Procesiranje, indeksiranje in iskanje po podatkih

Tretja naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

Žiga Kleine
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Email: zk5343@student.uni-lj.si

Tim Križnik
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Email: tk0730@student.uni-lj.si

Luka Železnik
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani
Email: lz6356@student.uni-lj.si

I. Povzetek

V tem poročilu bomo predstavili še zadnjo nalogo pri predmetu Iskanje in ekstrakcija podatkov s spleta. Pri tej nalogi smo tekst iz vnaprej podanih 1416 spletnih strani najprej predhodno procesirali, nato zgradili invertiran indeks s pomočjo SQLite podatkovne baze, na koncu pa še implementirali iskanje po našem invertiranem indeksu. Da smo lahko ovrednotili učinkovitost invertiranega indeksa, smo implementirali še iskanje po dokumentih brez uporabe invertiranega indeksa. Rezultate poizvedb na obeh implementacijah iskanja bomo ustrezno primerjali in predstavili.

II. Uvod

Invertirani indeks je podatkovna struktura s pomočjo katere lahko pospešimo iskanje po besedilu v večjem korpusu dokumentov. Deluje tako, da mapira besede v točne lokacije teh besed v izvornih besedilih.

Da lahko z invertiranim indeksom uspešno indeksiramo večje korpuse dokumentov, jih moramo ustrezno predhodno procesirati. Metode predhodnega procesiranja, uporabljene za potrebe naše implementacije invertiranega indeksa so pretvorba besedil v majhne črke, odstranjevanje nerelevantnih besed (angleško *stopwords*) in tokenizacija besedila.

III. Implementacija

A. Predhodno procesiranje in indeksiranje

Datoteke tipa HTML program *indexing.py* prebere eno za drugo in jo s razščlenjevalno knjižnico *BeautifulSoup* pretvori v objekt soup. Elementi tipa *script* in *style* so takoj na začetku s klicem metode *extract()* izločeni iz tega objekta. Hkrati tudi elementi tipa *Comment*, saj ne želimo indeksirati komentarjev v kodi. Iz objekta *soup* pridobimo ves tekst z ukazom: *soup.body.get_text(separator=' ')* . Ta vrne ločeno besedilo iz različnih HTML oznak, ki je pripravljeno na predprocesiranje.

Prvi korak predprocesiranja je tokenizacija pridobljenega besedila s funkcijo *word_tokenize()* knjižnice za obdelavo naravnega jezika *nltk.tokenize*. Rezultat je tabela blokov (ang. *tokens*) - med sabo ločene besede, števila in simboli na trenutno obravnavani strani. Nize v tej tabeli pretvorimo v nize z malimi črkami s *string.lower()* - obdelava besed poteka le na nizih z malimi črkami.

Pomemben korak na tem mestu je dodelitev indeksa besede v tej tabeli. Podatkovna struktura se spremeni v tabelo Python setov, npr. [('ta', 0), ('stran', 1), ('uporablja', 2), ('piškotke', 3) ...].

Ker želimo ohraniti pomensko bogate besede seznam še dodatno filtriramo z izločitvijo nerelevantnih slovenskih besed, ki se nahajajo v besednem korpusu paketa *nltk.corpus*. Korpus smo podaljšali še s seznamom besed v priloženi datoteki *stopwords.py* in sprotnimi abnormalijami (npr. niz *'x'*). Z regularnim izrazom izločimo še nize, ki vsebujejo številke *(!bool(re.search(r'', niz))* in nize simbolov *(bool(re.match('([a-zA-ZčšžćČŠŽĆ]+)', niz))*. Ostanejo le pomensko relevantne besede skupaj z njihovimi indeksi.

