Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Beti Piščanec, Sebastjan Mevlja in Lan Zukanović

Indeksiranje besed spletnih strani

Seminarska naloga pri predmetu Iskanje in ekstrakcija podatkov s spleta

MENTOR: doc. dr. Slavko Žitnik

1. Uvod

Seminarska naloga opisuje implementacijo preprostega inverznega indeksa besed spletnih strani in primerjavo hitrosti poizvedb z ali brez uporabe le tega. Aplikacija najprej prebere HTML datoteke in iz njih izvleče besedilo, ki ga obdela in indeksira. Indeks je nato uporabljen za pohitritev poizvedb. Za boljši pregled smo omejili iskanje in izpis rezultata na 5 najvišjih frekvenc.

2. Obdelava podatkov in indeksiranje

Pred indeksiranjem preberemo vse HTML datoteke. Iz njih s pomočjo BeautifulSoup knjižnice odstranimo JavaScript kodo, oblikovanja besedila, HTML komentarje in meta podatke, ki jih ne potrebujemo. Odstranili smo tudi vse enojne in dvojne narekovaje. Z uporabo iste knjižnice še izvlečemo besedilo, ki nas zanima. Besedilo z uporabo knjižnice nltk preoblikujemo v žetone in iz njih odstranimo odvečne besede (angl. stopwords), ki so bile na predlaganem seznamu. Besede tudi normaliziramo v male črke. Za vsako besedo poiščemo, kje v izvornem dokumentu je. Besede, imena dokumentov, frekvenco in indekse shranimo v podatkovno bazo SQLite. Gradnja indeksa traja približno 3 minute.

3. Iskanje z inverznim indeksom

Iskanje uporablja obstoječi inverzni indeks, ki smo ga zgradili vnaprej. Iskane besede so pretvorjene v seznam žetonov, ki je uporabljen za iskanje po podatkovni bazi. Poizvedba vrne seznam imen dokumentov, seštevkov frekvenc in indekse besed. Vsi dokumenti, ki so omenjeni v rezultatih poizvedbe so prebrani in shranjeni v slovar. Iz njih s pomočjo BeautifulSoup knjižnice odstranimo JavaScript kodo, oblikovanja besedila, HTML komentarje in meta podatke, ki jih ne potrebujemo. Odstranili smo tudi vse enojne in dvojne narekovaje. Z uporabo iste knjižnice še izvlečemo besedilo, ki nas zanima. Besedilo z uporabo knjižnice nltk preoblikujemo v žetone. Glede na najdene indekse najdemo še izvlečke iz vsakega dokumenta posebej in vse skupaj izpišemo. Iskanje »Sistem SPO« traja približno 12 milisekund.

4. Iskanje brez inverznega indeksa

Iskanje ne uporablja inverznega indeksa in zaporedoma odpre vse datoteke, podatke obdela in združi rezultate. Branje in obdelava podatkov je enako kot pri iskanju z inverznim indeksom. Iskane besede so pretvorjene v seznam žetonov, ki je uporabljen za iskanje. Preberemo vse HTML datoteke. Iz njih s pomočjo BeautifulSoup knjižnice odstranimo JavaScript kodo,

oblikovanja besedila, HTML komentarje in meta podatke, ki jih ne potrebujemo. Odstranili smo tudi vse enojne in dvojne narekovaje. Z uporabo iste knjižnice še izvlečemo besedilo, ki nas zanima. Besedilo z uporabo knjižnice nltk preoblikujemo v žetone in iz njih odstranimo odvečne besede (angl. stopwords), ki so bile na predlaganem seznamu. Za vsako besedo poiščemo, kje v izvornem dokumentu je. Glede na najdene indekse najdemo še izvlečke iz vsakega dokumenta posebej in vse skupaj izpišemo. Iskanje »Sistem SPO« traja približno 35 sekund.

5. Podatkovna baza

Podatkovna baza vsebuje 47278 unikatnih indeksiranih besed. Tabela 1 prikazuje dokumente z največjim seštevkom frekvenc besed, tabela 2 pa najpogostejše besede v celotnem indeksu.

