# Seminarska naloga 3 - Indeksiranje in poizvedovanje

Katjuša Jaklič, Luka Kuzman, Tjaša Mlakar
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Maj 2022

## 1. Uvod

Spletni indekser je program, ki ekstrahira podatke s spletnih strani in jih organizira tako, da doseže kar se da hitre rezultate oziroma odgovore glede na podano poizvedbo. Indeksiranje metapodatkov vključuje dodeljevanje ključnih besed, besednih zvez ali opisov posameznim spletnim mestom, na podlagi česar so le-ta pripravljena za nadaljni proces razvrščanja po relevantnosti.

## 2. Implementacija

V sklopu seminarske naloge smo implementirali spletni indekser in orodje za poizvedovanje. Implementacija je sestavljena iz treh osrednjih delov: procesiranja podatkov z indeksiranjem (`indexing.py`), pridobivanja podatkov z inverznim indeksom (`run-sqlite-search.py`) ter pridobivanja podatkov brez inverznega indeksa (`run-basic-search.py`). Pri poizvedovanju smo dali prednost stranem, kjer se je beseda pojavila največkrat (torej je imela poizvedovalna fraza največjo frekvenco).

Poleg tega smo implementirali tudi preprosti spletni iskalnik, prisoten v mapi `Poogle` in s tem uporabniku olajšali uporabo programa, hkrati pa smo v njem stestirali še nekaj konceptov.

### 2.1. Procesiranje podatkov z indeksiranjem

Pri sprehodu skozi datoteke smo pomočjo knjižnice BeautifulSoup iz njih pobrali tekstovno vsebino. Pred indeksiranjem pa smo morali tekst spremeniti v ustrezno obliko s filtriranjem nepomembnih podatkov. Le-to smo storili s pomočjo knjižnice *NLKT*, ki je tekst pretvorila v seznam žetonov (ang. *tokens*). Iz seznama smo nato odstranili stop besede (ang. *stop words*) in nekatere ostale žetone, za katere smo menili, da za poizvedbo niso pomembni. Preostale besede smo normalizirali tako, da smo jih spremenili v male črke.

### 2.2. Pridobivanje podatkov z inverznim indeksom

S pomočjo podatkov, ki smo jih shranili v podatkovno bazo, lahko naredimo poizvedbo. Po bazi poizvedemo s sledečo SQL kodo (Figure 1), ki nam vrne število pojavitev besed v določenem dokumentu. Tu spremenljivka `arguments` vsebuje besede, s katerimi poizvedujemo. Izvlečke smo pridobili z iskanjem besede, katere datoteka in indeks v tej datoteki se skladata z iskanim.

```
SELECT p.documentName AS docName,
 SUM(frequency) AS freq,
 GROUP_CONCAT(indexes) AS idxs
FROM Posting p
WHERE
    p.word IN (''' + arguments + ''')
GROUP BY p.documentName
ORDER BY freq DESC;
```

Figure 1: SQL skripta za pridobivanje podatkov z inverznim indeksom.

### 2.3. Pridobivanje podatkov brez inverznega indeksa

Pridobivanje podatkov smo poskusili tudi brez implementacije inverznega indeksa. Pri tem pristopu smo se le sprehodili skozi vse datoteke in v njih poiskali vsa ujemanja podanih argumentov. Izvlečke smo pridobili z branjem sosednjih elementov v seznamu žetonov.

## 3. Podatkovna baza

Baza je sestavljena iz 49.092 različnih besed, med katerimi so se najbolj pogosto pojavile besede, prikazane v tabeli 1.

TABLE 1: Tabela prvih 5 besed z največ pojavitvami.

| Beseda | Skupno število pojavitev |
|---|---|
| podatkov | 10.502 |
| slovenije | 9.856 |
| republike | 8.583 |
| podatki | 5.562 |
| dejavnosti | 5.560 |

Največ pojavitev v enem dokumentu je imela beseda proizvodnja, in sicer v dokumentu 'evem.gov.si.371.html', kar je razvidno iz spodnje tabele 2.

V besedilu je ostalo tudi nekaj žetonov, ki jih filtriranje ni odstranilo, vendar jih kljub temu ne bi označili za pomembne. Le te smo pri ogledu najpogostejših besed izpustili.

```
SELECT p.word, SUM(frequency) AS freq
FROM Posting p
GROUP BY p.word
ORDER BY freq DESC;
```

Figure 2: SQL skripta za prikaz najpogosteje pojavljajočih se besed.

