Poročilo tretje seminarske naloge

Mark Bogataj in Jakob Maležič
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Ljubljana, Slovenija
Mentor: asist. prof. dr. Slavko Žitnik

I. Uvod

V tem poročilu je predstavljena implementacija preprostega inverznega indeksa za iskanje relevantnih besedil.

II. Obdelava besedil in indeksiranje

A. Obdelava besedil

Najprej sva z uporabo knjižnice `BeautifulSoap` odstranila vse značke `script` in `noscript` ter izluščila zgolj besedilo brez značk. Nato sva zamenjala ločila, vse skoke v novo vrstico in ponavljajoče presledke z enojnim presledkom ter vse črke spremenila v male tiskane.

B. Indeksiranje

Indeksiranje sva naredila tako, da sva najprej obdelala besedilo kakor opisano nato pa sva z uporabo knjižnice `nltk` besedilo spremenila v seznam posameznih besed (tokenov) in odstranila končne besede (ang. stop words). Vsako izmed besed v seznamu sva nato poiskala v besedilu in si zapomnila njeno pozicijo. Slednje sva nato shranila v bazo in ustrezno posodobila število pojavitev besede.

C. Podatkovna baza

Po indeksiranju najina podatkovna baza vsebuje 45310 različnih besed. V tabeli `Posting` pa je 376162 zapisov. Beseda z največ pojavitvami je *podatkov* z 12258 pojavitvami, dokument z največjim številom pojavitev pa je *evem.gov.si/evem.gov.si.371.html* z 220898 pojavitvami.

III. Pridobivanje podatkov

A. Inverzni indeks

S tako zgrajeno podatkovno bazo lahko sedaj poiščeva rezultate poljubne poizvedbe. Vpisano poizvedbo najprej obdelava na enak način kakor besedilo, spremeniva v seznam posameznih besed in odstraniva končne besede. Ta seznam nato uporabiva za iskanje dokumentov z največ pojavitvami podanih besed. To narediva z naslednjo `SQL` poizvedbo:

```
SELECT
    SUM(frequency) AS "frequencies",
    documentName,
    group_concat(indexes)
FROM Posting
WHERE {"or ".join(["word LIKE ?"]*len(words_))}
GROUP BY documentName
ORDER BY frequencies DESC
LIMIT 10
```

Nato iterirava čez rezultate in iz vsakega dokumenta izbereva prvih pet pozicij pojavitve ene izmed besed. Za pozicije iz besedila tega dokumenta, ki sva ga obdelala enako kot pri indeksiranju z razliko, da nisva črk spremenila v male tiskane, izluščiva okolico in si shraniva tri besede pred in za iskano besedo. Slednje nato uporabiva pri izpisu rezultatov.

B. Zaporedno iskanje

Podobno kot pri inverznemu indeksu, sva poizvedbo najprej obdelala. Nato iterirava čez vse dokumente, jih obdelava enako kot prej in v besedilu poiščeva iskane besede ter v slovar shraniva njihovo pozicijo v besedilu in posodobiva število pojavitev. Shraniva pa tudi obdelano besedilo, brez spreminjanja v male tiskane črke, da nama kasneje ni potrebno ponovno odpirati dokumentov.

Nato dokumente urediva po številu pojavitev besed in na enak način kot pri inverznem indeksu kreirava okolice besed, ki jih nato uporabiva pri izpisu rezultatov.

IV. Izzivi

A. Prikaz iz originalnega besedila

Okolico iskanih besedil bi morala izluščiti iz originalnega besedila. To nama je povzročalo težave, saj se indeksi v obdelanem besedilu razlikujejo indeksom v neobdelanem besedilu. Zato sva pri iskanju okolic iskanih besed na podoben način obdelala besedilo. To bi lahko izboljšala tako, da bi najprej indeksirala vse besede in nato odstranila ustavitvene besede. Tako bi bili indeksi najdenih besed še vedno pravilni, pri izpisu okolice pa bi izpisala tudi končne besede.

B. Prekrivanje snippetov

Pri izpisu rezultatov se lahko okolice iskanih besed prekrivajo. Zato ni smiselno da vedno izpišemo vse okolice. To sva poskusila rešiti tako, da sva si shranila besede, za katere sva že našla okolice. Nato sva pri iskanju okolic drugih besede preverjala, ali vsebujejo kakšno shranjenih besed. Če vsebuje, sva takšno okolico preskočila, saj je le ta že shranjena pri prejšni besedi.

V. Rezultati

Iz tabele I lahko vidimo da je iskanje s pomočjo inverznega indeksa veliko hitreje kot zaporedno iskanje.

TABLE I
PRIMERJAVA ČASOV IZVAJAN.

