Naloga 3: Indeksiranje in poizvedovanje

Blaž Marolt, Rok Šolar, Anže Veršnik

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko

1 Uvod

Tretja domača naloga je zahtevala implementacijo gradnje invertnega indeksa ter rangiranje shranjenih dokumentov glede na podano poizvedbo. Na voljo smo imeli 1416 spletnih strani iz štirih domen s katerimi smo zgradili invertni indeks. V tem delu bomo predstavili našo implementacijo gradnje invertnega indeksa ter implementacijo pridobivanja rezultatov za vnešeno poizvedbo. Pri pridobivanju rezultatov bomo primerjali hitrosti med iskanjem po zgrajenem indeksu ter iskanjem brez indeksa (sekvenčno pregledovanje dokumentov). Nato bomo predstavili statistične podatke dobljene podatkovne baze (invertnega indeksa) ter prikazali rezultate šestih poizvedb.

2 Implementacija

2.1 Procesiranje podatkov

Vsak dokument najprej odpremo s kodiranjem UTF-8. Dobljen HTML posredujemo knjižnici `BeautifulSoup`, ki nam s funkcijo `findAll(text=True)` vrne seznam vseh besedil znotraj dokumenta. Iz dobljenega seznama nato filtriramo:

- Elemente, katerih ime starša je `style, script, head, title, meta, [document] ali noscript`,
- elemente, ki so HTML komentarji,
- elemente, ki ne vsebujejo ničesar ali pa samo znak za novo vrstico.

Vsako besedilo iz dobljenega seznama nato tokeniziramo (funkcija `word_tokenize` iz knjižnica `nltk`) ter vrnemo seznam besed. Vsako besedo nato obdelamo na sledeč način:

- Iz začetka ali konca besede odstranimo ločila (definirana v `string.punctuation`). Ločila znotraj besede dovolimo.
- Vse črke pretvorimo v male.
- Odstranimo. oz ignoriramo besede, ki niso ključne (ang. stopwords).

2.2 Indeksiranje

Prvi korak pri implementaciji je indeksiranje, kjer besedo vpisujemo v podatkovno bazo. Indeksiranje smo implementirali tako, da odpremo dokument in

nato pridobimo tekst iz dokumenta HTML. Nato za vsako besedo najprej preverimo, ali je ključna (stopword) in če je, jo zapišemo v podatkovno bazo v tabelo IndexWord. Pri vsaki besedi pogledamo če smo tako besedo v dokumentu že videli. Če je še nismo videli preverimo ponovitve te besede v dokumentu, zberemo število ponovitev in shranjujemo indekse, kje besedo najdemo. Po pregledu dokumenta shranimo v tabelo število ponovitev, ime ter indekse ponovitve besede v dokumentu.

2.3 Poizvedovanje

Vnešeno poizvedbo obdelamo po postopku, opisanem v 2.1, brez obdelave kode HTML. Rezultate za obdelano poizvedbo dobimo z eno poizvedbo SQL, generirano z naslednjim ukazom:

```
sql = '''
SELECT p.documentName AS docName, SUM(frequency) AS freq, GROUP_CONCAT(indexes) AS idxs
FROM Posting p
WHERE
    p.word IN ({seq})
GROUP BY p.documentName
ORDER BY freq DESC;'''.format(seq=','.join(['?']*len(query)))
```

Pri tem je `query` seznam besed iz obdelane poizvedbe.
Za vsak rezultat nato odpremo pripadajoč dokument HTML ter generiramo odrezek (ang. snippet). Ta je definiran kot besedilo, dolgo največ 200 znakov, kjer so izpisane do največ tri besede pred in tri besede za najdenimi besedami iz poizvedbe.