TABLE 2: Tabela prvih 5 besed z največ pojavitvami v enem dokumentu.

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

TABLE 3: Tabela prvih 5 datotek z največ besedami.

| Dokument | Število brsed |
|---|---|
| evem.gov.si.371.html | 80.966 |
| podatki.gov.si.340.html | 27.433 |
| e-prostor.gov.si.166.html | 11.163 |
| e-prostor.gov.si.147.html | 5.762 |
| e-prostor.gov.si.218.html | 4.500 |

## 4. Rezultati

Čas poizvedbe smo primerjali tako s pomočjo inverznega indeksa kot tudi brez njega. Kot pričakovano se izkaže, da je pridobivanje podatkov brez inverznega indeksa potekalo precej počasneje. Le-to si lahko ogledamo na primeru šestih poizvedb:

- PREDELOVALNE DEJAVNOSTI
- TRGOVINA
- SOCIAL SERVICES
- AVTOMOBIL
- SLOVAR SLOVENSKEGA KNJIŽNEGA JEZIKA
- ZDRAVILO

Ker so ponekod rezultati precej dolgi, smo se odločili, da za boljšo preglednost priložimo le prvih pet rezultatov vsake poizvedbe, prvih šest pojavitev besede v dokumentu ter en izrez besedila (ang. *snippet*). Rezultati poizvedb z omenjenimi omejitvami so prikazani na zadnjih treh straneh.

Oglejmo si povprečen čas pridobitve rezultata poizvedbe, pridobljen s pomočjo zgornjih šestih poizvedb:

- z inverznim indeksom: $0,62575$ sekund (brez ustvarjanja izvlečkov $0,10505$ sekund)
- brez inverznega indeksa: $30,63822$ sekund

Opazimo lahko izredno pohitritev iskanja z inverznim indeksom v primerjavi z odsotnostjo le-tega. Pri iskanju izvlečka besedila ob poizvedbi z inverznim indeksom se je proces poizvedbe malo podaljšal, saj smo v podatkovni bazi morali poiskati tudi vse besede, ki se pojavljajo okoli indeksa poizvedene besede, kar prikažemo s primerjavo časa v oklepaju. Le-tega v drugem primeru ni, saj le izpišemo žetone, ki ležijo poleg iskanega. V primeru, da bi želeli izpisati še več izvlečkov, bi se čas poizvedbe prvega primera v našem pristopu podaljšal.

## 5. Psevdo-iskalnik

Kot zadnje smo se s pomočjo knjižnice Flask odločili implementirati iskalnik (ang. *search engine*), s tem pa želeli upredmetiti koncepte te naloge oziroma jih predstaviti na način, razumljiv vsem uporabnikom. Ob zagonu in obisku naslova LOCALHOST:5000 se nam prikaže polje za iskanje, prikazano na Figure 3 s poizvedbo TRGOVINA.


Figure 3: Začetna stran psevdo-iskalnika.

V iskalno polje vnesemo poljubno poizvedbo (na podoben način, kot bi to storili z vnosom poizvedbe v ukazno vrstico), nato pa nam program vrne prvih 20 rezultatov poizvedbe, kot je prikazano na Figure 4. Pri tem smo se odločili za ustvarjanje hiperpovezave z naslovom, ki jo pridobimo iz same HTML datoteke. Zraven naslova prikažemo lokacijo datoteke, kot tudi število zadetkov, torej število ujemanj poizvedbe z besedami v datoteki.


Figure 4: Prvih 20 rezultatov poizvedbe TRGOVINA.

Za prikaz izreza smo se odločili za drugačen pristop. Ker se pri gradnji izreza, omenjenega v poglavju 2.2, omejimo le na besede, ki jih nismo filtrirali kot nepomembne informacije, so pogosto izrezi izgledali nesmiselni. Tu se zato lotimo drugačnega pristopa. Ker smo vsako datoteko tu že tako ali tako prebrali, da smo iz nje pridobili naslov, smo le-to uporabili tudi za pridobivanje celotnega teksta. Tvorba izvlečka je nato potekala iz originalnega teksta; v originalnem besedilu smo poiskali besedo. Tu nam indeks pojavitve besede, kot smo ga shranili v podatkovno bazo, ne pride prav. S tovrstnim načinom smo velikokrat dobili bolj smiselen izvleček, vendar pa je proces zaradi tega trajal dlje časa.

