Obdelava, indeksiranje in pridobivanje podatkov

Anja Brelih
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Večna pot 113, 1000 Ljubljana
ab0555@student.uni-lj.si

1 Uvod

Pri tretjem projektnem delu pri predmetu Iskanje in ekstrakcija podatkov s spleta je bila naša naloga implementirati metode za obdelavo in indeksiranje podatkov spletnih strani ter nad podatki oziroma spletnimi stranmi izvajali poizvedbe.

2 Implementacija

Podane smo imeli *html* datoteke spletnih strani iz štirih različnih domen (*e-prostor.gov.si*, *e-uporava.gov.si*, *evem.gov.si* ter *podatki.gov.si*). V prvem koraku smo iz spletnih strani izdelali obrnjeni indeks, ki smo ga shranili v podatkovno bazo. V drugem koraku smo izvajali poizvedbe z uporabo obrnjenega indeksa. V tretjem koraku smo za primerjavo izvajali poizvedbe brez uporabe obrnjenega indeksa – v tem primeru smo ob vsaki poizvedbi pregledali vse dokumente z uporabo enakih postopkov kot smo jih implementirali pri prvem koraku.

2.1 Obdelava in indeksiranje podatkov

Pri obdelavi in indeksiranju podatkov smo v prvem koraku kreirali podatkovno bazo, v kolikor ta še ni obstajala. V naslednjem koraku smo zagnali zanko, ki je prebrala vsako datoteko ter jo s knjižnico *BeautifulSoup* zapisala v *lxml* format. Zanka je spodaj opisane postopke izvedla nad vsemi *html* datotekami.

Prebrano datoteko smo zapisali z malimi tiskanimi črkami in s knjižnico *nltk* pridobili žetone iz vsebine datoteke. Žetoni predstavljajo besede, ki bodo zapisane v naš obrnjeni indeks, vendar pa jih je pred zapisom v podatkovno bazo potrebno še prečistiti.

Vsak pridobljen žeton smo primerjali s seznamom irelevantnih slovenskih in angleških besed in ga, v kolikor beseda obstaja v enem od teh seznamov, odstranili. Preverili smo tudi, da sam žeton ne predstavlja ločilo. Ker smo opazili, da se še vedno pojavljajo besede oziroma znaki, ki ne sodijo v indeks, smo ustvarili tudi seznam nedovoljenih besed in znakov, ki se jih odstrani iz pridobljenega seznama žetonov.

Nad prečiščenim seznamom žetonov smo nato zagnali zanko, ki je za vsak unikaten žeton preštela ponovitve ter zapisala položaj le-teh. Vsak unikaten žeton smo nato zapisali v podatkovno bazo skupaj z informacijami o imenu dokumenta, številom ponovitev žetona ter njihovih položajih v seznamu prečiščenih žetonov.

Ko smo izvedli postopek nad vsemi datotekami, je bil naš obrnjeni indeks zaključen. V podatkovni bazi se nahaja *49.023* unikatnih besed, s *379.209* vnosi z informacijami o dokumentu, ponavljanju ter položajih oziroma indeksu. Beseda, z največjo frekvenco v enem dokumentu je »proizvodnja« s *2.266* ponovitvami.

Ob implementaciji poizvedbe z obrnjenim indeksom, pri kreiranju odrezkov besedila, smo ugotovili, da zapisovanje datoteke v *lxml* format s knjižnico *BeautifulSoup* (1)

```
lxml=BeautifulSoup(file.read(), 'lxml')
```
(1)

ni bila najboljša izbira. Pridobili vsebino datoteke skupaj z *html* značkami in njihovimi atributi. Večina teh se je pri kreiranju in čiščenju žetonov odpravila, vendar vseeno obstaja kakšen vnos v podatkovno bazo, ki je odveč. Rezultat v podatkovni bazi je seveda uporaben, vendar lahko z manjšim popravkom iz datoteke *html* pridobimo le vsebino (2)

```
text=BeautifulSoup(file.read().getText(separator=" ")
```
(2).

Pri izdelavi odrezkov besedila smo tako uporabili metodo, ki iz *html* datoteke pridobi le vsebino, da smo si olajšali delo oziroma to sploh omogočili v zgledni obliki.

2.2 Poizvedba z obrnjenim indeksom

Pri poizvedbi z obrnjenim indeksom se, glede na vneseno poljubno število besed v iskalni niz, v podatkovni bazi poišče vnose, ki vsebujejo vsaj eno od teh besed. Rezultat se razvrsti po padajočem vrstnem redu skupnih ponovitev iskanih besed v dokumentu (seštevek ponovitev vseh besed iz iskalnega niza v dokumentu).