Skozi podatkovno strukturo besed se pred pisanjem v bazo sprehodimo z namenom, da preštejemo frekvence ponovitev. Seznam indeksov, kjer se beseda ponovi ustrezno daljšamo in struktura dobi obliko, npr. *[['uporablja', [2, 45]], ['piškotke', [3]], ...]*.

```
e-prostor.gov.si.12.html
[['uporablja', ['2', '511', '540', '574', '595', '628']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '1
e-prostor.gov.si.120.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '117', '155', '227', '305', '483', '5
e-prostor.gov.si.121.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '1626']], ['strani', ['9', '650', '78
e-prostor.gov.si.122.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '1126']], ['strani', ['9', '749']], [
e-prostor.gov.si.123.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '241']], ['strani', ['9']], ['soglaša
e-prostor.gov.si.124.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '646']], ['strani', ['9']], ['soglaša
e-prostor.gov.si.125.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '904']], ['strani', ['9']], ['soglaša
e-prostor.gov.si.126.html
[['uporablja', ['2']], ['piškotke', ['3']], ['nadaljevanjem', ['6']], ['uporabe', ['7', '415']], ['strani', ['9']], ['soglaša
```

Fig. 1. Primeri podatkovne strukture z besedami in indeksi ponovitev.

Podatkovno bazo SQLite (s knjižnico sqlite3) napolnimo z zbranimi podatki. Vstavljanje v *IndexWord* poteka s poizvedbo *'INSERT OR IGNORE INTO IndexWord(word) VALUES (?)'*, ki ustavi vpisovanje istih besed najdenih na različnih straneh.*IndexWord* je torej seznam vseh indeksiranih besed (ang. *vocabulary*). Tabela *Posting* ima zapis nanašujoč se na specifične strani: *(beseda, nazivDokumenta, frekvenca, indeksi)*.

B. Iskanje po podatkih z invertiranim indeksom

Iskanje z invertiranim indeksom poteka z obrnjenim seznamom - s seznama terminov, ki so v slovarju, poiščemo dokumente v katerih se nahaja in dodatne metapodatke/indekse.

Program *run-sqlite-search.py* kot vhod sprejme niz, ki ga predobdela, tako kot pridobljen tekst s spletnih strani v fazi predhodnega procesiranja - tokenizacija, pretvorba v male črke in izločitev nerelevantnih besed. Tuple nizov iz poizvedbe posredujemo metodi, ki dostopa do zapisa *Posting* v podatkovni bazi *(fig 2)*.

```
def search_Postings(query):
    c = conn.cursor()
    sql_query = '''
        SELECT p.documentName AS docName, SUM(frequency) AS freq, GROUP_CONCAT(indexes) AS idxs
        FROM Posting p
        WHERE
            p.word IN ({seqe})
        GROUP BY p.documentName
        ORDER BY freq DESC;
    '''.format(
        seqe=','.join(['?'] * len(query)))
    data_to_insert = query


    try:
        res = c.execute(sql_query, data_to_insert)
        conn.commit()
        return res
    except sqlite3.Error as error:
        print("Error Search: " + error.args[0])
    return
```

Fig. 2. Metoda za iskanje v SQLite podatkovni bazi.

Iz vrnjenih zadetkov v rezultatu preberemo naziv dokumenta, št. pojavitev in indekse (položaj pojavitev). Metodi *get_snipet()* posredujemo te podatke, da odpre dokument zapisan v *Posting-u*, in ga obdela popolnoma enako kot v fazi predprocesiranja: odstrani oznake *script, style, Comment,* pridobi besedilo z *soup.get_text* in tokeniziraj z enako metodo. Indeksi so posledično enaki, kot so bili pridobljeni v fazi predprocesiranja in se uporabijo za izpis okolice besede (ang. snippet) - okolica je pridobljena iz pravkar ustvarjene tabele. Rezultati večbesednih poizvedb so združeni po dokumentih.

C. Iskanje po podatkih brez invertiranega indeksa

Za implementacijo iskalnika brez invertiranega indeska smo morali spremeniti in dodelati nekatere elemente prejšnih dveh implementacij. Ker v tem primeru ne uporabljamo baze, smo lahko ta del povsem izpustili. Namesto v bazo, smo podatke shranjevali v lokalno spremenljivko za vsako spletno stran. Ta del je zato pohitril shranjevanje predprocesiranja in indeksiranja podatkov, ampak posledično deluje le za eno stran naenkrat.