## 6. Zaključek

V seminarski nalogi smo uspešno implementirali indekser ter z njegovo pomočjo uspešno izvajali poizvedbe v naboru HMTL datotek, poizvedovanje pa smo implementirali tudi brez prisotnosti le-tega. Nato smo rezultate za obe vrsti poizvedb med seboj primerjali in ugotavljali posamezno učinkovitost. Izkaže se da kljub dejstvu, da moramo na začetku porabiti nekaj časa za izgradnjo indeksa, se pri hitrosti pridobivanja poizvedb le-to izredno obrestuje.

Zanimiva je bila eksperimentacija tvorjenja izvlečkov, ki smo se je lotili na kar nekaj načinov. Opazili smo, da ima vsak izmed njih svoje pomanjkljivosti, vendar menimo, da pridobivanje izvlečkov v psevdo-iskalniku deluje najbolje oziroma je najbližje tistemu v realnih aplikacijah, četudi zaradi tokenizacije ne moramo direktno dostopati do indeksov, kjer se beseda pojavi. Prav tako je zaradi branja vsake datoteke, iz katere gradimo izvleček, izvedba nekoliko počasnejša. Le-to bi lahko pohitrili z shranjevanjem celotne HTML datoteke v podatkovni bazi (ali pa le izvlečka, ki bi ga povezali z indeksom), vendar bi podatkovna baza tedaj zasedla precej več prostora. Poizvedovanje bi dodatno izboljšalo tudi dajanje večje prioritete besednim zvezam (recimo SLOVAR SLOVENSKEGA KNJIŽNEGA JEZIKA) in združevanje različnih oblik besede (recimo sklanjatve) v eno.

```
Hits: 1284
        Doc: '\evem.gov.si\evem.gov.si.371.html'
        Indexes: 39, 44, 61, 63, 198, 243, ...
        Snippet: '... samostojne republika domov dejavnosti republika samostojne
        portal ...'
Hits: 75
        Doc: '\evem.gov.si\evem.gov.si.377.html'
        Indexes: 91, 107, 331, 335, 344, 411, ...
        Snippet: '... straže semena strokovni dejavnosti dekan strokovni
        strokovni ...'
Hits: 40
        Doc: '\podatki.gov.si\podatki.gov.si.340.html'
        Indexes: 1486, 1686, 1796, 2660, 2690, 3167, ...
        Snippet: '... razvoj jeziku zdravje dejavnosti varnost družba
        zobozdravstvena ...'
Hits: 39
        Doc: '\evem.gov.si\evem.gov.si.452.html'
        Indexes: 1, 5, 38, 43, 50, 58, ...
        Snippet: '... storitvene dejavnosti storitvene storitvene storitvene ...'
Hits: 31
        Doc: '\evem.gov.si\evem.gov.si.653.html'
        Indexes: 103, 196, 200, 204, 208, 213, ..
        Snippet: '... licenca izpit sistemov dejavnosti specializirane prodajalne
        prevoz ...'
```

Figure 5: Poizvedba besedne zveze PREDELOVALNE DEJAVNOSTI.

```
Hits: 364
        Doc: '\evem.gov.si\evem.gov.si.371.html'
        Indexes: 2847, 9214, 11885, 11891, 22087, 22091, ...
        Snippet: '... proizvodnja proizvodnja dejavnosti trgovina proizvodnja
        dejavnosti proizvodnja ...'
Hits: 94
        Doc: '\evem.gov.si\evem.gov.si.651.html'
        Indexes: 215, 219, 223, 228, 783, 786, ...
        Snippet: '... škodljivih izdelkov energetskih trgovina plinov kmetijskih
        oseb ...'
Hits: 92
        Doc: '\evem.gov.si\evem.gov.si.21.html'
        Indexes: 32, 56, 60, 67, 69, 74, ...
        Snippet: '... zapiram e-vem evempodročja trgovina trgovina trgovina
        trgovina ...'
Hits: 82
        Doc: '\podatki.gov.si\podatki.gov.si.340.html'
        Indexes: 281, 290, 384, 420, 453, 890, ...
        Snippet: '... slovenije družba finance trgovina d.o.o slovenije družba ...'
Hits: 13
        Doc: '\evem.gov.si\evem.gov.si.623.html'
        Indexes: 0, 41, 47, 54, 59, 65, 71, 74, 78, 86, 93, 115, 227
        Snippet: 'trgovina trgovina trgovina trgovina ...'
```

