Seminar 3

Nejc Ribič, Klemen Jesenovec

veliki traven, 2019

1 Uvod

Pri tretji seminarski nalogi smo implementirali invertni indeks z že pripravljenimi stranmi [1].

Poročilo je sestavljeno iz štirih delov. V prvem delu najprej predstavimo pristop čiščenja podatkov. Nato sledi predstavitev gradnje invertnega indeksa. Sledi predstavitev implementacije iskanja po invertnem indeksu. Pred koncem predstavimo še implementacijo sekvenčnega iskanja. Na koncu sledita še zaključek in sklep ter kot bonus pogalavje še seznam zaslonskih slik.

Izvorna koda je dostopna na GitHub strani [2].

2 Čiščenje podatkov

Čiščenja podatkov smo se lotili v treh korakih. Datoteke smo najprej rekurzivno prebirali. Nato smo besede iz dokumenta tokenizirali z uporabo python knjižnice *nltk*. Sledil je postopek spreminjanja vseh črk v besedah v male črke. Preden smo odstranili nepomembne besede, smo vse žetone oštevilčili, saj lahko tako enostavno najdemo mesto pojavitve besede in njeno soseščino. Besede smo še dodtano očistili z regularnim izrazom "[a-zA-Z0-9][a-zA-Z0-9-]+ ". Nato pa je sledil postopek kreiranja indeksa. To smo storili tako, da smo prešeli pojavitve posamezne besede v listu žetonov in si shranili tudi soseščine.

Dodana vrednost shranjevanja soseščine je predstavljena v opisu implementacije invertnega indeksa.

3 Implementacija invertnega indeksa

Implementacije invertnega indeksa smo se lotili v treh delih. V prvem delu smo najprej implementirali vse potrebne metode (C-Create, U-Update, R-Read, D-Delete) za manipulacijo z našo podatkovno bazo (*SQLite*). Nato smo dodali še tako imenovano poslovno logiko shranjevanja in polnjena *IndexWord*a ter *Posting*a. Pri čemer smo za potrebe shranjenvanja soseščine dodali še novo polje v tebelo *Posting*. Polje se imenuje *neighbourhood*.

Nato smo iz očiščenih dokumentov (predstavljeno v poglavju 2) napolnili tabeli (*IndexWord*, *Posting*). To smo storili tako, da smo se z zanko sprehodili preko vseh unikatnih besed, ki smo jih pridobili s čiščenjem dokumentov. Nato smo za vsako besedo vstavili pojavitve in soseščine. Nato smo vse te informacije shranili v invertni indeks - f: frekvenca ponovitve besede, w: beseda, i: indeksi besede v dokumentu ter n: soseščina besede.

4 Iskanje po invertnem indeksu

Kot smo že omenili, smo za potrebe iskanja po invertnem indeksu razširli podatkovno shemo *Posting* z atributom *neighbourhood*.

Spremenljivko *query* smo najprej očistili ter lematizirali po postopkih, kot je omenjeno v navodilu za tretji seminar. Nato smo za vsako besedo izvedli iskanje v invertnem indeksu ter vse rezultate razvrstili po frekvenčni oceni. Končne rezultate smo nato izpisali v lepši in formatirani obliki z izpisom soseščine (atr. *neighbourhood*). Povprečen čas izvajanja poizvedbe je 20ms. Rezultati so predstavljeni na slikah 1, 2, 3, 4, 5, 6.

5 Implementacija sekvenčnega iskanja

Sekvenčnega iskanja smo se lotili v dveh delih. Najprej smo se sprehodili preko vseh dokumentov in v teh dokumentih za vsako besedo v naši poizvedi (*query*) poiskali število teh pojavitev. V drugem delu smo nato te najdene frekvence razvrstili v frekvenčno padajočem vrstnem redu.