3 Dobljena podatkovna baza

V tem poglavju bomo predstavili nekaj statističnih podatkov o dobljenem invertnem indeksu. Vseh indeksiranih besed je 47195. Povprečno ima vsak indeksiran dokument 273 unikatnih besed. V tabeli 1 je prikazanih deset najbolj pogostih besed in njihova frekvenca, v tabeli 2 pa deset dokumentov, ki imajo največje število unikatnih besed ter pripadajoče število unikatnih besed.

4 Primerjava hitrosti

V nadaljevanju bomo predstavili primerjavo hitrosti med iskanjem z invertnim indeksom ter sekvenčnim iskanjem. Obe metodi smo testirali tako, da smo merili le čas v katerem metoda vrne rezultat s številom besed iz poizvedbe v posameznem dokumentu. Kot lahko vidimo v tabeli 3 je iskanje z invertnim indeksom bistveno hitrejše. Poizvedbe z invertnim indeksom so tako hitre, da jih med sabo ne moremo primerjati zaradi vpliva ostalih procesov na računalniku, medtem ko se pri sekvenčnem iskanju lepo vidi vpliv količine besed v poizvedbi.

Beseda	**Frekvenca**
podatkov	11049
slovenije	9928
republike	8572
dejavnosti	5565
podatki	4940
portalu	4702
navigation	4474
krepko	4277
navadno	4242
pogoji	4180

Tabela 1. 10 najpogostejših besed v invertnem indeksu.

Dokument	**Št. unikatnih besed**
evem.gov.si.371.html	13256
podatki.gov.si.340.html	6540
e-prostor.gov.si.166.html	6101
e-prostor.gov.si.218.html	4497
e-prostor.gov.si.57.html	1749
evem.gov.si.398.html	1638
evem.gov.si.651.html	1293
e-uprava.gov.si.56.html	1204
e-prostor.gov.si.150.html	1067
e-uprava.gov.si.44.html	1056

Tabela 2. 10 dokumentov z največ unikatnih besed v invertnem indeksu.

Poizvedba	Invertni indeks [s]	Sekvenčno pregledovanje [s]
predelovalne dejavnosti	0,395	142,147
trgovina	0,031	97,180
social services	0,015	129,017
inšpekcijski pregled	0,023	70,648
recept	0,005	65,785
praksa	0,015	66,10388

Tabela 3. Primerjava časa pri iskanju z invertnim indeksom ter sekvenčnim iskanjem.

5 Rezultati poizvedb

V nadaljevanju sledijo rezultati poizvedb. Pri poizvedbah smo naredili omejitev na največ 10 izpisov, zaradi dolžine rezultata. Že samo rezultati poizvedbe predelovalne dejavnosti so v poročilu zasedli 42 strani. Rezultati so prikazani v obliki: frekvenca, ime dokumenta in odrezek (snippet).

5.1 Predelovalne dejavnosti

1291 evem.gov.si.371.html ... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti . V iskalnik ... 645 od 645 dejavnosti Izpisanih je od ... Izpisanih je od dejavnosti A KMETIJSTVO IN ...

75 evem.gov.si.377.html ... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdravstveni dejavnosti Dimnikar Diplomirana medicinska ... I v zdravstveni dejavnosti Laboratorijski sodelavec II ...

40 podatki.gov.si.340.html ... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER JUDOVSKE ... ŠOLSKIH IN OBŠOLSKIH DEJAVNOSTI Center urbane kulture ... in druge zdravstvene dejavnosti, d.o.o . ...

38 evem.gov.si.452.html ... e-VEM eVEM › Dejavnosti › Druge storitvene ... › Druge storitvene dejavnosti, drugje nerazvrščene ...) Druge storitvene dejavnosti, drugje nerazvrščene ... SKD šifra zajema dejavnosti in storitve, ...

31 evem.gov.si.653.html ... Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske ali televizijske dejavnosti Dovoljenje za izvajanje ... za izvajanje sevalne dejavnosti Dovoljenje za izvajanje ...