IV. Rezultati

A. Podatkovna baza

Podatkovna baza ima v tabeli *IndexWord* 33370 unikatnih zapisov / besed. V tabeli *Posting* pa 346846 zapisov. Najpogostejše besede so: *proizvodnja, gl, spada, dejavnosti, d.o.o, skupnost, krajevna, ministrstvo, šola, dejavnost, vir* (*fig* 3)... Najpogostejši dokumenti so: *evem.gov.si.371.html, podatki.gov.si.340.html, e-prostor.gov.si.57.html...*(*fig* 4).

B. Poizvedbe po podatkih z invertiranim indeksom

V priloženih posnetkih zaslona (fig. 5- 10) so rezultati poizvedb za besede “predelovalne dejavnosti”, “trgovina”, “social services”, ”hackaton”, ”izredni študent” in ”vzdrževanje stiskalnic za iverne plošče”. Z uporabo invertiranega indeksa iskanje ene besede traja ponavadi nekaj milisekund (maks.

	word	documentName	frequency ▲1
	Filter	Filter	Filter
1	proizvodnja	evem.gov.si/evem.gov.si.371.html	2266
2	gl	evem.gov.si/evem.gov.si.371.html	1668
3	spada	evem.gov.si/evem.gov.si.371.html	1338
4	dejavnosti	evem.gov.si/evem.gov.si.371.html	1287
5	d.o.o	podatki.gov.si/podatki.gov.si.340.html	967
6	xsd	e-prostor.gov.si/e-prostor.gov.si....	927
7	skupnost	podatki.gov.si/podatki.gov.si.340.html	809
8	krajevna	podatki.gov.si/podatki.gov.si.340.html	754
9	ministrstvo	evem.gov.si/evem.gov.si.371.html	589
10	šola	podatki.gov.si/podatki.gov.si.340.html	582
11	dejavnost	evem.gov.si/evem.gov.si.371.html	545
12	vir	evem.gov.si/evem.gov.si.371.html	527
13	om	e-prostor.gov.si/e-prostor.gov.si....	511
14	ipd	evem.gov.si/evem.gov.si.371.html	488
15	družina	podatki.gov.si/podatki.gov.si.340.html	475

Fig. 3. Najpogostejše besede v podatkovni bazi.

```
SELECT      `documentName`,
            COUNT(`documentName`) AS `freq`
    FROM    `Posting`
    GROUP BY `documentName`
    ORDER BY `freq` DESC
    LIMIT   10;
```

	documentName	freq
1	evem.gov.si/evem.gov.si.371.html	12245
2	podatki.gov.si/podatki.gov.si.340.html	6507
3	e-prostor.gov.si/e-prostor.gov.si....	1620
4	evem.gov.si/evem.gov.si.398.html	1481
5	evem.gov.si/evem.gov.si.651.html	1265
6	e-uprava.gov.si/e-uprava.gov.si....	1159
7	evem.gov.si/evem.gov.si.653.html	1007
8	e-uprava.gov.si/e-uprava.gov.si....	964
9	evem.gov.si/evem.gov.si.377.html	925
10	e-prostor.gov.si/e-prostor.gov.si....	901

Fig. 4. Najpogostejši dokumenti v podatkovni bazi.

pribl. 10ms). V redkih primerih se tudi zgodi da zaradi zagona baze ali inicializiranja, iskanje doseže okoli 100ms. Iskanje ene besede v vseh datotekah se meri v milisekundah in ne sekundah, kot je opisano v nadaljevanju v *razdelku C*.
Izpis programa se za iskani niz izpiše združeno po dokumentih in s frekvenco vseh besed v poizvdebi. Poizvedbe z besedami, ki se tematsko nanašajo na spletne strani prikažejo zelo veliko rezultatov (npr. beseda *dejavnosti* se pojavi v zelo veliko dokumentih). Dodatno bi lahko izboljšali še izpis okolice iskanih besed (+3/-3 besede).