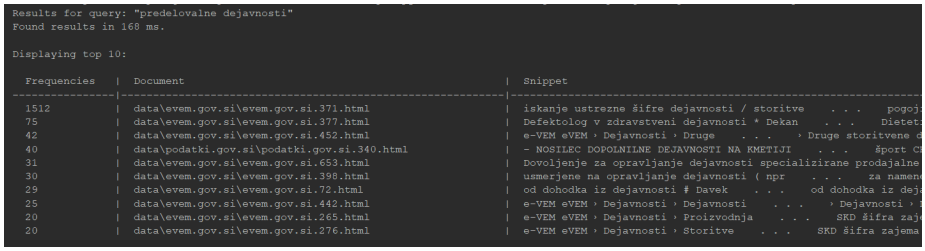
Povprečen čas iskanja z uporabo te metode je 64 sekund na besedo (poizvedba z dvema besedama traja okoli 128 sekund). Rezultati pozvedb so prikazani na slikah 7, 8, 9.

V splošnem je ta pristop terjal precej časa. Približno 3000 krat počasnejše je iskanje na ta način, kot na način iskanja z invertnim indeksom.

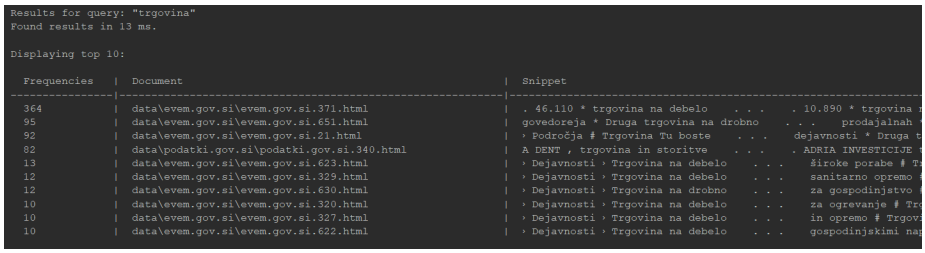
6 Zaključek

V seminarski nalogi smo se naučili ogromno novega znanja, predvsem na področju hitrega iskanja in delovanja invertnega indeksa. Zanimivo je bilo implementirati iskanje po invertnem indeksu in v splošnem čiščenje html dokumentov ter mapiranje dejanskih vrednosti z vrednostmi v invertnem indeksu. V splošnem smo z rezultati zadovoljni.

7 Zaslonske slike rezultatov

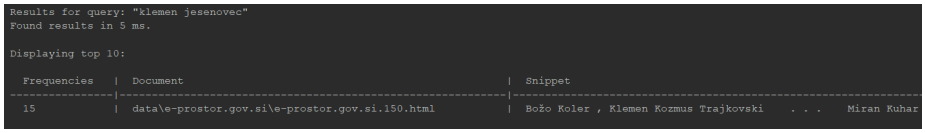


Slika 1: Poizvedba z invertnim indeksom: predelovalne dejavnosti.

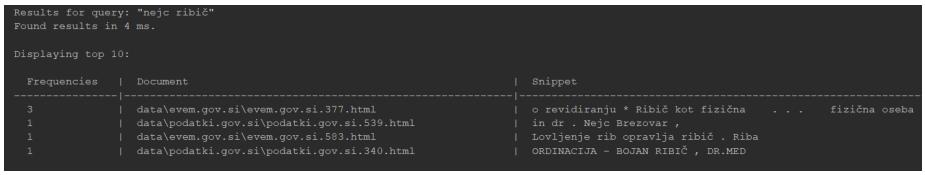


Slika 2: Poizvedba z invertnim indeksom: trgovina.

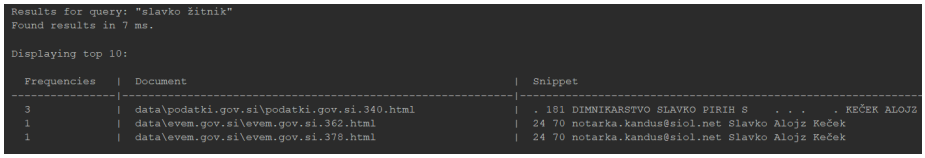
Slika 3: Poizvedba z invertnim indeksom: social services.



