IEPS: Data indexing and retrieval

Authors: Julijan Jug, Jaka Jenko

I. Introduction

For the third assignment of the Web Information Extraction and Retrieval course, we implemented data indexing and retrieval system. First we constructed an index of 1416 web pages and execute some queries on it. Than we compared the results to manually searching the files.

II. Data preprocessing and indexing

For indexing part we created a database with tables IndexWord and Posting. IndexWord table contains all words found in the web pages while Posting table contains the locations of these words.

For the preprocessing we used BeautifulSoup library with which we parsed the HTML, removed script, iframe, and style tags and extracted the text. The text was converted to lower case and tokenized using nltk library. We removed the tokens that were stopwords or punctuation marks. For the stopwords set we used the set provided in the assignment instructions.

The tokens were stored in an dictionary with their indices and later inserted in the database.

III. Data retrieval with inverted index

For the data retrieval using inverted index we executed a single sql query. Firstly we preprocessed the search query (tokenization, stopwords removal). The SQL query returns a list of items with documentName, frequency and indices and looks like this:

```
    SELECT p.documentname, sum(p.frequency) as
frequency, string_agg(p.indexes, ',') AS indexes
            FROM ieps3.indexword  w INNER JOIN
iepsЗ.posting p ON w.word = p.word
            WHERE w.word IN "query"
            group by p.documentname
            order by 2 desc
```

The results were then simply formated and extracted snippets from the original document. The snippets include 3 words before and after the searched term. Punctuation characters are not included in this snippets.

The results for all search queries are listed in Appendix A. we executed these queries:

- predelovalne dejavnosti
- trgovina
- social services
- zaporna kazen
- fakulteta za računalništvo in informatiko
- republika slovenija

IV. Data retrieval without inverted index

For data retrieval without the inverted index we used the same preprocessing procedure as for the inverted index version. Next we simply iterated through all of the web page files and stored the indexes of the word occurrences in an dictionary. At the end we sorted the dictionary and formatted the output.

The results for all search queries are listed in Appendix B.

V. Database description

This are some facts about the constructed index database:

- The database includes 49.066 distinct words.
- Total words count is 811.725.
- Most frequent word is 'podatkov' with 10.502 occurrences.
- Second most frequent word is 'slovenije' with 9.856 occurrences.
- Third most frequent word is 'republike' with 8.583 occurrences.
- The word with most occurrences in a single document is 'proizvodnja' in 'evem.gov.si.371.html', with 2.266 occurrences.
- Word that appears in the most document is 'pogoji' with 1.398 documents.
- Document with the most diverse vocabulary is 'evem.gov.si.371.html' with 13.184 words.

VI. Conclusion

We have successfully implemented both approaches to data retrieval from web pages. If we now compare the results we see that the inverted index outperform the basic search. Queries using inverted index take at most 23ms, usually less then 10ms. And basic file iteration searches takes much longer, on average around 2 minutes (120.000ms). There does not seem to be any differences in the outputed query results. The benefit of having an inverted index is really apparent here. Although the down side is that the index has to be first constructed and then updated on aregular bases. Basic search method does not have this problem.

References

Appendix A
Inverted index results

Results for a query: "predelovalne dejavnosti"

Results found in 0.78ms

Frequencies	Document	Snippet
1288	evem.gov.si/evem.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti V iskalnik vpišite ... 645 od 645 dejavnosti Izpisanih je od ... Izpisanih je od dejavnosti A KMETIJSTVO IN ... pogojih za opravljanje dejavnosti · P ...
75	evem.gov.si/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 ... II v zdravstveni dejavnosti Laboratorijski tehnik Lad ...
40	podatki.gov.si/podatki.gov.si.340.html	... KALAN 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 DENTIM zobozdravstvo ... DERMA ...
39	evem.gov.si/evem.gov.si.452.html	... nastavitve Druge storitvene dejavnosti drugje nerazvrščene 96.090 ... drugje nerazvrščene 96.090 Dejavnosti eVEM Republika Slovenija ... e-VEM eVEM>Dejavnosti>Druge storitvene dejavnosti drugje nerazvrščene 96.090 ... 96.090 Druge storitvene dej ...
31	evem.gov.si/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 ... za izvajanje sevalne dejavnosti Dovoljenje za izvaj ...