C. Poizvedbe po podatkih brez invertiranega indeksa

Na naši napravi je indeskiranje in iskanje določene ene besede vzelo od 100 do 300 milisekund. Večje število iskanih besed tudi vpliva na ta čas, najbolj pa je pomembna dolžina teksta strani in posledično število različnih besed. Iskanje ene besede v vseh datotekah je skupno vzelo okoli 260 sekund.

Results for a query: "predelovalne dejavnosti"
Results found in 9.012093127ms.

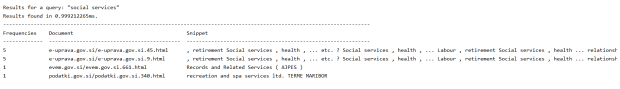
Frequencies	Document	Snippet
1291	evem.gov.si/evem.gov.si.371.html	iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti . V iskalnik ... 045 od 045 dejavnosti I
75	evem.gov.si/evem.gov.si.377.html	Defektolog v zdravstveni dejavnosti Oskar ocirema direktor ... Dietetik v zdravstveni dejavnosti Olimikar Diplomirana medicinska ...] v
60	podatki.gov.si/podatki.gov.si.360.html	- NOSILEC DOPOLNILNE DEJAVNOSTI NA KRETIZI BREGAR ... Sport CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER JUDOVSKE ... ŠOLSKIH IN OBŠOLSKIH D
58	evem.gov.si/evem.gov.si.452.html	e-VEM eVEM › Dejavnosti › Druge storitvene ... › Druge storitvene dejavnosti , drugje nerazvrščene ...] Druge storitvene dejavnosti , d
51	evem.gov.si/evem.gov.si.453.html	Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske ali televizijske dejavnosti Dovoljenje za izvajanje ... za
30	evem.gov.si/evem.gov.si.398.html	usmerjene na opravljanje dejavnosti (npr . ... za namene opravljanja dejavnosti ipd . V ... mesecev z opravljanjem dejavnosti v Sloveni
28	evem.gov.si/evem.gov.si.72.html	od dohodka iz dejavnosti Davek od dohodka ... od dohodka iz dejavnosti Ko začnete opravljati ... od dohodka iz dejavnosti . Za dohodek .
23	evem.gov.si/evem.gov.si.442.html	e-VEM eVEM › Dejavnosti › Dejavnosti za ... › Dejavnosti › Dejavnosti za nego telesa ... (96.040) Dejavnosti za nego telesa ... SKD šif
18	evem.gov.si/evem.gov.si.28.html	za opravljanje gospodarske dejavnosti . Lastnosti zasebnega ... niso za posamezne dejavnosti ali posamezne vrste ... ustanovitev in opra
17	evem.gov.si/evem.gov.si.265.html	e-VEM eVEM › Dejavnosti › Proizvodnja mesa ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpis
17	evem.gov.si/evem.gov.si.460.html	e-VEM eVEM › Dejavnosti › Druge nerazvrščene ... Druge nerazvrščene predelovalne dejavnosti (32.990) ... Druge nerazvrščene predelo
16	evem.gov.si/evem.gov.si.266.html	e-VEM eVEM › Dejavnosti › Proizvodnja mesnih ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predp
16	evem.gov.si/evem.gov.si.272.html	e-VEM eVEM › Dejavnosti › Storitve za ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpisi in
16	evem.gov.si/evem.gov.si.276.html	e-VEM eVEM › Dejavnosti › Storitve za ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpisi in
16	evem.gov.si/evem.gov.si.392.html	. 6 . Dejavnosti podjetja Ob ustanovitvi ...] in druge dejavnosti , ki jih ... opravljati . Število dejavnosti , ki jih ... ni omejeno
16	evem.gov.si/evem.gov.si.42.html	› Načini opravljanja dejavnosti Načini opravljanja dejavnosti ... dejavnosti Načini opravljanja dejavnosti Najpogostejši način za ... za
16	evem.gov.si/evem.gov.si.564.html	e-VEM eVEM › Dejavnosti › Proizvodnja sadnih ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predp
16	evem.gov.si/evem.gov.si.574.html	e-VEM eVEM › Dejavnosti › Zagarje , ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpisi in ...
15	evem.gov.si/evem.gov.si.140.html	e-VEM eVEM › Dejavnosti › Proizvodnja orodja ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predp
15	evem.gov.si/evem.gov.si.267.html	e-VEM eVEM › Dejavnosti › Proizvodnja olja ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpis
15	evem.gov.si/evem.gov.si.522.html	e-VEM eVEM › Dejavnosti › Spremljajoče storitvene ... ' Spremljajoče storitvene dejavnosti v kopenskem prometu , ...) Spremljajoče storit
15	evem.gov.si/evem.gov.si.546.html	e-VEM eVEM › Dejavnosti › Mlekarstvo in ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpisi i
14	evem.gov.si/evem.gov.si.133.html	e-VEM eVEM › Dejavnosti › Dirkanje in ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpisi in
14	evem.gov.si/evem.gov.si.161.html	e-VEM eVEM › Dejavnosti › Druga površinska ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpis
14	evem.gov.si/evem.gov.si.199.html	e-VEM eVEM › Dejavnosti › Stavbno mizarstvo ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Predpi
14	evem.gov.si/evem.gov.si.224.html	e-VEM eVEM › Dejavnosti › Tehnično preizkuševanje ... SKD šifra zajema dejavnosti in storitve , ... začetek in opravljanje dejavnosti . Pr