Prvih nekaj rezultatov smo nato izpisali skupaj z odrezki besedila. Za namen izdelave odrezkov smo dokument še enkrat prebrali ter iz *html* datoteke prebrali le besedilo. Nato smo iz besedila še enkrat pridobili žetone – tokrat brez zapisovanja besedila z malimi tiskanimi črkami ter čiščenja žetonov. Ta korak je bil potreben, če smo želeli pridobiti smiselne odrezke besedila, čeprav to pomeni, da so shranjeni indeksi besed v besedilu neuporabljeni. V kolikor bi za kreiranje odrezkov uporabili shranjene indekse ter žetone z enako obdelavo, kot pri zapisu v podatkovno bazo, bi bili odrezki zapisani z malimi tiskanimi črkami, ne bi vsebovali ločil, veznih členov, itd.

Za vsako besedo v poizvedbi, z izjemo irelevantnih slovenskih besed (npr. vezni členi), smo poiskali indekse v seznamu žetonov ter nato prvih nekaj odrezkov oblikovali tako, da smo zaobjeli iskano besedo s tremi žetoni pred ter tremi žetoni po besedi, v kolikor je to možno glede na položaj iskane besede. V kolikor se je ta

nahajala na ali blizu začetka oziroma konca, smo za odrezek vzeli primerno število besed manj.

V samih odrezkih so se pojavile manjše napake – dvoji presledki ali presledki okoli ločil, ki smo jih odstranili oziroma popravili.

Na koncu smo rezultat poizvedbe zapisali na standardni izhod v predpisanem formatu skupaj s časom, ki je bil potreben za poizvedbo ter formatiranje odrezkov besedila.

2.3 Poizvedba brez obrnjenega indeksa

Pri poizvedbi brez obrnjenega indeksa smo združili postopke opisane poglavjih 2.1 ter 2.2, z izjemo pisanja in poizvedovanja v podatkovno bazo. Tako smo pri vsaki poizvedbi posebej pregledali vse dokumente ter si shranjevali število ponovitev besed za vsak dokument.

Za namen pridobitve identičnih rezultatov poizvedb kot pri obrnjenem indeksu, smo uporabili enak postopek za zapisovanje datoteke v *lxml* format s knjižnico *BeautifulSoup,* kot pri obdelavi in indeksiranju podatkov. Če bi uporabili metodo, ki iz *html* datoteke pridobi le vsebino ter shranjevali žetone kot spremenljivko, bi se lahko izognili ponovnemu branju datoteke pri izdelavi odrezkov besedila.

3 Poizvedbe

Izvedli smo šest različnih poizvedb z obrnjenim indeksom in brez obrnjenega indeksa. Rezultati pri obeh metodah so identični z izjemo časa poizvedbe ter drugega vrstnega reda dokumentov z enakimi frekvencami (iz podatkovne baze rezultat znotraj enakih frekvenc razporedimo po imenu dokumenta).

Pridobili smo rezultate za poizvedbe »predelovalne dejavnosti« (Priloga 1 in 2), »trgovina« (Priloga 3 in 4), »social services« (Priloga 5 in 6), »državni organi« (Priloga 7 in 8), »program« (Priloga 9 in 10) ter »Vloga za prijavo začasnega prebivališča« (Priloga 11 in 12). V prilogah je, zaradi preglednosti, prikazan le en odrezek besedila na dokument.

Število izpisanih rezultatov ter število odrezkov besedila se nastavlja parametrično na začetku programske kode. V kolikor je število rezultatov ali odrezkov manj od nastavljene vrednosti, se jih izpiše toliko, kot jih obstaja.

Sami odrezki besedila občasno še vključujejo kakšno napako pri formatiranju. Skozi testiranje smo poskušali najti in odpraviti napake, vendar še vedno napake niso izključljive.

V povprečju so poizvedbe brez obrnjenega indeksa potrebovale 165*x* več časa kot poizvedbe z obrnjenim indeksom. Največji faktor razlike med enakima poizvedbama je 470, najmanjši pa 28. Čas, ki smo ga porabili pri poizvedbah brez obrnjenega indeksa se ne glede na iskani niz giblje med 52 in 56 sekundami.

4 Zaključek

Nekatere pomanjkljivosti, ki smo jih zaznali pri implementaciji procesiranja, indeksiranja ter pridobivanja podatkov so opisane že skozi same razlage uporabljenih postopkov. Manjša izboljšava je možna pri pridobivanju vsebine iz *html* datoteke; pri oblikovanju odrezkov besedila je možno implementirati bolj robustno rešitev, čeprav tudi predstavljena rešitev daje primerne rezultate.

V splošnem implementirane rešitve delujejo kot je zastavljeno. Pri pregledu rezultatov je predvsem zanimiva časovna primerjava med eno in drugo poizvedbo.