Results for a query: "trgovina"

Results found in 4.1ms

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	... organizacij gl 46.110 trgovina na debelo s ... juh gl 10.890 trgovina na debelo z ... ipd. gl 10.890 trgovina na debelo s ... jedmi gl 46.380 trgovina na drobno s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ... z nevarnimi kemikalij ...
94	evem.gov.si/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 ... Nepremičninsko posredovanje Nespe ...
92	evem.gov.si/evem.gov.si.21.html	... Moj e-VEM eVEM>Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno zunaj ... tržnic 47.990 Nespecializirana trgovina na debelo Trgovina ... trgovina na debe ...
82	podatki.gov.si/podatki.gov.si.340.html	... d.o.o A DENT trgovina in storitve d.o.o ... d.o.o ADRIA INVESTICIJE trgovina posredništvo storitve in ... storitve d.o.o AHATSERVIS trgovina in storitve d.o.o ... vzdrževanje d.o.o ALBA trgovina in proizvodnja d.o.o ... proizvodnja storitve in t ...
13	evem.gov.si/evem.gov.si.623.html	... varovanja zasebnosti.xSpremememba nastavitve Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... porabe Sem spada trgovina na debelo z ... plutovinastimi izdelki ipd trgovina na debelo s ... in deli zanja trgovina na debelo s . ...

Results for a query: "social services"

Results found in 3.0ms

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.45.html	... Labour retirement Social services health death Taxes ... relationship etc. Social services health death How ... culture Labour retirement Social services health death ... employment relationship etc. Social services health death ... I obtain fin ...
5	e-uprava.gov.si/e-uprava.gov.si.9.html	... Labour retirement Social services health death Taxes ... relationship etc. Social services health death How ... culture Labour retirement Social services health death ... employment relationship etc. Social services health death ... I obtain fin ...
1	evem.gov.si/evem.gov.si.661.html	... Records and Related Services AJPES and the ...
1	podatki.gov.si/podatki.gov.si.340.html	... recreation and spa services ltd. TERME MARIBOR ...

Results for a query: "zaporna kazen"

Results found in 6.1ms

Frequencies	Document	Snippet
7	evem.gov.si/evem.gov.si.371.html	... gospodarstvo na zaporno kazen najmanj treh mesecev ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni v kazenskem ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni zaposlen na ... ...
4	evem.gov.si/evem.gov.si.227.html	... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni v kazenskem ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni zaposlen na ...
1	evem.gov.si/evem.gov.si.22.html	... pravnomočno obsojena na kazen zapora zaradi kaznivega ...
1	evem.gov.si/evem.gov.si.23.html	... pravnomočno obsojena na kazen zapora zaradi določenih ...

Results for a query: "fakulteta za računalništvo in informatiko"

Results found in 15.0ms

Frequencies	Document	Snippet
42	podatki.gov.si/podatki.gov.si.340.html	... INSTITUTUM STUDIORUM HUMANITATIS FAKULTETA ZA PODIPLOMSKI HUMANISTIČNI ... d.o.o Evropska pravna fakulteta EVROPSKI CENTER ZA ... MARIBORU Evro-sredozemska univerza FAKULTETA ZA DIZAJN samostojni ... Univerze na Primorskem FAKULTETA ZA DRŽAVNE I ...
12	e-prostor.gov.si/e-prostor.gov.si.150.html	... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakultet ...
9	podatki.gov.si/podatki.gov.si.14.html	... upravo Združenje za informatiko in telekomunikacije IKT ... za računalništvo in informatiko V okviru Festivala ... za računalništvo in informatiko v Ljubljani potekala ... za računalništvo in informatiko Nadaljujte z branjem ... za računalništvo ...
9	podatki.gov.si/podatki.gov.si.12.html	... upravo Združenje za informatiko in telekomunikacije IKT ... za računalništvo in informatiko V okviru Festivala ... za računalništvo in informatiko v Ljubljani potekala ... za računalništvo in informatiko Nadaljujte z branjem ... za računalništvo ...
8	podatki.gov.si/podatki.gov.si.534.html	... študentov Fakultete za računalništvo in informatiko | ... študentov Fakultete za računalništvo in informatiko Predstavitev ... študentov Fakultete za računalništvo in informatiko V ... Ljubljani Fakultete za računalništvo in informatiko so ... z ...

Results for a query: "republika slovenija"