Fig. 5. Rezultati poizvedbe 'predelovalne dejavnosti'.

Results for a query: "trgovina"
Results found in 5.092375457ms.

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	gl . 46.110 trgovina na debelo s ... gl . 10.890 trgovina na debelo ž ... gl . 10.890 trgovina na debelo s ... gl . 46.380 trgovina na dro
96	evem.gov.si/evem.gov.si.051.html	Druga govorjenja Druga trgovina na drobno v ... specializiranih prodajalnah Druga trgovina na drobno v ... nespecializiranih prodajalnah D
92	evem.gov.si/evem.gov.si.21.html	eVEM : Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina
82	podatki.gov.si/podatki.gov.si.340.html	A DENT _ trgovina in storitve _ ... . ADRIA INVESTICIJE trgovina _ posredništvo _ ... d.o.o . ANATSEVSIS trgovina in storitve _ ... d.o.o
13	evem.gov.si/evem.gov.si.023.html	› Dejavnosti › Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... Sem spada : trgovina na debelo z ... izdelki ipd .
12	evem.gov.si/evem.gov.si.329.html	› Dejavnosti › Trgovina na debelo z ... in sanitarno opremo Trgovina na debelo z ... Sem spada : trgovina na debelo z ... z neobdelanim le
12	evem.gov.si/evem.gov.si.630.html	› Dejavnosti › Trgovina na drobno v ... predmeti za gospodinjstvo Trgovina na drobno v ... spada : specializirana trgovina na drobno s ...
10	evem.gov.si/evem.gov.si.320.html	› Dejavnosti › Trgovina na debelo s ... napravami za ogrevanje Trgovina na debelo s ... Sem spada : trgovina na debelo s ... izdelki , kli
10	evem.gov.si/evem.gov.si.327.html	› Dejavnosti › Trgovina na debelo z ... napravami in opremo Trgovina na debelo z ... Sem spada : trgovina na debelo s ... koles in koles t
10	evem.gov.si/evem.gov.si.622.html	› Dejavnosti › Trgovina na debelo z ... električnimi gospodinjskimi napravami Trgovina na debelo z ... Sem spada : trgovina na debelo z ...
9	evem.gov.si/evem.gov.si.328.html	› Dejavnosti › Trgovina na debelo s ... in plinastimi gorivi Trgovina na debelo s ... ta podvarstvo spada trgovina na debelo s ... Sem spad
9	evem.gov.si/evem.gov.si.343.html	› Dejavnosti › Trgovina na drobno v ... prodajalnah s tekstilom Trgovina na drobno v ... spada : specializirana trgovina na drobno s ... s
9	evem.gov.si/evem.gov.si.620.html	› Dejavnosti › Trgovina na debelo z ... napravami in deli Trgovina na debelo z ... Sem spada : trgovina na debelo z ... ventili in cevmi t
8	evem.gov.si/evem.gov.si.316.html	› Dejavnosti › Trgovina na debelo s ... (46.720) Trgovina na debelo s ... Sem spada : trgovina na debelo z ... debelo z rudami trgovina
8	evem.gov.si/evem.gov.si.323.html	› Dejavnosti › Trgovina na debelo s ... debelo s tekstilom Trgovina na debelo s ... Sem spada : trgovina na debelo s ... debelo s preja tr