30 evem.gov.si.398.html ... usmerjene na opravljanje dejavnosti (npr . ... za namene opravljanja dejavnosti ipd . V ... mesecev z opravljanjem dejavnosti v Sloveniji nismo ... uporabljali za opravljanje dejavnosti . Ali se ...

29 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 ... Za dohodek iz dejavnosti se šteje dohodek ... neodvisnim samostojnim opravljanjem dejavnosti, ne glede ...

23 evem.gov.si.442.html ... e-VEM eVEM › Dejavnosti › Dejavnosti za ... › Dejavnosti › Dejavnosti za nego telesa ... (96.040) Dejavnosti za nego telesa ... SKD šifra zajema dejavnosti in storitve, ... začetek in opravljanje dejavnosti . Predpisi in ...

18 evem.gov.si.28.html ... za opravljanje gospodarske dejavnosti . Lastnosti zasebnega ... niso za posamezne dejavnosti ali posamezne vrste ... ustanovitev in opravljanje dejavnosti zavoda . Ime ... iz opravljanja nepridobitne dejavnosti se ne obdavči ...

17 evem.gov.si.265.html ... e-VEM eVEM › Dejavnosti › Proizvodnja mesa ... SKD šifra zajema dejavnosti in storitve, ... začetek in opravljanje dejavnosti . Predpisi in ... za opravljanje dopolnilne dejavnosti na kmetiji in ...

5.2 trgovina

364 evem.gov.si.371.html ... gl . 46.110 trgovina na debelo s ... gl . 10.890 trgovina na debelo z ... gl . 10.890 trgovina na debelo s ... gl . 46.380 trgovina na drobno s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ...

96 evem.gov.si.651.html ... Druga govedoreja Druga trgovina na drobno v ... specializiranih prodajalnah Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno v ... z živili Druga trgovina na drobno zunaj ...

92 evem.gov.si.21.html ... eVEM › Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno zunaj ... 47.990) Nespecializirana trgovina na debelo Trgovina ...

82 podatki.gov.si.340.html ... A DENT, trgovina in storitve, ... ADRIA INVESTICIJE trgovina, posredništvo, ... d.o.o . AHATSERVIS trgovina in storitve, ... d.o.o . ALBA trgovina in proizvodnja, ... , storitve in trgovina d.o.o . ALMA ...

13 evem.gov.si.623.html ... › Dejavnosti › Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... Sem spada : trgovina na debelo z ... izdelki ipd . trgovina na debelo s ... in deli zanja trgovina na debelo s ...

12 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 lesom trgovina na debelo s ... primarne obdelave lesa trgovina na debelo s ...

12 evem.gov.si.630.html ... › Dejavnosti › Trgovina na drobno v ... predmeti za gospodinjstvo Trgovina na drobno v ... spada : specializirana trgovina na drobno s ... s pohištvom specializirana trgovina na drobno s ...

10 evem.gov.si.320.html ... › Dejavnosti › Trgovina na debelo s ... napravami za ogrevanje Trgovina na debelo s ... Sem spada : trgovina na debelo s ... izdelki, ključavnicami trgovina na debelo z ... izdelki za pritrjevanje trgovina na debelo s ...

10 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 trgovina na debelo z ... z industrijskimi roboti trgovina na debelo z ...

10 evem.gov.si.622.html ... › Dejavnosti › Trgovina na debelo z ... električnimi gospodinjskimi napravami Trgovina na debelo z ... Sem spada : trgovina na debelo z ... z zabavno elektroniko trgovina na debelo s ... in podobnimi ploščami trgovina na debelo z ...

5.3 social services

5 e-uprava.gov.si.45.html ... , retirement Social services, health, ... ? Social services, health, ... Labour, retirement Social services, health ... etc . ? Social services, health ... I obtain financial social assistance ? How ...

5 e-uprava.gov.si.9.html ... , retirement Social services, health, ... ? Social services, health, ... Labour, retirement Social services, health ... etc . ? Social services, health ... I obtain financial social assistance ? How ...

1 evem.gov.si.661.html ... Records and Related Services (AJPES) ...