Ime dokumenta	**Seštevek frekvenc**
evem.gov.si/evem.gov.si.371.html	103237
podatki.gov.si/podatki.gov.si.340.html	32109
e-prostor.gov.si/e-prostor.gov.si.166.html	11202
e-prostor.gov.si/e-prostor.gov.si.147.html	9628
podatki.gov.si/podatki.gov.si.511.html	6918
evem.gov.si/evem.gov.si.398.html	5197
e-prostor.gov.si/e-prostor.gov.si.57.html	4567

Tabela 1: dokumenti z največ besedami.

Beseda	**Seštevek frekvenc**
,	53583
.	25494
:	15171
(	14946
)	14909
-	13176
podatkov	11089
slovenije	10507
republike	8583
dejavnosti	6093

Tabela 2: najpogostejše besede.

6. Poizvedbe

Rezultati poizvedb so omejeni na 5 vrstic. Zaradi lepšega prikaza poročilo ne vsebuje vseh izvlečkov dokumenta, vendar samo nekaj začetnih. Za iskanje je bil uporabljen inverzni indeks.

6.1 Poizvedba »predelovalne dejavnosti«

Rezultati so bili najdeni v osmih milisekundah. Dokument z največ zadetki vsebuje 1291 iskanih besed.

```
Results for a query: "predelovalne dejavnosti"
Results found in 8 ms.

Frequencies Document                                Snippet
----------- ----------------------------------------- ------------------------------------------------------------------------------
1291        evem.gov.si/evem.gov.si.371.html          ... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opra
75          evem.gov.si/evem.gov.si.377.html          ... Defektolog v zdravstveni dejavnosti Dekan oziroma direktor ... Dietetik v zdr
40          podatki.gov.si/podatki.gov.si.340.html    ... - NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI BREGAR ... šport CENTER INTERESNIH
40          evem.gov.si/evem.gov.si.452.html          nastavitve Druge storitvene dejavnosti , drugje nerazvrščene ... 96.090 ) / Dejav
31          evem.gov.si/evem.gov.si.653.html          ... Dovoljenje za opravljanje dejavnosti specializirane prodajalne z ... radijske
```

Slika 1: rezultat poizvedbe »predelovalne dejavnosti«.

6.2 Poizvedba »trgovina«

Rezultati so bili najdeni v šestih milisekundah. Dokument z največ zadetki vsebuje 364 iskanih besed.

```
Results for a query: "trgovina"
Results found in 6 ms.

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 z ... gl
96          evem.gov.si/evem.gov.si.651.html          ... Druga govedoreja Druga trgovina na drobno v ... specializiranih prodajalnah D
92          evem.gov.si/evem.gov.si.21.html           ... eVEM › Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina
82          podatki.gov.si/podatki.gov.si.340.html    ... A DENT , trgovina in storitve , ... . ADRIA INVESTICIJE trgovina , posredniš
14          evem.gov.si/evem.gov.si.623.html          x Sprememba nastavitve Trgovina na debelo z ... › Dejavnosti › Trgovina na debelo
```

Slika 2: rezultat poizvedbe »trgovina«.