8	evem.gov.si/evem.gov.si.334.html	› Dejavnosti › Trgovina na debelo z ... semeni in krmo Trgovina na debelo z ... Sem spada : trgovina na debelo z ... invjem in semeni trgo
7	evem.gov.si/evem.gov.si.315.html	› Dejavnosti › Trgovina na debelo s ... (46.460) Trgovina na debelo s ... Sem spada : trgovina na debelo z ... farmacevtskimi izdelki ,
7	evem.gov.si/evem.gov.si.319.html	› Dejavnosti › Trgovina na debelo s ... priključki , opremo Trgovina na debelo s ... Sem spada : trgovina na debelo s ... s kmetijskimi tr
7	evem.gov.si/evem.gov.si.339.html	› Dejavnosti › Trgovina na drobno v ... z rabljenim blagom Trgovina na drobno v ... Sem spada : trgovina na drobno s ... drobno s starinam
7	evem.gov.si/evem.gov.si.350.html	› Dejavnosti › Trgovina na drobno v ... barvami in steklom Trgovina na drobno v ... spada : specializirana trgovina na drobno ' ... spada
7	evem.gov.si/evem.gov.si.450.html	› Dejavnosti › Trgovina , vzdrževanje in ... motornih koles ; trgovina z njihovimi deli ... (45.400) Trgovina , vzdrževanje in ... motor
7	evem.gov.si/evem.gov.si.617.html	› Dejavnosti › Trgovina na debelo s ... preprogami in svetili Trgovina na debelo s ... Sem spada : trgovina na debelo s ... pohištvom za g
6	evem.gov.si/evem.gov.si.330.html	› Dejavnosti › Trgovina na debelo z ... raki , mehkužci Trgovina na debelo z ... spada tudi : trgovina na debelo s ... s pripravljenimi je
6	evem.gov.si/evem.gov.si.331.html	› Dejavnosti › Trgovina na debelo z ... ostanki in odpadki Trgovina na debelo z ... ostanki in odpadki Trgovina na debelo z ... obdelave)
6	evem.gov.si/evem.gov.si.339.html	› Dejavnosti › Trgovina na debelo z ... z računalniškimi napravami Trgovina na debelo z ... Sem spada : trgovina na debelo z ... spada tud
6	evem.gov.si/evem.gov.si.352.html	› Dejavnosti › Trgovina na drobno v ... raki , mehkužci Trgovina na drobno v ... spada tudi : trgovina na drobno z ...) stanju . Trgovina

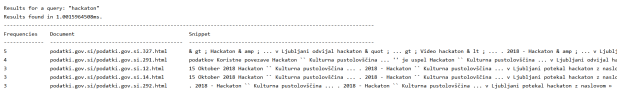
Fig. 6. Rezultati poizvedbe 'trgovina'.



Results for a query: "social services"
Results found in 0.999212265ms.

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.45.html	, retirement Social services , health , ... etc. ? Social services , health , ... Labour , retirement Social services , health ... relationsh
5	e-uprava.gov.si/e-uprava.gov.si.9.html	, retirement Social services , health , ... etc. ? Social services , health , ... Labour , retirement Social services , health ... relationsh
1	evem.gov.si/evem.gov.si.661.html	Records and Related Services (63985)
1	podatki.gov.si/podatki.gov.si.340.html	recreation and spa services ltd. TERME MARIBOR

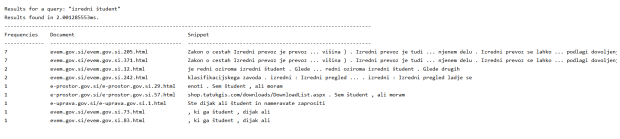
Fig. 7. Rezultati poizvedbe 'social services'.