1 podatki.gov.si.340.html ... recreation and spa services ltd . TERME ...

5.4 inšpekcijski pregled

12 evem.gov.si.371.html ... diagnostičnih preparatov za pregled in vivo proizvodnja ... industrijsko rabo) Pregled tovornega dvigala Pregled ... Pregled tovor-

nega dvigala Pregled invalidske ploščadi Popravila ... osebnim vozilom na pregled v zdravstveni dom ...

9 evem.gov.si.14.html ... delu Preventivni zdravstveni pregled Usposabljanje iz požarnega ... tveganja je sistematičen pregled vseh vidikov dela ... . Preventivni zdravstveni pregled S preventivnimi zdravstvenimi ... na predhodni zdravstveni pregled : pred prvo ...

5 e-prostor.gov.si.150.html ... projekta . Podrobnejši pregled rezultatov projekta je ... ki vključuje : pregled končnih poročil in ... prilog), pregled prispevkov na konferencah ... delavnicah projekta, pregled objav v strokovnih ...

3 e-prostor.gov.si.16.html ... tudi : toponomastični pregled zemljepisnih imen, ... določenem ozemlju . Pregled vključuje terensko in ... pravopisni in slovnični pregled v skladu s ...

3 evem.gov.si.242.html ... in opremo ; pregled obstoječih ladij, ... osnovni : Osnovni pregled je obvezen za ... izredni : Izredni pregled ladje se opravi ...

2 e-prostor.gov.si.18.html ... dodatnih gradivih je pregled izbrane literature, ... parametrov EPSG in pregled podatkov, pomembnih ...

2 evem.gov.si.224.html ... snovi v zrak Pregled hidrantov (kontrolne ... mleka ; Tehnični pregled (pavšalni npr ...

2 evem.gov.si.398.html ... “ Prikaži osnovni pregled ”, ki ... 44,06 EUR Za pregled vpisa v sodnem ...

2 podatki.gov.si.189.html ... Okolje in prostor Pregled toka odpadkov, ... z naslovom “ Pregled toka odpadkov, ...

1 e-prostor.gov.si.103.html ... vprašalnikov REN in pregled podatkov . Z ...

5.5 recept

1 evem.gov.si.645.html ... zdravniški ali veterinarski recept, in sicer ...

5.6 praksa

2 podatki.gov.si.340.html ... SPLOŠNE MEDICINE PRIMARNA PRAKSA, splošna medicina ... d.o.o . VETERINARSKA PRAKSA TENETIŠE, D.O.O ...

1 evem.gov.si.48.html ... in odgovori Sodna praksa ZFPPIPP ZIZ Pravna ...

1 evem.gov.si.75.html ... zaposlitvi . Sodna praksa je dodala še ...

1 podatki.gov.si.308.html ... tuja in domača praksa na področju odpiranja ...

1 podatki.gov.si.553.html ... dostopnosti obstaja večletna praksa Informacijskega pooblaščenca . ...

6 Zaključek

V tem delu smo implementirali gradnjo in poizvedovanje po invertnem indeksu. Poleg tega smo implementirali tudi sekvenčno iskanje besed po vseh dokumentih, brez indeksa, s čimer smo ugotovili, da je poizvedovanje z invertnim indeksom odločno hitrejše. Pri poizvedovanju z invertnim indeksom je najpočasnejši del

generiranje odrezkov (ang. snippets), pri čimer moramo odpreti in prebrati vsako datoteko iz rezultata, zato generiranje odrezkov nismo upoštevali pri primerjavi hitrosti. Prikazali smo še nekaj statističnih podatkov o dobljeni podatkovni bazi (o invertnem indeksu). Možna nadgradnja sistema bi vsebovala lematizacijo besed, kar bi močno izboljšalo rezultate iskanja, saj že v tabeli 1 opazimo dve besedi, ki bi morali biti združeni (podatkov in podatki).