6.3 Poizvedba »social services«

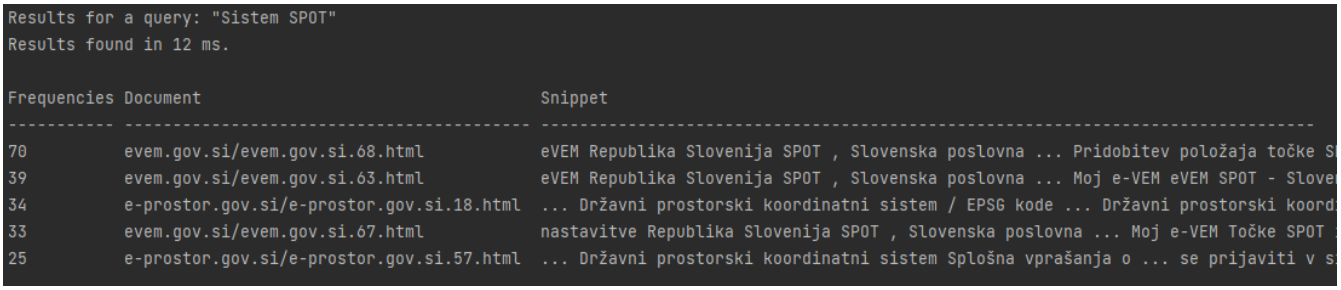
Rezultati so bili najdeni v štirih milisekundah. Dokument z največ zadetki vsebuje 5 iskanih besed.

```
Results for a query: "social services"
Results found in 4 ms.

Frequencies Document                                Snippet
----------- ----------------------------------------- ------------------------------------------------------------------------------
5           e-uprava.gov.si/e-uprava.gov.si.9.html    ... Labour , retirement Social services , health , ... relationship etc. ? Social
5           e-uprava.gov.si/e-uprava.gov.si.45.html   ... Labour , retirement Social services , health , ... relationship etc. ? Social
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 ) ...

Process finished with exit code 0
```

Slika 3: rezultat poizvedbe »social services«.

6.4 Poizvedba »Sistem SPOT«

Rezultati so bili najdeni v dvanajstih milisekundah. Dokument z največ zadetki vsebuje 70 iskanih besed.

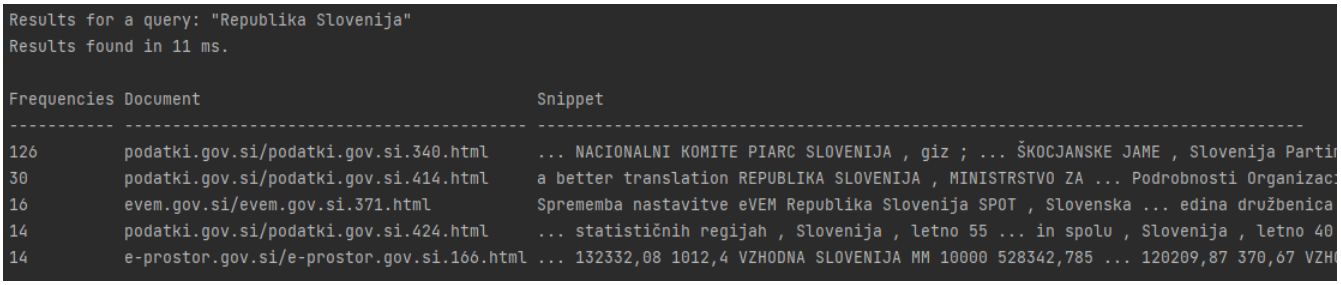


```
Results for a query: "Sistem SPOT"
Results found in 12 ms.

Frequencies Document                                Snippet
----------- ----------------------------------------- ------------------------------------------------------------------------------
70          evem.gov.si/evem.gov.si.68.html           eVEM Republika Slovenija SPOT , Slovenska poslovna ... Pridobitev položaja točke S
39          evem.gov.si/evem.gov.si.63.html           eVEM Republika Slovenija SPOT , Slovenska poslovna ... Moj e-VEM eVEM SPOT - Slove
34          e-prostor.gov.si/e-prostor.gov.si.18.html ... Državni prostorski koordinatni sistem / EPSG kode ... Državni prostorski koord
33          evem.gov.si/evem.gov.si.67.html           nastavitve Republika Slovenija SPOT , Slovenska poslovna ... Moj e-VEM Točke SPOT
25          e-prostor.gov.si/e-prostor.gov.si.57.html ... Državni prostorski koordinatni sistem Splošna vprašanja o ... se prijaviti v s
```

Slika 4: rezultat poizvedbe »Sistem SPOT«.