Literatura

[1] NLTK. Documentation. https://www.nltk.org/. Dostopano: maj 2022

[2] GeeksforGeeks. Python – Compute the frequency of words after removing stop words and stemming. https://www.geeksforgeeks.org/python-compute-the-frequency-of-words-after-removing-stop-words-and-stemming/. Dostopano: maj 2022

[3] Rabexc. A simple way to generate snippets in python. https://rabexc.org/posts/html-snippets-in-python. Dostopano: maj 2022

[4] Stackoverflow. Best way to strip punctuation from a string. https://stackoverflow.com/questions/265960/best-way-to-strip-punctuation-from-a-string. Dostopano: maj 2022

[5] GeeksforGeeks. Removing stop words with NLTK in Python. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/. Dostopano: maj 2022

[6] Python Tutorials. NLTK stop words. https://pythonspot.com/nltk-stop-words/. Dostopano: maj 2022

Priloge

Priloga 1: Poizvedba »predelovalne dejavnosti« z obrnjenim indeksom

Results for a query: "predelovalne dejavnosti"

753 results found in 1972.66 ms.

Frequencies	Document	Snippet
1288	evem.gov.si/evem.gov.si.371.html	... za infrastrukturo C PREDELOVALNE DEJAVNOSTI 10 ...
75	evem.gov.si/evem.gov.si.377.html	... Defektolog v zdravstveni dejavnosti Dekan oziroma ...
40	podatki.gov.si/podatki.gov.si.340.html	... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI ...
39	evem.gov.si/evem.gov.si.452.html	... Druge storitvene dejavnosti, drugje ...
31	evem.gov.si/evem.gov.si.653.html	... Dovoljenje za opravljanje dejavnosti specializirane prodajalne ...

Priloga 2: Poizvedba »predelovalne dejavnosti« brez obrnjenega indeksa

Results for a query: "predelovalne dejavnosti"

753 results found in 55108.34 ms.

Frequencies	Document	Snippet
1288	evem.gov.si/evem.gov.si.371.html	... za infrastrukturo C PREDELOVALNE DEJAVNOSTI 10 ...
75	evem.gov.si/evem.gov.si.377.html	... Defektolog v zdravstveni dejavnosti Dekan oziroma ...
40	podatki.gov.si/podatki.gov.si.340.html	... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI ...
39	evem.gov.si/evem.gov.si.452.html	... Druge storitvene dejavnosti, drugje ...
31	evem.gov.si/evem.gov.si.653.html	... Dovoljenje za opravljanje dejavnosti specializirane prodajalne ...

Priloga 3: Poizvedba »trgovina« z obrnjenim indeksom

Results for a query: "trgovina"

125 results found in 1912.16 ms.

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	... gl. 46.110 trgovina na debelo ...
94	evem.gov.si/evem.gov.si.651.html	... Druga govedoreja Druga trgovina na drobno ...
92	evem.gov.si/evem.gov.si.21.html	... eVEM › Področja Trgovina Tu boste ...
82	podatki.gov.si/podatki.gov.si.340.html	... A DENT, trgovina in storitve ...
13	evem.gov.si/evem.gov.si.623.html	Trgovina na debelo ...

Priloga 4: Poizvedba »trgovina« brez obrnjenega indeksa

Results for a query: "trgovina"

125 results found in 54277.06 ms.

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	... gl. 46.110 trgovina na debelo ...
94	evem.gov.si/evem.gov.si.651.html	... Druga govedoreja Druga trgovina na drobno ...
92	evem.gov.si/evem.gov.si.21.html	... eVEM › Področja Trgovina Tu boste ...
82	podatki.gov.si/podatki.gov.si.340.html	... A DENT, trgovina in storitve ...
13	evem.gov.si/evem.gov.si.623.html	Trgovina na debelo ...

Priloga 5: Poizvedba »social services« z obrnjenim indeksom

Results for a query: "social services"

4 results found in 506.69 ms.

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.9.html	... Labour, retirement Social services , ...
5	e-uprava.gov.si/e-uprava.gov.si.45.html	... Labour, retirement Social services , ...
1	podatki.gov.si/podatki.gov.si.340.html	... recreation and spa services ltd. TERME ...
1	evem.gov.si/evem.gov.si.661.html	... Records and Related Services (AJPES ...

Priloga 6: Poizvedba »social services« brez obrnjenega indeksa

Results for a query: "social services"

4 results found in 53233.95 ms.

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.45.html	... Labour, retirement Social services , ...
5	e-uprava.gov.si/e-uprava.gov.si.9.html	... Labour, retirement Social services , ...
1	evem.gov.si/evem.gov.si.661.html	... Records and Related Services (AJPES ...
1	podatki.gov.si/podatki.gov.si.340.html	... recreation and spa services ltd. TERME ...

Priloga 7: Poizvedba »državni organi« z obrnjenim indeksom

Results for a query: "državni organi"

1248 results found in 163.22 ms.