Algoritem / Poizvedba	predelovalne dejavnosti	trgovina	social services	Republika Slovenija	davek	vloge in obvestila
Inverzni indeks	3.34s	3.42s	1.03s	3.38s	0.25s	2.51s
Zaporedno iskanje	39.94s	40.84s	40.51s	39.65s	41.01s	39.96s

Results for a query: predelovalne dejavnosti

Results found in: 3.34 s.

Frequencies	Document	Snippet
1570	evem.gov.si/evem.gov.si.371.html	iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljar
78	evem.gov.si/evem.gov.si.377.html	Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravst
49	evem.gov.si/evem.gov.si.452.html	Druge storitvene dejavnosti drugje nerazvrščene (... 090) / Dejavnosti / eVEM R
40	podatki.gov.si/podatki.gov.si.340.html	- NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH DEJ
31	evem.gov.si/evem.gov.si.653.html	Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske ali
31	evem.gov.si/evem.gov.si.398.html	usmerjene na opravljanje dejavnosti (npr : ... za namene opravljanja dejavnosti
30	evem.gov.si/evem.gov.si.72.html	od dohodka iz dejavnosti Republika Slovenija SPOT ... od dohodka iz dejavnosti Da
30	evem.gov.si/evem.gov.si.442.html	Dejavnosti za nego telesa ... 040) / Dejavnosti / eVEM Republika ... e-VEM eVEM
25	evem.gov.si/evem.gov.si.460.html	Drugje nerazvrščene predelovalne dejavnosti (32 990 ... 990) / Dejavnosti / eVE
23	evem.gov.si/evem.gov.si.276.html	620) / Dejavnosti / eVEM Republika ... e-VEM eVEM › Dejavnosti › Storitve za ...

Fig. 1. Rezultati za poizvedbo ”predelovalne dejavnosti”.

Results for a query: trgovina

Results found in: 3.42 s.

Frequencies	Document	Snippet
368	evem.gov.si/evem.gov.si.371.html	gl 46 110 trgovina na debelo s ... gl 10 890 trgovina na debelo z ... g
96	evem.gov.si/evem.gov.si.651.html	Druga govedoreja Druga trgovina na drobno v ... specializiranih prodaja
92	evem.gov.si/evem.gov.si.21.html	eVEM › Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trg
82	podatki.gov.si/podatki.gov.si.340.html	o A DENT trgovina in storitve d ... o ADRIA INVESTICIJE trgovina posred
14	evem.gov.si/evem.gov.si.623.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... izdelk
13	evem.gov.si/evem.gov.si.630.html	Trgovina na drobno v ... › Dejavnosti › Trgovina na drobno v ... predme
13	evem.gov.si/evem.gov.si.329.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... in sar
11	evem.gov.si/evem.gov.si.622.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... elektr
11	evem.gov.si/evem.gov.si.327.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... naprav
11	evem.gov.si/evem.gov.si.320.html	Trgovina na debelo s ... › Dejavnosti › Trgovina na debelo s ... naprav

Fig. 2. Rezultati za poizvedbo ”trgovina”.

Results for a query: social services

Results found in: 1.03 s.

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

Fig. 3. Rezultati za poizvedbo ”social services”.

Results for a query: davek

Results found in: 0.25 s.

Frequencies	Document	Snippet
18	evem.gov.si/evem.gov.si.7.html	dajatve trošarine in davek na dodano vrednost ... Sloveniji DDV (Davek na do
6	evem.gov.si/evem.gov.si.9.html	/ Davki / Davek od dohodka pravnih ... › Davki › Davek od dohodka pravnih ...
6	evem.gov.si/evem.gov.si.71.html	/ Davki / Davek na dodano vrednost ... › Davki › Davek na dodano vrednost ...
5	evem.gov.si/evem.gov.si.72.html	/ Davki / Davek od dohodka iz ... › Davki › Davek od dohodka iz ... dohodka i
4	e-uprava.gov.si/e-uprava.gov.si.52.html	kdaj se plačuje davek na promet nepremičnin ... Kdaj moram plačevati davek na
2	evem.gov.si/evem.gov.si.90.html	• DDV • Davek od dohodka pravnih ... pravnih oseb • Davek od dohodka iz
2	evem.gov.si/evem.gov.si.69.html	• DDV • Davek od dohodka pravnih ... pravnih oseb • Davek od dohodka iz
2	evem.gov.si/evem.gov.si.662.html	• DDV • Davek od dohodka pravnih ... pravnih oseb • Davek od dohodka iz
2	evem.gov.si/evem.gov.si.633.html	• DDV • Davek od dohodka pravnih ... pravnih oseb • Davek od dohodka iz
2	evem.gov.si/evem.gov.si.60.html	• DDV • Davek od dohodka pravnih ... pravnih oseb • Davek od dohodka iz

Fig. 4. Rezultati za poizvedbo ”davek”.

Results for a query: Republika Slovenija

Results found in: 3.38 s.