Results for a query: "hackaton"
Results found in 1.001519645086s.

Frequencies	Document	Snippet
5	podatki.gov.si/podatki.gov.si.327.html	& gt ; Hackaton & amp ; ... » ljubljani odvijal hackaton & quot ; ... gt ; Video hackaton & lt ; ... . 2018 - Hackaton & amp ; ... » ljubl
4	podatki.gov.si/podatki.gov.si.291.html	podatkov Koristne povezave Hackaton `` Kulturna postolovišćina ... '' je uspel Hackaton `` Kulturna postolovišćina ... » ljubljani odvijal ha
3	podatki.gov.si/podatki.gov.si.12.html	15 Oktober 2018 Hackaton `` Kulturna postolovišćina ... . 2018 - Hackaton `` Kulturna postolovišćina ... » ljubljani potekal hackaton z nasl
3	podatki.gov.si/podatki.gov.si.16.html	15 Oktober 2018 Hackaton `` Kulturna postolovišćina ... . 2018 - Hackaton `` Kulturna postolovišćina ... » ljubljani potekal hackaton z nasl
3	podatki.gov.si/podatki.gov.si.292.html	. 2018 - Hackaton `` Kulturna postolovišćina ... . 2018 - Hackaton `` Kulturna postolovišćina ... » ljubljani potekal hackaton z naslovom «

Fig. 8. Rezultati poizvedbe 'hackaton'.



Results for a query: "izredni študent"
Results found in 2.000126553ms.

Frequencies	Document	Snippet
7	evem.gov.si/evem.gov.si.395.html	Zakon o cestah Izredni prevoz je prevoz ... višine) . Izredni prevoz je tudi ... njenem delu . Izredni prevoz se lahko ... podlagi dovoljen
7	evem.gov.si/evem.gov.si.371.html	Zakon o cestah Izredni prevoz je prevoz ... višine) . Izredni prevoz je tudi ... njenem delu . Izredni prevoz se lahko ... podlagi dovoljen
2	evem.gov.si/evem.gov.si.12.html	je redni oziroma izredni študent . Glede ... , redni oziroma izredni študent . Glede drugih
2	evem.gov.si/evem.gov.si.242.html	klasifikacijskega zavoda . izredni) Izredni pregled ... . izredni) Izredni pregled ljudje so se
1	e-prostor.gov.si/e-prostor.gov.si.29.html	enoti . Sem študent , ali moram
1	e-prostor.gov.si/e-prostor.gov.si.57.html	shop.tutakgis.com/downloads/DownloadList.aspx . Sem študent , ali moram
1	e-uprava.gov.si/e-uprava.gov.si.1.html	Ste dijak ali študent in nameravate zaposliti
1	evem.gov.si/evem.gov.si.73.html	, ki ga študent , dijak ali
1	evem.gov.si/evem.gov.si.83.html	, ki ga študent , dijak ali

Fig. 9. Rezultati poizvedbe 'izredni študent'.

Results for a query: "vzdrževanje stiskalnic za iverne plošče"
Results found in 2.000313878ms.

