

# 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 poglavje 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 dodatno 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 prešli 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 polnjenja *IndexWorda* ter *Postinga*. Pri čemer smo za potrebe shranjevanja soseščine dodali še novo polje v tabelo *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 razširili 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 poizvedbi (*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 poizvedb 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

```
Results for query: "predelovalne dejavnosti"
Found results in 168 ms.

Displaying top 10:
```

Frequencies	Document	Snippet
1512	data\evem.gov.si\evem.gov.si.371.html	iskanje ustrezne šifre dejavnosti / storitve . . . pogoj
75	data\evem.gov.si\evem.gov.si.377.html	Defektolog v zdravstveni dejavnosti * Dekan . . . Dieteti
42	data\evem.gov.si\evem.gov.si.452.html	e-VEM eVEM * Dejavnosti * Druge . . . Druge storitvene d
40	data\podatki.gov.si\podatki.gov.si.340.html	- NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI . . . šport C
31	data\evem.gov.si\evem.gov.si.653.html	Dovoljenje za opravljanje dejavnosti specializirane prodajalne
30	data\evem.gov.si\evem.gov.si.398.html	usmerjene na opravljanje dejavnosti ( mpt . . . za namena
29	data\evem.gov.si\evem.gov.si.72.html	od dohodka iz dejavnosti # Davač . . . od dohodka iz dej
25	data\evem.gov.si\evem.gov.si.442.html	e-VEM eVEM * Dejavnosti * Dejavnosti . . . Dejavnosti *
20	data\evem.gov.si\evem.gov.si.265.html	e-VEM eVEM * Dejavnosti * Proizvodnja . . . SKD šifra zaj
20	data\evem.gov.si\evem.gov.si.276.html	e-VEM eVEM * Dejavnosti * Storitve . . . SKD šifra zajema

Slika 1: Poizvedba z invertnim indeksom: predelovalne dejavnosti.

```
Results for query: "trgovina"
Found results in 13 ms.

Displaying top 10:
```

Frequencies	Document	Snippet
364	data\evem.gov.si\evem.gov.si.371.html	. 46.110 * trgovina na debelo . . . 10.890 * trgovina na
95	data\evem.gov.si\evem.gov.si.651.html	govodoreja * Druga trgovina na drobno . . . prodajalnah *
92	data\evem.gov.si\evem.gov.si.21.html	* Področja # Trgovina Tu boste . . . dejavnosti * Druga t
82	data\podatki.gov.si\podatki.gov.si.340.html	A DENT , trgovina in storitve . . . ADRIA INVESTICIJE *
13	data\evem.gov.si\evem.gov.si.623.html	* Dejavnosti * Trgovina na debelo . . . široke porabe # Tr
12	data\evem.gov.si\evem.gov.si.325.html	* Dejavnosti * Trgovina na debelo . . . sanitarno opremo *
12	data\evem.gov.si\evem.gov.si.630.html	* Dejavnosti * Trgovina na drobno . . . za gospodinjstvo *
10	data\evem.gov.si\evem.gov.si.320.html	* Dejavnosti * Trgovina na debelo . . . za ogrevanje # Tr
10	data\evem.gov.si\evem.gov.si.327.html	* Dejavnosti * Trgovina na debelo . . . in opremo # Trgov
10	data\evem.gov.si\evem.gov.si.622.html	* Dejavnosti * Trgovina na debelo . . . gospodinjstvi n

Slika 2: Poizvedba z invertnim indeksom: trgovina.

```
Results for query: "social services"
Found results in 3 ms.

Displaying top 10:
```

Frequencies	Document	Snippet
3	data\ev-uprava.gov.si\ev-uprava.gov.si.45.html	, retirement * Social services , . . . ? # Social serv
3	data\ev-uprava.gov.si\ev-uprava.gov.si.9.html	, retirement * Social services , . . . ? # Social serv
2	data\ev-uprava.gov.si\ev-uprava.gov.si.45.html	retirement * Social services , health . . . # Social s
2	data\ev-uprava.gov.si\ev-uprava.gov.si.9.html	retirement * Social services , health . . . # Social s
1	data\evem.gov.si\evem.gov.si.661.html	Records and Related Services ( AJPES . . . # Social s
1	data\podatki.gov.si\podatki.gov.si.340.html	recreation and spa services ltd .

Slika 3: Poizvedba z invertnim indeksom: social services.

```
Results for query: "klemen jesenovec"
Found results in 5 ms.

Displaying top 10:
```

Frequencies	Document	Snippet
15	data\ev-prostor.gov.si\ev-prostor.gov.si.150.html	Božo Koler , Klemen Kozmus Trajkovski . . . Miran Ruhar

Slika 4: Poizvedba z invertnim indeksom: Klemen Jesenovec.

Results for query: "nejc ribič"  
Found results in 4 ms.

Displaying top 10:

Frequencies	Document	Snippet
3	data\evem.gov.si\evem.gov.si.377.html	o revidiranju * Ribič kot fizična . . . fizična oseba
1	data\podatki.gov.si\podatki.gov.si.539.html	in dr . Nejc Brezovar ,
1	data\evem.gov.si\evem.gov.si.583.html	Lovljenje rib opravlja ribič . Riba
1	data\podatki.gov.si\podatki.gov.si.340.html	ORDINACIJA - BOJAN RIBIČ , DR.MED

Slika 5: Poizvedba z invertnim indeksom: Nejc Ribič.

Results for query: "slavko žitnik"  
Found results in 7 ms.

Displaying top 10:

Frequencies	Document	Snippet
3	data\podatki.gov.si\podatki.gov.si.340.html	. 181 DIMNIKARSTVO SLAVKO PIRIH S . . . KEČEK ALOJZ
1	data\evem.gov.si\evem.gov.si.362.html	24 70 notarka.kandus@siol.net Slavko Alojz Keček
1	data\evem.gov.si\evem.gov.si.378.html	24 70 notarka.kandus@siol.net Slavko Alojz Keček

Slika 6: Poizvedba z invertnim indeksom: Slavko Žitnik.

Searching . . . This is gonna take a while.  
Results for query: "trgovina"  
Found results in 64742 ms.

Displaying top 10:

Frequencies	Document
248	data\evem.gov.si\evem.gov.si.371.html
76	data\podatki.gov.si\podatki.gov.si.340.html
9	data\evem.gov.si\evem.gov.si.623.html
8	data\evem.gov.si\evem.gov.si.329.html
8	data\evem.gov.si\evem.gov.si.630.html
7	data\evem.gov.si\evem.gov.si.651.html
6	data\evem.gov.si\evem.gov.si.21.html
6	data\evem.gov.si\evem.gov.si.320.html
6	data\evem.gov.si\evem.gov.si.327.html
6	data\evem.gov.si\evem.gov.si.622.html

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\\_processing\\_and\\_indexing](https://github.com/ribicnejc/data_processing_and_indexing). [*Dostopano* :  
*velikitraven*, 2019].