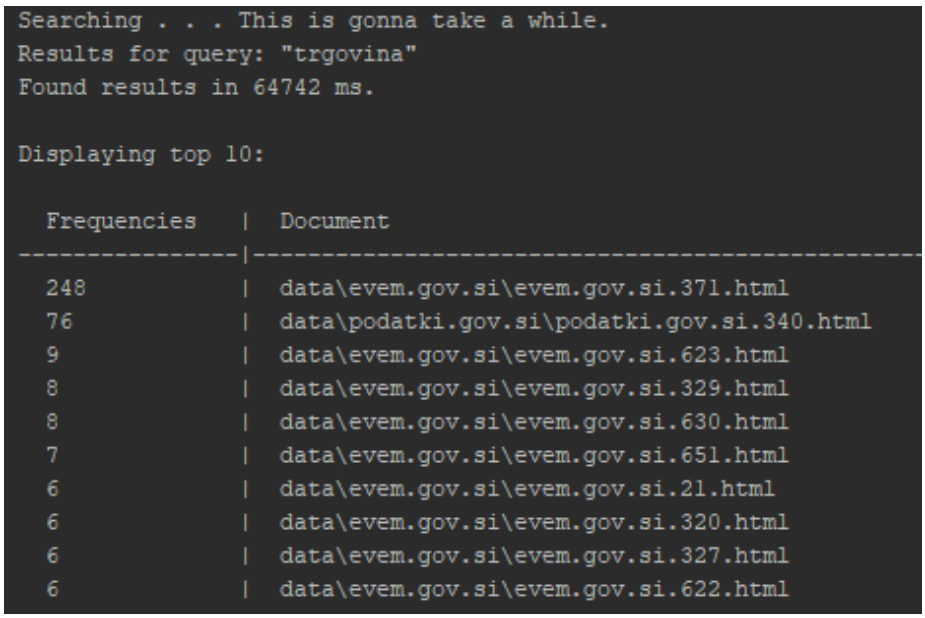
Slika 4: Poizvedba z invertnim indeksom: Klemen Jesenovec.



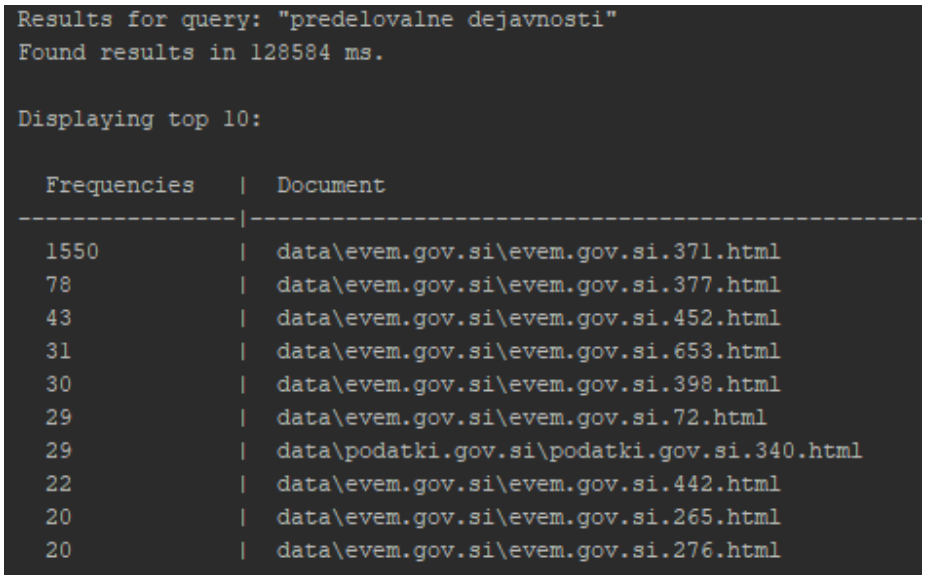
Slika 5: Poizvedba z invertnim indeksom: Nejc Ribič.



Slika 6: Poizvedba z invertnim indeksom: Slavko Žitnik.



Slika 7: Sekvenčna poizvedba: trgovina.

