angl. posting). Povezave z največjimi frekvencami so navedene v tabeli I. Največ pojavitev v enem dokumentu je imale beseda *proizvodnja*.

Beseda	Dokument	Število pojavitev
proizvodnja	evem.gov.si/evem.gov.si.371.html	2.266
gl	evem.gov.si/evem.gov.si.371.html	1.668
spada	evem.gov.si/evem.gov.si.371.html	1.338
dejavnosti	evem.gov.si/evem.gov.si.371.html	1.284
d.o.o	podatki.gov.si/podatki.gov.si.340.html	967

Tabela I

BESEDE IN DOKUMENTI Z NAJVEČJIM ŠTEVILOM POJAVITEV.

### B. Rezultati testnih poizvedb

Primerjali smo časovno zahtevnost poizvedb z uporabo inverznega indeksa in brez njegove uporabe za naslednje poizvedbe (Tabela II): predelovalne dejavnosti, trgovina, social

services, računalništvo, vozniško dovoljenje in geodetski vestnik. Rezultati poizvedb so zaradi velikosti omejeni na prvih 5 rezultatov in priloženi na zadnjih straneh poročila (Slika 1-6). Rezultati z uporabo indeksa in brez njega so enaki, zato smo priložili samo rezultate z uporabo indeksa.

Poizvedba	Čas z uporabo indeksa	Čas brez uporabe indeksa
“predelovalne dejavnosti”	0.141s	53.871s
“trgovina”	0.009s	48.993s
“social services”	0.003s	48.818s
“računalništvo”	0.001s	51.585s
“vozniško dovoljenje”	0.043s	50.359s
“geodetski vestnik”	0.016s	50.331s

Tabela II

BESEDE IN DOKUMENTI Z NAJVEČJIM ŠTEVILOM POJAVITEV.

Povprečni čas poizvedovanja z uporabo inverznega indeksa je 0.036 sekund. Povprečni čas poizvedovanja brez uporabe inverznega indeksa je 50.326 sekund. Kot pričakovano se pridobivanje podatkov izredno pohitri, če uporabimo inverzni indeks.

#### IV. ZAKLJUČEK

V seminarski nalogi smo uspešno implementirali in poizvedovali z inverznim indeksom. Polega tega smo implementirali sekvenčno iskanje poizvedovalnih besed po vseh dokumentih brez indeksa. Gradnja indeksa sicer traja nekaj časa, vendar ko imamo indeks zgrajen, poteka poizvedovanje skrajno hitro. Če poizvedujemo brez indeksa se čas poizvedovanja poveča kar 1400-krat. Zato se izredno obrestuje nekaj časa na začetku za gradnjo indeksa.

Frequencies	Document	Snippet
1284	evem.gov.si/evem.gov.si.371.html	...ustrezne šifre dejavnosti /storitve informacij...pogojih opravljanje dejavnosti iskalnik vpišite...ok izpisa
75	evem.gov.si/evem.gov.si.377.html	...defektolog zdravstveni dejavnosti dekan direktor...dietetik zdravstveni dejavnosti dimnikar diplomirana...so
40	podatki.gov.si/podatki.gov.si.340.html	...nosilec dopolnilne dejavnosti kmetiji bregar...center interesnih dejavnosti ptuj center...šolskih obšolskih
39	evem.gov.si/evem.gov.si.452.html	nastavitve storitvene dejavnosti druge nerazvrščene...druge nerazvrščene dejavnosti evem republika...evem/dej
31	evem.gov.si/evem.gov.si.653.html	...dovoljenje opravljanje dejavnosti specializirane prodajalne...radijske televizijske dejavnosti dovoljenje iz

Slika 1. Rezultat poizvedbe **predelovalne dejavnosti**.

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	...organizacij gl trgovina debelo kmetijskimi...juh gl trgovina debelo me
94	evem.gov.si/evem.gov.si.651.html	...dozimetrija govedoreja trgovina drobnno specializiranih...specializiran
92	evem.gov.si/evem.gov.si.21.html	...e-vem evem:področja trgovina našli informacije...seznam dejavnosti trg
82	podatki.gov.si/podatki.gov.si.340.html	...d.o.o dent trgovina storitve d.o.o...adria investicije trgovina posred
13	evem.gov.si/evem.gov.si.623.html	zasebnosti.xsprememba nastavitve trgovina debelo izdelki...široke porabe

Slika 2. Rezultat poizvedbe **trgovina**.