Frequencies	Document	Snippet
128	podatki.gov.si/podatki.gov.si.340.html	- v likvidaciji REPUBLIKA SLOVENIJA REPUBLIKA SLOVENIJA ... likvidaciji REPUBLIKA SLO
30	podatki.gov.si/podatki.gov.si.414.html	REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... Podrobnosti Organizacija : REPUBLIKA SLOVENIJA
16	evem.gov.si/evem.gov.si.371.html	eVEM Republika Slovenija SPOT Slovenska ... edina družbenica je Republika Slovenija S
14	podatki.gov.si/podatki.gov.si.424.html	in statističnih regijah Slovenija letno 55 ogledov ... poškodbe in spolu Slovenija le
14	e-prostor.gov.si/e-prostor.gov.si.166.html	1012 4 VZHODNA SLOVENIJA MM 10000 528342 ... 370 67 VZHODNA SLOVENIJA MM 10000 540391
13	podatki.gov.si/podatki.gov.si.5.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogledov REPUBLIKA SLOVENIJ
13	podatki.gov.si/podatki.gov.si.408.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogledov REPUBLIKA SLOVENIJ
13	podatki.gov.si/podatki.gov.si.407.html	(277) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogledov REPUBLIKA SLOVENIJ
13	podatki.gov.si/podatki.gov.si.34.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogledov REPUBLIKA SLOVENIJ
13	podatki.gov.si/podatki.gov.si.142.html	(33) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogledov REPUBLIKA SLOVENIJA

Fig. 5. Rezultati za poizvedbo ”Republika Slovenija”.

Results for a query: vloge in obvestila

Results found in: 2.51 s.

Frequencies	Document	Snippet
15	evem.gov.si/evem.gov.si.403.html	status oddane vloge Obvestila glede sprememb statusov ... omogoča elektronsko oddajo
14	evem.gov.si/evem.gov.si.37.html	Navodila za izpolnjevanje vloge za pridobitev potrdil ... Navodila za izpolnjevanje
9	e-uprava.gov.si/e-uprava.gov.si.44.html	vsebine brez vnaprejšnjega obvestila Da bi uporabnikom ... spreminjata brez predhodn
8	evem.gov.si/evem.gov.si.371.html	začne na podlagi vloge poslovnega subjekta s ... od prejema popolne vloge na podlagi
8	e-uprava.gov.si/e-uprava.gov.si.22.html	v nastavitvah obveščanja Obvestila Prikazana so vsa ... Prikazana so vsa obvestila n
6	evem.gov.si/evem.gov.si.398.html	z dnem oddaje vloge preko portala e-VEM ... portala e-VEM oddam vloge za prijavo v .
5	evem.gov.si/evem.gov.si.84.html	tujca vloži naslednje vloge po enotnem dovoljenju ... o tujcih) Vloge za enotna dov
5	evem.gov.si/evem.gov.si.372.html	: Postopek oddaje vloge za pridobitev obrtnega ... dovoljenja Postopek oddaje vloge
4	podatki.gov.si/podatki.gov.si.16.html	vsebine brez vnaprejšnjega obvestila Uporabnik je torej ... spreminjata brez predhod
4	evem.gov.si/evem.gov.si.368.html	leti pred vložitvijo vloge za pridobitev statusa ... leti pred vložitvijo vloge za p

Fig. 6. Rezultati za poizvedbo ”vloge in obvesila”.

Results for a query: predelovalne dejavnosti

Results found in: 39.94 s.

Frequencies	Document	Snippet
1570	evem.gov.si.371.html	za infrastrukturo C PREDELOVALNE DEJAVNOSTI 10 Proizvodnja ... 32 Druge raznovrstne pre
78	evem.gov.si.377.html	Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni d
49	evem.gov.si.452.html	Druge storitvene dejavnosti drugje nerazvrščene (... 090) / Dejavnosti / eVEM Republi
40	podatki.gov.si.340.html	- NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH DEJAVNOST
31	evem.gov.si.398.html	usmerjene na opravljanje dejavnosti (npr : ... za namene opravljanja dejavnosti ipd V
31	evem.gov.si.653.html	Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske ali telev
30	evem.gov.si.442.html	Dejavnosti za nego telesa ... 040) / Dejavnosti / eVEM Republika ... e-VEM eVEM › Deja
30	evem.gov.si.72.html	od dohodka iz dejavnosti Republika Slovenija SPOT ... od dohodka iz dejavnosti Davek od
25	evem.gov.si.460.html	Drugje nerazvrščene predelovalne dejavnosti (32 ... › Drugje nerazvrščene predelovalne
23	evem.gov.si.265.html	110) / Dejavnosti / eVEM Republika ... e-VEM eVEM › Dejavnosti › Proizvodnja mesa ...

Fig. 7. Rezultati brez inverznega indeksa za poizvedbo ”predelovalne dejavnosti”.