Figure 6: Poizvedba besedne zveze TRGOVINA.

```
Hits: 5
        Doc: '\e-uprava.gov.si\e-uprava.gov.si.9.html'
        Indexes: 72, 233, 241,73, 234
        Snippet: '... personal documents certificates social of residence death ...'
Hits: 5
        Doc: '\e-uprava.gov.si\e-uprava.gov.si.45.html'
        Indexes: 72, 233, 241,73, 234
        Snippet: '... personal documents certificates social of residence death ...'
Hits: 1
        Doc: '\podatki.gov.si\podatki.gov.si.340.html'
        Indexes: 22707
        Snippet: '... recreation and spa services ltd. terme maribor ...'
Hits: 1
        Doc: '\evem.gov.si\evem.gov.si.661.html'
        Indexes: 495
        Snippet: '... records and related services ajpes and the ...'
```

Figure 7: Poizvedba besedne zveze SOCIAL SERVICES.

```
Hits: 5
        Doc: '\podatki.gov.si\podatki.gov.si.200.html'
        Indexes: 7, 310, 371, 404, 465
        Snippet: '... povprečno povprečno povprečno avtomobil povprečno povprečno
        avtomobil ...'
Hits: 2
        Doc: '\e-uprava.gov.si\e-uprava.gov.si.41.html'
        Indexes: 140, 142
        Snippet: '... avtomobila storiti prodati avtomobil postopek avtomobil
        kupiti ...'
Hits: 1
        Doc: '\podatki.gov.si\podatki.gov.si.196.html'
        Indexes: 623
        Snippet: '... poraba goriva osebni avtomobil delež osebnih avtomobilov ...'
Hits: 1
        Doc: '\podatki.gov.si\podatki.gov.si.35.html'
        Indexes: 624
        Snippet: '... poraba goriva osebni avtomobil delež osebnih avtomobilov ...'
Hits: 1
        Doc: '\podatki.gov.si\podatki.gov.si.442.html'
        Indexes: 654
        Snippet: '... poraba goriva osebni avtomobil delež osebnih avtomobilov ...'
```

Figure 8: Poizvedba besedne zveze AVTOMOBIL.

```
Hits: 15
        Doc: '\evem.gov.si\evem.gov.si.371.html'
        Indexes: 59550, 60464, 64665, 64671, 65032, 65055, ...
        Snippet: '... oseba pridobila dovoljenje slovenskega inštituta revizijo
        opravljanje ...'
Hits: 4
        Doc: '\podatki.gov.si\podatki.gov.si.319.html'
        Indexes: 456,306,457, 459
        Snippet: '... območjih občin organu slovenskega jezika uradna jezika ...'
Hits: 4
        Doc: '\podatki.gov.si\podatki.gov.si.307.html'
        Indexes: 0, 171, 296, 298
        Snippet: 'slovar slovar slovar podatki ...'
Hits: 4
        Doc: '\e-prostor.gov.si\e-prostor.gov.si.150.html'
        Indexes: 1922, 1955, 1987, 2027
        Snippet: '... 12. strokovno srečanje slovenskega združenja geodezijo
        geofiziko ...'
Hits: 2
        Doc: '\podatki.gov.si\podatki.gov.si.340.html'
        Indexes: 4504, 26239
        Snippet: '... okolje prostor zavod slovenskega prostor zavod republike ...'
```

Figure 9: Poizvedba besedne zveze SLOVAR SLOVENSKEGA KNJIŽNEGA JEZIKA.

```
Hits: 1
    Doc: '\evem.gov.si\evem.gov.si.371.html'
    Indexes: 72342
    Snippet: '... učijo izdelovati lastno zdravilo društvo kmečkih gospodinj ...'
```

Figure 10: Poizvedba besedne zveze ZDRAVILO.