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.9.html	...labour retirement social services health...employment relationship soc
5	e-uprava.gov.si/e-uprava.gov.si.45.html	...labour retirement social services health...employment relationship soc
1	podatki.gov.si/podatki.gov.si.340.html	...and spa services ltd. terme...
1	evem.gov.si/evem.gov.si.661.html	...and related services ajpes and...

Slika 3. Rezultat poizvedbe **social services**.

Frequencies	Document	Snippet
4	podatki.gov.si/podatki.gov.si.12.html	...študentov fakultete računalništvo informatiko okviru...aprilu fakultet
4	podatki.gov.si/podatki.gov.si.14.html	...študentov fakultete računalništvo informatiko okviru...aprilu fakultet
4	podatki.gov.si/podatki.gov.si.534.html	...študentov fakultete računalništvo informatiko opsi...študentov fakulte
3	podatki.gov.si/podatki.gov.si.295.html	...aprilu fakulteti računalništvo informatiko ljubljani...razvila fakulte
2	podatki.gov.si/podatki.gov.si.327.html	...študentov fakultete računalništvo informatiko lt...ljubljan fakultete

Slika 4. Rezultat poizvedbe **računalništvo**.

Frequencies	Document	Snippet
188	evem.gov.si/evem.gov.si.371.html	...uporabnik pridobiti dovoljenje pristojnem organu...potrebno pridobiti dovoljenje izda agencija...ribolov potrebno dovoljenje ministrstva kmetijstvo...ribičev posebno dovoljenje gospodarski ribolov...gospodarski ribolov dovoljenje gospodarski ribolov...
107	evem.gov.si/evem.gov.si.653.html	...medicine licenca dovoljenje opravljanje dejavnosti...zdravili drobno dovoljenje ministrstva zdravje...ministrstva zdravje dovoljenje ministrstva zdravje...ministrstva zdravje dovoljenje poenostavljenem postopku...poenostavljenem postopku dovoljenje unije dovoljenje...
16	evem.gov.si/evem.gov.si.398.html	...knjige sklepov dovoljenje atvp spremembo...ozs obrtno dovoljenje pridobiti obrtno...pridobiti obrtno dovoljenje poslovni subjekt...pridobiti obrtno dovoljenje področje obrtnih...register obrtno dovoljenje postane član...
10	evem.gov.si/evem.gov.si.84.html	...veljavno enotno dovoljenje prebivanje delo...srbije delovno dovoljenje bilateralnem sporazumu...ebdp enotno dovoljenje združuje nadomešča...združuje nadomešča dovoljenje prebivanje izdajale...enote delovno dovoljenje izdajali zavodu...
7	evem.gov.si/evem.gov.si.43.html	...pridobiti ustrezno dovoljenje začne opravljati...prostor uporabno dovoljenje poslovni prostor...preverite ima dovoljenje uporabo ustrezne...praviloma uporabno dovoljenje glede dejavnost...dejavnost uporabno dovoljenje izda pristojna...

Slika 5. Rezultat poizvedbe **vozniško dovoljenje**.

Frequencies	Document	Snippet
74	e-prostor.gov.si/e-prostor.gov.si.150.html	...slovenije geodetski vestnik letn št...koper geodetski vestnik letn št...croatia geodetski vestnik letn št...sloveniji geodetski vestnik letn št...sloveniji geodetski vestnik letn št...
30	e-prostor.gov.si/e-prostor.gov.si.18.html	...gis geodetski vestnik berk s...slovenijo geodetski vestnik boldin d...epsg geodetski vestnik kete p...6500d96/xyz slovenski geodetski sistem trirazsežnih...6423d96/φλh slovenski geodetski sistem trirazsežnih...
27	e-prostor.gov.si/e-prostor.gov.si.38.html	...slovenije geodetski vestnik berk s...d48 geodetski vestnik hočev m...mreže geodetski vestnik medved k...slovenija geodetski vestnik medved k...etrs89 geodetski vestnik medved k...
20	e-prostor.gov.si/e-prostor.gov.si.42.html	...izmere geodetski vestnik koler b...slovenije geodetski vestnik koler b...sloveniji geodetski vestnik koler b...slovenije geodetski vestnik koler b...sistema geodetski vestnik kuhar m...
15	e-prostor.gov.si/e-prostor.gov.si.46.html	...slovenije geodetski vestnik berk s...sistem geodetski vestnik berk s...d96/tm geodetski vestnik berk s...pomurju geodetski vestnik dornik snoj...s-transformacija geodetski vestnik sedež d...

Slika 6. Rezultat poizvedbe **geodetski vestnik**.