Results found in 23.15ms

Frequencies	Document	Snippet
126	podatki.gov.si/podatki.gov.si.340.html	... NACIONALNI KOMITE PIARC SLOVENIJA giz v angleškem ... PARK ŠKOCJANSKE JAME Slovenija Partim finančna in ... informiranje d.o.o RADIOTELEVIZIJA SLOVENIJA javni zavod Ljubljana ... v likvidaciji REPUBLIKA SLOVENIJA REPUBLIKA SLOVENIJA MINISTRSTVO ...
30	podatki.gov.si/podatki.gov.si.414.html	... better translation REPUBLIKA SLOVENIJA MINISTRSTVO ZA DELO ... Podrobnosti Organizacija REPUBLIKA SLOVENIJA MINISTRSTVO ZA DELO ... ENAKE MOŽNOSTI REPUBLIKA SLOVENIJA MINISTRSTVO ZA DELO ... 149 ogledov REPUBLIKA SLOVENIJA MINISTRSTVO ZA DELO .. ...
16	evem.gov.si/evem.gov.si.371.html	... nastavitve eVEM Republika Slovenija SPOT Slovenska poslovna ... družbenica je Republika Slovenija Službe se zagotavljajo ... je ustanovila Republika Slovenija ter na podlagi ... je ustanovila Republika Slovenija ter na podlagi ... je ustanovila ...
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 VZHODNA SLOVENIJA MM 10000 540391,312 ... 81709,09 374,87 VZHODNA SLOVENIJA MM 10000 521281,03 ... 33984,42 491,7 VZHODNA SLOVENIJA MM 10000 408572,817 ... 55648,42 741, ...
14	podatki.gov.si/podatki.gov.si.424.html	... in statističnih regijah Slovenija letno 55 ogledov ... poškodbe in spolu Slovenija letno 40 ogledov ... splošnih zobnih ambulantah Slovenija letno 32 ogledov ... splošnih zobnih ambulantah Slovenija Nadaljujte z branjem ... spolu in starosti Slo ...

Appendix B
Basic search results

Results for a query: "predelovalne dejavnosti"

Results found in 115298.8ms

Frequencies	Document	Snippet
1288	evem.gov.si/evem.gov.si.371.html	... iskanje ustrezne šifre dejavnosti /storitve in informacij ... pogojih za opravljanje dejavnosti V iskalnik vpišite ... 645 od 645 dejavnosti Izpisanih je od ... Izpisanih je od dejavnosti A KMETIJSTVO IN ... pogojih za opravljanje dejavnosti · P ...
75	evem.gov.si/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 ... II v zdravstveni dejavnosti Laboratorijski tehnik Lad ...
40	podatki.gov.si/podatki.gov.si.340.html	... KALAN 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 DENTIM zobozdravstvo ... DERMA ...
39	evem.gov.si/evem.gov.si.452.html	... nastavitve Druge storitvene dejavnosti drugje nerazvrščene 96.090 ... drugje nerazvrščene 96.090 Dejavnosti eVEM Republika Slovenija ... e-VEM eVEM>Dejavnosti>Druge storitvene dejavnosti drugje nerazvrščene 96.090 ... 96.090 Druge storitvene dej ...
31	evem.gov.si/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 ... za izvajanje sevalne dejavnosti Dovoljenje za izvaj ...

Results for a query: "trgovina"

Results found in 102237.7ms

Frequencies	Document	Snippet
364	evem.gov.si/evem.gov.si.371.html	... organizacij gl 46.110 trgovina na debelo s ... juh gl 10.890 trgovina na debelo z ... ipd. gl 10.890 trgovina na debelo s ... jedmi gl 46.380 trgovina na drobno s ... Skladiščenje nevarnih kemikalij Trgovina na debelo z ... z nevarnimi kemikalij ...
94	evem.gov.si/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 ... Nepremičninsko posredovanje Nespe ...
92	evem.gov.si/evem.gov.si.21.html	... Moj e-VEM eVEM>Področja Trgovina Tu boste našli ... Seznam dejavnosti Druga trgovina na drobno v ... nespecializiranih prodajalnah Druga trgovina na drobno zunaj ... tržnic 47.990 Nespecializirana trgovina na debelo Trgovina ... trgovina na debe ...
82	podatki.gov.si/podatki.gov.si.340.html	... d.o.o A DENT trgovina in storitve d.o.o ... d.o.o ADRIA INVESTICIJE trgovina posredništvo storitve in ... storitve d.o.o AHATSERVIS trgovina in storitve d.o.o ... vzdrževanje d.o.o ALBA trgovina in proizvodnja d.o.o ... proizvodnja storitve in t ...
13	evem.gov.si/evem.gov.si.623.html	... varovanja zasebnosti.xSpremememba nastavitve Trgovina na debelo z ... izdelki široke porabe Trgovina na debelo z ... porabe Sem spada trgovina na debelo z ... plutovinastimi izdelki ipd trgovina na debelo s ... in deli zanja trgovina na debelo s . ...

Results for a query: "social services"

Results found in 115452.4ms

Frequencies	Document	Snippet
5	e-uprava.gov.si/e-uprava.gov.si.9.html	... culture Labour retirement Social services health death ... Labour retirement Social services health death Taxes ... employment relationship etc. Social services health death ... relationship etc. Social services health death How ... I obtain fin ...
5	e-uprava.gov.si/e-uprava.gov.si.45.html	... culture Labour retirement Social services health death ... Labour retirement Social services health death Taxes ... employment relationship etc. Social services health death ... relationship etc. Social services health death How ... I obtain fin ...
1	evem.gov.si/evem.gov.si.661.html	... Records and Related Services AJPES and the ...
1	podatki.gov.si/podatki.gov.si.340.html	... recreation and spa services ltd. TERME MARIBOR ...