6.5 Poizvedba »Republika Slovenija«

Rezultati so bili najdeni v enajstih milisekundah. Dokument z največ zadetki vsebuje 126 iskanih besed.

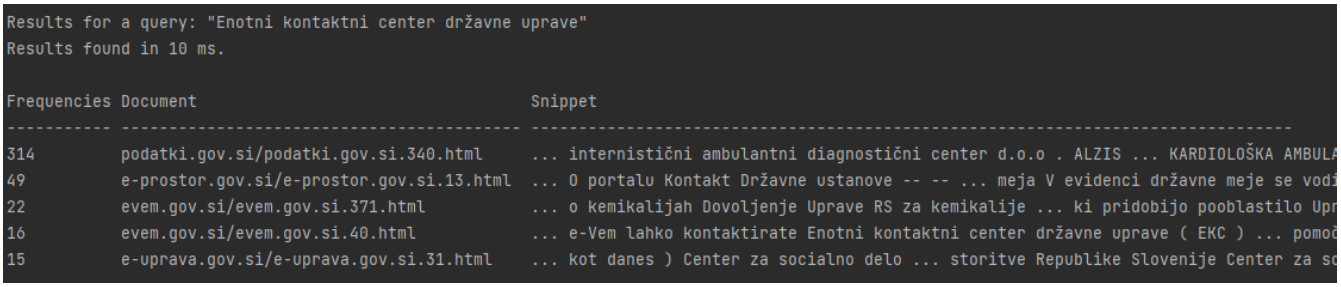


```
Results for a query: "Republika Slovenija"
Results found in 11 ms.

Frequencies Document                                Snippet
----------- ----------------------------------------- ------------------------------------------------------------------------------
126         podatki.gov.si/podatki.gov.si.340.html    ... NACIONALNI KOMITE PIARC SLOVENIJA , giz ; ... ŠKOCJANSKE JAME , Slovenija Parti
30          podatki.gov.si/podatki.gov.si.414.html    a better translation REPUBLIKA SLOVENIJA , MINISTRSTVO ZA ... Podrobnosti Organizac
16          evem.gov.si/evem.gov.si.371.html          Sprememba nastavitve eVEM Republika Slovenija SPOT , Slovenska ... edina družbenica
14          podatki.gov.si/podatki.gov.si.424.html    ... statističnih regijah , Slovenija , letno 55 ... in spolu , Slovenija , letno 40
14          e-prostor.gov.si/e-prostor.gov.si.166.html ... 132332,08 1012,4 VZHODNA SLOVENIJA MM 10000 528342,785 ... 120209,87 370,67 VZH
```

Slika 4: rezultat poizvedbe »Republika Slovenija«.

6.6 Poizvedba »Enotni kontaktni center državne uprave«

Rezultati so bili najdeni v desetih milisekundah. Dokument z največ zadetki vsebuje 314 iskanih besed.



```
Results for a query: "Enotni kontaktni center državne uprave"
Results found in 10 ms.

Frequencies Document                                Snippet
----------- ----------------------------------------- ------------------------------------------------------------------------------
314         podatki.gov.si/podatki.gov.si.340.html    ... internistični ambulantni diagnostični center d.o.o . ALZIS ... KARDIOLOŠKA AMBULA
49          e-prostor.gov.si/e-prostor.gov.si.13.html ... O portalu Kontakt Državne ustanove -- -- ... meja V evidenci državne meje se vodi
22          evem.gov.si/evem.gov.si.371.html          ... o kemikalijah Dovoljenje Uprave RS za kemikalije ... ki pridobijo pooblastilo Upr
16          evem.gov.si/evem.gov.si.40.html           ... e-Vem lahko kontaktirate Enotni kontaktni center državne uprave ( EKC ) ... pomoč
15          e-uprava.gov.si/e-uprava.gov.si.31.html   ... kot danes ) Center za socialno delo ... storitve Republike Slovenije Center za so
```

Slika 4: rezultat poizvedbe »Enotni kontaktni center državne uprave«.