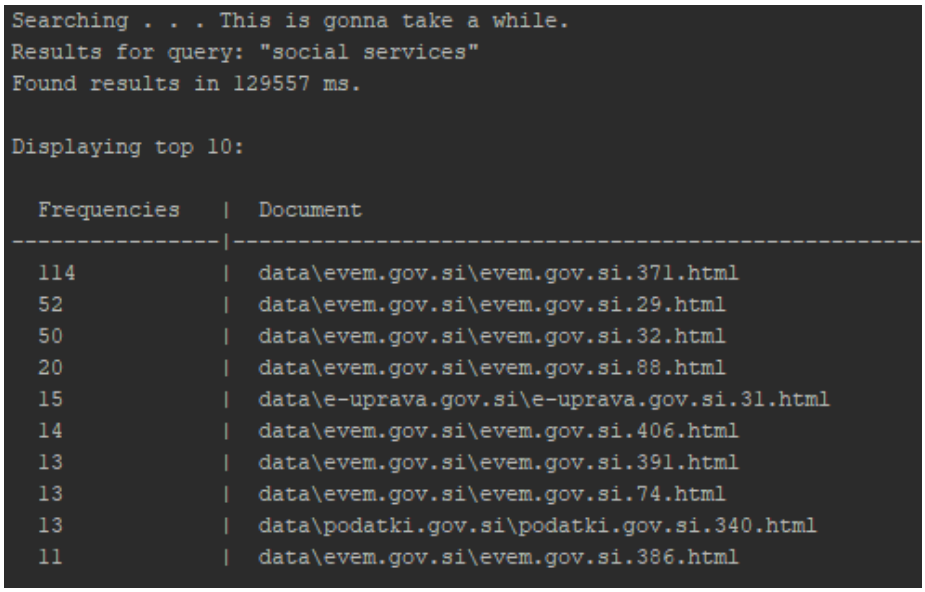


```
Results for query: "predelovalne dejavnosti"
Found results in 128584 ms.

Displaying top 10:

  Frequencies   |  Document
----------------|-------------------------------------------------
  1550          |  data\evem.gov.si\evem.gov.si.371.html
  78            |  data\evem.gov.si\evem.gov.si.377.html
  43            |  data\evem.gov.si\evem.gov.si.452.html
  31            |  data\evem.gov.si\evem.gov.si.653.html
  30            |  data\evem.gov.si\evem.gov.si.398.html
  29            |  data\evem.gov.si\evem.gov.si.72.html
  29            |  data\podatki.gov.si\podatki.gov.si.340.html
  22            |  data\evem.gov.si\evem.gov.si.442.html
  20            |  data\evem.gov.si\evem.gov.si.265.html
  20            |  data\evem.gov.si\evem.gov.si.276.html
```

Slika 8: Sekvenčna poizvedba: predelovalne storitve.



```
Searching . . . This is gonna take a while.
Results for query: "social services"
Found results in 129557 ms.

Displaying top 10:

  Frequencies   |  Document
----------------|-------------------------------------------------
  114           |  data\evem.gov.si\evem.gov.si.371.html
  52            |  data\evem.gov.si\evem.gov.si.29.html
  50            |  data\evem.gov.si\evem.gov.si.32.html
  20            |  data\evem.gov.si\evem.gov.si.88.html
  15            |  data\e-uprava.gov.si\e-uprava.gov.si.31.html
  14            |  data\evem.gov.si\evem.gov.si.406.html
  13            |  data\evem.gov.si\evem.gov.si.391.html
  13            |  data\evem.gov.si\evem.gov.si.74.html
  13            |  data\podatki.gov.si\podatki.gov.si.340.html
  11            |  data\evem.gov.si\evem.gov.si.386.html
```

Slika 9: Sekvenčna poizvedba: social services.

Literatura

[1] Programming assignment 3. Dosegljivo: http://zitnik.si/teaching/wier/PA3.html. [Dostopano: veliki traven, 2019].

[2] ribicnejc/web-data-extraction. Dosegljivo: https://github.com/ribicnejc/data$_{p}rocessing_{a}nd_{i}ndexing$. [$Dostopano : velikitraven$, 2019].