Results for a query: "zaporna kazen"

Results found in 90188.7ms

Frequencies	Document	Snippet
7	evem.gov.si/evem.gov.si.371.html	... gospodarstvo na zaporno kazen najmanj treh mesecev ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni v kazenskem ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni zaposlen na ... ...
4	evem.gov.si/evem.gov.si.227.html	... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni v kazenskem ... katero je zagrožena kazen zapora šestih mesecev ... mesecev ali hujša kazen ni zaposlen na ...
1	evem.gov.si/evem.gov.si.23.html	... pravnomočno obsojena na kazen zapora zaradi določenih ...
1	evem.gov.si/evem.gov.si.22.html	... pravnomočno obsojena na kazen zapora zaradi kaznivega ...

Results for a query: "fakulteta za računalništvo in informatiko"

Results found in 83304.0ms

Frequencies	Document	Snippet
42	podatki.gov.si/podatki.gov.si.340.html	... INSTITUTUM STUDIORUM HUMANITATIS FAKULTETA ZA PODIPLOMSKI HUMANISTIČNI ... d.o.o. Center za računalništvo v mehaniki kontinuuma ... d.o.o Evropska pravna fakulteta EVROPSKI CENTER ZA ... MARIBORU Evro-sredozemska univerza FAKULTETA ZA DIZAJN sam ...
12	e-prostor.gov.si/e-prostor.gov.si.150.html	... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakulteta za gradbeništvo in ... Univerza v Ljubljani Fakultet ...
9	podatki.gov.si/podatki.gov.si.12.html	... upravo Združenje za informatiko in telekomunikacije IKT ... študentov Fakultete za računalništvo in informatiko V ... za računalništvo in informatiko V okviru Festivala ... na Fakulteti za računalništvo in informatiko v ... za računalništvo in i ...
9	podatki.gov.si/podatki.gov.si.14.html	... upravo Združenje za informatiko in telekomunikacije IKT ... študentov Fakultete za računalništvo in informatiko V ... za računalništvo in informatiko V okviru Festivala ... na Fakulteti za računalništvo in informatiko v ... za računalništvo in i ...
8	podatki.gov.si/podatki.gov.si.534.html	... študentov Fakultete za računalništvo in informatiko | ... za računalništvo in informatiko | OPSI Odprti ... študentov Fakultete za računalništvo in informatiko Predstavitev ... za računalništvo in informatiko Predstavitev seminarskih nalog ... š ...

Results for a query: "republika slovenija"

Results found in 127778.2ms

Frequencies	Document	Snippet
126	podatki.gov.si/podatki.gov.si.340.html	... NACIONALNI KOMITE PIARC SLOVENIJA giz v angleškem ... PARK ŠKOCJANSKE JAME Slovenija Partim finančna in ... informiranje d.o.o RADIOTELEVIZIJA SLOVENIJA javni zavod Ljubljana ... d.o.o v likvidaciji REPUBLIKA SLOVENIJA REPUBLIKA SLOVENIJA ... v ...
30	podatki.gov.si/podatki.gov.si.414.html	... a better translation REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... better translation REPUBLIKA SLOVENIJA MINISTRSTVO ZA DELO ... Organizacije Podrobnosti Organizacija REPUBLIKA SLOVENIJA MINISTRSTVO ZA ... Podrobnosti Organizacija REPUBLIKA SLOVENIJA ...
16	evem.gov.si/evem.gov.si.371.html	... zasebnosti.xSprememba nastavitve eVEM Republika Slovenija SPOT Slovenska ... nastavitve eVEM Republika Slovenija SPOT Slovenska poslovna ... edina družbenica je Republika Slovenija Službe se ... družbenica je Republika Slovenija Službe se zagota ...
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 VZHODNA SLOVENIJA MM 10000 540391,312 ... 81709,09 374,87 VZHODNA SLOVENIJA MM 10000 521281,03 ... 33984,42 491,7 VZHODNA SLOVENIJA MM 10000 408572,817 ... 55648,42 741, ...
14	podatki.gov.si/podatki.gov.si.424.html	... in statističnih regijah Slovenija letno 55 ogledov ... poškodbe in spolu Slovenija letno 40 ogledov ... splošnih zobnih ambulantah Slovenija letno 32 ogledov ... splošnih zobnih ambulantah Slovenija Nadaljujte z branjem ... spolu in starosti Slo ...