Frequencies	Document	Snippet
248	evem.gov.si/evem.gov.si.371.html	proizvodnja stiskalnic za iverne plošče in podobno ... vzdrževanje stiskalnic za iverne plošče in podobno ... vzdrževanje stiskalnic za ive
71	evem.gov.si/evem.gov.si.335.html	vzdrževanje stiskalnic za iverne plošče in podobno ... vzdrževanje stiskalnic za iverne plošče in podobno ... stiskalnic za iverne plošče i
35	evem.gov.si/evem.gov.si.651.html	vzdrževanje stiskalnic za iverne plošče ipd . ... stiskalnic za iverne plošče ipd . Popravila ... Popravila in vzdrževanje stiskalnic za iv
31	evem.gov.si/evem.gov.si.35.html	vzdrževanje stiskalnic za iverne plošče ipd . ... stiskalnic za iverne plošče ipd . Popravila ... Popravila in vzdrževanje stiskalnic za iv
29	podatki.gov.si/podatki.gov.si.360.html	podjetje za proizvodnjo stiskalnic , strojnih delov ... , upravljanje in vzdrževanje d.o.o . ALBA ... AMSAT Montaža in vzdrževanje elektron
27	evem.gov.si/evem.gov.si.377.html	vzdrževanje stiskalnic za iverne plošče in podobno ... stiskalnic za iverne plošče in podobno Nosilec ... popravila in vzdrževanje stiskalni
16	evem.gov.si/evem.gov.si.227.html	› Dejavnosti › Vzdrževanje železniških vozil Vzdrževanje ... Vzdrževanje železniških vozil Vzdrževanje železniških vozil Vzdrževanje ... Vo
11	evem.gov.si/evem.gov.si.244.html	Dejavnosti › Redno vzdrževanje pristaniške infrastrukture Redno ... pristaniške infrastrukture Redno vzdrževanje pristaniške infrastrukture
8	evem.gov.si/evem.gov.si.103.html	vzdrževanje stiskalnic za iverne plošče ipd . ... vzdrževanje stiskalnic za iverne plošče ipd . ... stiskalnic za iverne plošče ipd . Popra
8	evem.gov.si/evem.gov.si.468.html	sodov popravila in vzdrževanje kotlov za pridobivanje ... akumulatorjev popravila in vzdrževanje jadrnih reaktorjev , ... izotopov popravi
7	evem.gov.si/evem.gov.si.130.html	› Popravila in vzdrževanje drugih prevoznih sredstev ...) Popravila in vzdrževanje drugih prevoznih sredstev ... vzdrževanje železniških w
7	evem.gov.si/evem.gov.si.536.html	Dejavnosti › Redno vzdrževanje objektov za varnost ... varnost plovbe Redno vzdrževanje objektov za varnost ... varnost plovbe Redno vzdrže
6	e-prostor.gov.si/e-prostor.gov.si.13.html	za označevanje in vzdrževanje državne meje ... , , označevanje in vzdrževanje državne meje , ... , označevanje in vzdrževanje slovensko - m
6	evem.gov.si/evem.gov.si.445.html	› Popravila in vzdrževanje ladij in čolnov ...) Popravila in vzdrževanje ladij in čolnov ... popravila in splošno vzdrževanje ladij . popr
4	evem.gov.si/evem.gov.si.133.html	› Popravila in vzdrževanje zračnih in vesoljskih ...) Popravila in vzdrževanje zračnih in vesoljskih ... popravila in vzdrževanje zrakop
4	evem.gov.si/evem.gov.si.18.html	v železniškem prometu Vzdrževanje javne železniške infrastrukture ... na nacionalni ravni Vzdrževanje železniških vozil Redno ... železniš
4	evem.gov.si/evem.gov.si.235.html	› Dejavnosti › Vzdrževanje javne železniške infrastrukture ... javne železniške infrastrukture Vzdrževanje javne železniške infrastrukture
4	evem.gov.si/evem.gov.si.653.html	Koncesijska pogodba za vzdrževanje pristaniške infrastrukture Kvalificirani ... usposobljenosti delavcic za vzdrževanje železniških vozil P

Fig. 10. Rezultati poizvedbe 'vzdrževanje stiskalnic za iverne plošče'.

V. Zaključek

Kot vidimo, lahko z uporabo ustrezno zgrajenega invertiranega indeksa močno pospešimo iskanje po besedilu, še posebej v večjih korpusih. Izziv je predvsem predhodno procesiranje spletnih strani, še posebej če to opravljamo na spletu, ki se spreminja zelo dinamično. V tem primeru je potrebno invertiran indeks ustrezno posodabljati.