Results for a query: trgovina

Results found in: 40.84 s.

Frequencies	Document	Snippet
368	evem.gov.si.371.html	gl 46 110 trgovina na debelo s ... gl 10 890 trgovina na debelo z ...
96	evem.gov.si.651.html	Druga govedoreja Druga trgovina na drobno v ... specializiranih proda
92	evem.gov.si.21.html	eVEM › Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga t
82	podatki.gov.si.340.html	o A DENT trgovina in storitve d ... o ADRIA INVESTICIJE trgovina posre
14	evem.gov.si.623.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... izdel
13	evem.gov.si.329.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... in s
13	evem.gov.si.630.html	Trgovina na drobno v ... › Dejavnosti › Trgovina na drobno v ... predm
11	evem.gov.si.320.html	Trgovina na debelo s ... › Dejavnosti › Trgovina na debelo s ... napra
11	evem.gov.si.327.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... napra
11	evem.gov.si.622.html	Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo z ... elekt

Fig. 8. Rezultati brez inverznega indeksa za poizvedbo ”trgovina”.

Results for a query: social services

Results found in: 40.51 s.

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

Fig. 9. Rezultati brez inverznega indeksa za poizvedbo ”social services”.

Results for a query: davek

Results found in: 41.01 s.

Frequencies	Document	Snippet
18	evem.gov.si.7.html	dajatve trošarine in davek na dodano vrednos
6	evem.gov.si.71.html	/ Davki / Davek na dodano vrednost ... › Dav
6	evem.gov.si.9.html	/ Davki / Davek od dohodka pravnih ... › Dav
5	evem.gov.si.72.html	/ Davki / Davek od dohodka iz ... › Davki ›
4	e-uprava.gov.si.52.html	kdaj se plačuje davek na promet nepremičnin
2	e-prostor.gov.si.1.html	katere se obračuna davek na dodano vrednost
2	e-prostor.gov.si.121.html	katere se plača davek na promet nepremičnin
2	e-prostor.gov.si.57.html	katere se plača davek na promet nepremičnin
2	evem.gov.si.1.html	• DDV • Davek od dohodka pravnih ... pravnih
2	evem.gov.si.10.html	• DDV • Davek od dohodka pravnih ... pravnih

Fig. 10. Rezultati brez inverznega indeksa za poizvedbo ”davek”.

Results for a query: Republika Slovenija

Results found in: 39.65 s.

Frequencies	Document	Snippet
128	podatki.gov.si.340.html	- v likvidaciji REPUBLIKA SLOVENIJA REPUBLIKA SLOVENIJA ... lik
30	podatki.gov.si.414.html	REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... Podrobnosti Organizacija
16	evem.gov.si.371.html	eVEM Republika Slovenija SPOT Slovenska ... edina družbenica je
14	e-prostor.gov.si.166.html	1012 4 VZHODNA SLOVENIJA MM 10000 528342 ... 370 67 VZHODNA SLO
14	podatki.gov.si.424.html	in statističnih regijah Slovenija letno 55 ogledov ... poškodbe
13	podatki.gov.si.142.html	(33) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogled
13	podatki.gov.si.34.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogle
13	podatki.gov.si.407.html	(277) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogle
13	podatki.gov.si.408.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogle
13	podatki.gov.si.5.html	(220) REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... varstva 149 ogle

Fig. 11. Rezultati brez inverznega indeksa za poizvedbo ”Republika Slovenija”.

Results for a query: vloge in obvestila

Results found in: 39.96 s.

Frequencies	Document	Snippet
15	evem.gov.si.403.html	omogoča elektronsko oddajo vloge za pridobitev ali ... in mo
14	evem.gov.si.37.html	Navodila za izpolnjevanje vloge za pridobitev potrdil ... Na
9	e-uprava.gov.si.44.html	opisano storitev Uporabniki vloge lahko izpolnijo in ... pri
8	e-uprava.gov.si.22.html	v vaše podatke vloge premoženje in sodelovanje ... pozabili
8	evem.gov.si.371.html	začne na podlagi vloge poslovnega subjekta s ... od prejema
6	evem.gov.si.398.html	z dnem oddaje vloge preko portala e-VEM ... portala e-VEM od
5	evem.gov.si.372.html	: Postopek oddaje vloge za pridobitev obrtnega ... dovoljenj
5	evem.gov.si.84.html	tujca vloži naslednje vloge po enotnem dovoljenju ... o tujc
4	evem.gov.si.368.html	leti pred vložitvijo vloge za pridobitev statusa ... leti pr
4	podatki.gov.si.16.html	vsebine brez vnaprejšnjega obvestila Uporabnik je torej ...

Fig. 12. Rezultati brez inverznega indeksa za poizvedbo ”vloge in obvestila”.