Frequencies	Document	Snippet
42	podatki.gov.si/podatki.gov.si.106.html	DRŽAVNI ZBOR REPUBLIKE ...
34	podatki.gov.si/podatki.gov.si.373.html	... organizacije. Organizacija DRŽAVNI ZBOR REPUBLIKE ...
34	podatki.gov.si/podatki.gov.si.23.html	... organizacije. Organizacija DRŽAVNI ZBOR REPUBLIKE ...
32	podatki.gov.si/podatki.gov.si.111.html	... manj Tip organizacije Državni organi ...
31	podatki.gov.si/podatki.gov.si.367.html	... Tip organizacije Državni organi ...

Priloga 8: Poizvedba »državni organi« brez obrnjenega indeksa

Results for a query: "državni organi"

1248 results found in 52651.21 ms.

Frequencies	Document	Snippet
42	podatki.gov.si/podatki.gov.si.106.html	DRŽAVNI ZBOR REPUBLIKE ...
34	podatki.gov.si/podatki.gov.si.23.html	... organizacije. Organizacija DRŽAVNI ZBOR REPUBLIKE ...
34	podatki.gov.si/podatki.gov.si.373.html	... organizacije. Organizacija DRŽAVNI ZBOR REPUBLIKE ...
32	podatki.gov.si/podatki.gov.si.111.html	... manj Tip organizacije Državni organi ...
31	podatki.gov.si/podatki.gov.si.367.html	... Tip organizacije Državni organi ...

Priloga 9: Poizvedba »program« z obrnjenim indeksom

Results for a query: "program"

144 results found in 1453.74 ms.

Frequencies	Document	Snippet
14	evem.gov.si/evem.gov.si.371.html	... 1) hranijo program ali programe ...
13	e-prostor.gov.si/e-prostor.gov.si.2.html	... slovenskem jeziku SPLETNI PROGRAM SITRIK PROGRAM ...
13	e-prostor.gov.si/e-prostor.gov.si.33.html	... slovenskem jeziku SPLETNI PROGRAM SITRIK PROGRAM ...
11	e-prostor.gov.si/e-prostor.gov.si.57.html	... v PREG in program od mene ...
6	e-prostor.gov.si/e-prostor.gov.si.110.html	... Aplikacije SPLETNI PROGRAM SITRIK je ...

Priloga 10: Poizvedba »program« brez obrnjenega indeksa

Results for a query: "program"

144 results found in 53766.1 ms.

Frequencies	Document	Snippet
14	evem.gov.si/evem.gov.si.371.html	... 1) hranijo program ali programe ...
13	e-prostor.gov.si/e-prostor.gov.si.2.html	... slovenskem jeziku SPLETNI PROGRAM SITRIK PROGRAM ...
13	e-prostor.gov.si/e-prostor.gov.si.33.html	... slovenskem jeziku SPLETNI PROGRAM SITRIK PROGRAM ...
11	e-prostor.gov.si/e-prostor.gov.si.57.html	... v PREG in program od mene ...
6	e-prostor.gov.si/e-prostor.gov.si.110.html	... Aplikacije SPLETNI PROGRAM SITRIK je ...

Priloga 11: Poizvedba »Vloga za prijavo začasnega prebivališča« z obrnjenim indeksom

Results for a query: "Vloga za prijavo začasnega prebivališča"

190 results found in 112.08 ms.

Frequencies	Document	Snippet
36	e-uprava.gov.si/e-uprava.gov.si.58.html	... začasno prebivališče. Vloga za prijavo ...
13	evem.gov.si/evem.gov.si.398.html	... , sicer se vloga šteje za ...
9	evem.gov.si/evem.gov.si.88.html	... če gre za prijavo delavca , ...
8	podatki.gov.si/podatki.gov.si.175.html	... objekta, državi prebivališča in uporabi ...
7	evem.gov.si/evem.gov.si.406.html	... kot zavezanec za prijavo (delodajalec ...

Priloga 12: Poizvedba »Vloga za prijavo začasnega prebivališča« brez obrnjenega indeksa

Results for a query: "Vloga za prijavo začasnega prebivališča"

190 results found in 52770.15 ms.

Frequencies	Document	Snippet
36	e-uprava.gov.si/e-uprava.gov.si.58.html	... začasno prebivališče. Vloga za prijavo ...
13	evem.gov.si/evem.gov.si.398.html	... , sicer se vloga šteje za ...
9	evem.gov.si/evem.gov.si.88.html	... če gre za prijavo delavca , ...
8	podatki.gov.si/podatki.gov.si.175.html	... objekta, državi prebivališča in uporabi ...
7	e-uprava.gov.si/e-uprava.gov.si.37.html	... iz kazenskih evidenc. Vloga za pridobitev ...