Indexer

Timotej Kovač

I. Introduction

In this paper we describe the results gathered while testing two approaches in finding relevant documents to a search query. The two approaches used were the invert index and a manual checking one. We describe the implementation of both of them, pre-processing steps that were taken before and the results that we got.

II. Indexer implementation

A. Data processing with indexing

All of the data pre-processing was done in file `processing.py`. There html content is read directly from file and a DOM tree is constructed with the help of 'lxml.html' library. This is then cleaned of all scripts and nav and footer sub trees. After that only the body is forwarded to the next step of the extraction process. The result is then tokenized with the help of 'nltk' library, special characters are removed with the help of regular expressions library 're' and the resulting words are checked with the stop words black list. After this process what remains are only the words that have some meaning and are deemed useful for indexing. This was done for all of the websites that were available to us.

After that the unique words were gathered and tuples were created in order to insert them into the database. First two tables `IndexWord` and `Posting` were created. After that all of the data was inserted with two insert statements. This was necessary as individual insert statements which were first executed during the word tokenization proved to be inefficient.

B. Data retrieval with inverted index

For the data retrieval part with the help of mySQL database we started by retrieving the user input query words. First we checked if the database was present and if necessary we ran the data processing step with indexing (as described in the previous section). After that multiple select statements were executed to gather the frequency, document and indices of particular words appearing in the previously retrieved html pages. After that all of the data for individual query words were merged and sorted from the most frequent ones to least. Snippets were extracted to be attached to the end result. At the end the retrieved data was formatted and sent to the standard output.

C. Data retrieval without inverted index

For this part we used some of the same approaches as when we used the inverted index database. The main difference here was that we didn't use the database so we didn't have to create it and fill it with data. Nevertheless we still had to gain information about the appearance of words in html pages. So we started with this extraction and then manually checked each result if any of the query words appeared in it. If this was the case we saved the frequency, document and indices and proceeded with the search. After searching all of the documents we proceeded with merging the results, adding snippets, sorting them, formatting the output and sending everything to the standard output.

III. Database description

From the html pages we have extracted a total of 48.081 unique words that were then saved to IndexWord table. Word with the highest frequency on a given site was 'proizvodnja' with a frequency of 2266 on site evem.gov.si. Word with the highest frequency over all the sites was 'slovenije' with a frequency of 9105.

IV. Results

From the results it is apparent that the approach using the inverted index data is far superior to the one that manually checks all of the documents. Time needed for each approach can be seen in table IV.

query/method	inverted index	manual checking
predelovalne dejavnosti	19ms (20ms)	92085ms
trgovina	4ms (5ms)	92414ms
social services	69ms (5ms)	92335ms
robot	124ms (5ms)	92516ms
davčna olajšava 140ms (5ms)	93280ms	
podatki	24ms (20ms)	93650ms

Table I
Comparison of time needed to gain a query result using the inverted index and manual checking methods. In brackets is the time needed only for the SQL query without any post-processing.

From the table it can be seen that manual checking roughly takes the same amount of time regardless of the input. That is because it must check all of the document as it does not know which contain any of the query words. The inverted index method takes a much smaller amount of time. This time does vary a lot but always performs in under 200ms while the manual approach takes a minute and a half.

The actual results gathered using the inverted index for the above mentioned queries can be seen in the following subsections. The outputs have been cut to only keep the top 10 results and maximum of 5 snippets.

A. *predelovalne dejavnosti*

Result of query 'predelovalne dejavnosti'.

```
Results for a query: "predelovalne dejavnosti"

    Results found in 1021ms.

    Frequencies Document                                Snippet
    ----------- ----------------------------------------- ----------------------------------------------------------------
    1287            evem.gov.si/evem.gov.si.371.html            vir ministrstvo infrastrukturo predelovalne dejavnosti 10 ... tehnologijo
    32 raznovrstne predelovalne dejavnosti 32110 ... 32990 drugje nerazvrščene predelovalne dejavnosti spada ... ustrezne postavke področja
    predelovalne dejavnosti predelava ... iskanje ustrezne šifre dejavnosti storitve informacij
    74          evem.gov.si/evem.gov.si.377.html            straže defektolog zdravstveni dejavnosti dekan direktor ... detektiv dietetik
    zdravstveni dejavnosti dimnikar diplomirana ... laboratorijski sodelavec zdravstveni dejavnosti laboratorijski sodelavec ...
    laboratorijski sodelavec zdravstveni dejavnosti laboratorijski tehnik ... kuhar logoped zdravstveni dejavnosti magister farmacije
    40          podatki.gov.si/podatki.gov.si.340.html          kalan nosilec dopolnilne dejavnosti kmetiji bregar ... šport center
    interesnih dejavnosti ptuj center ... center šolskih obšolskih dejavnosti center urbane ... dentiko zobozdravstvene zdravstvene
    dejavnosti doo dentim ... derma san zdravstvene dejavnosti prodaja storitve
    36          evem.gov.si/evem.gov.si.452.html            prijava evemdejavnostidruge storitvene dejavnosti drugje nerazvrščene ...
    nerazvrščene 96090 storitvene dejavnosti drugje nerazvrščene ... skd šifra zajema dejavnosti storitve predpisani ... pogoji začetek
    opravljanje dejavnosti predpisi pogoji ... razvoj tehnologijo lista dejavnosti običajno opravljajo
    30          evem.gov.si/evem.gov.si.653.html            licenca dovoljenje opravljanje dejavnosti specializirane prodajalne ...
    izvajanje radijske televizijske dejavnosti dovoljenje izvajanje ... dovoljenje izvajanje sevalne dejavnosti dovoljenje izvajanje ...
    dovoljenje izvajanje sevalne dejavnosti dovoljenje izvajanje ... dovoljenje izvajanje sevalne dejavnosti državi dovoljenje
    28          evem.gov.si/evem.gov.si.398.html            aktivnostmi usmerjene opravljanje dejavnosti npr pripravljalna ... nabavah
    namene opravljanja dejavnosti ipd obdobju ... 12 mesecev opravljanjem dejavnosti sloveniji presegli ... 11000 uporabljali opravljanje
    dejavnosti identificirati namene ... nabavah namene opravljanja dejavnosti ipd vloga
    28          evem.gov.si/evem.gov.si.72.html         evemvodenje podjetjadavkidavek dohodka dejavnosti davek dohodka ... dejavnosti
    davek dohodka dejavnosti začnete opravljati ... zavezanec davek dohodka dejavnosti dohodek dejavnosti ... dohodka dejavnosti dohodek
    dejavnosti šteje dohodek ... neodvisnim samostojnim opravljanjem dejavnosti glede namen
    20          evem.gov.si/evem.gov.si.442.html            nego telesa 96040 dejavnosti nego telesa ... skd šifra zajema dejavnosti
    storitve predpisani ... pogoji začetek opravljanje dejavnosti predpisi pogoji ... predpisi pogoji lista dejavnosti običajno opravljajo
    ... opravljajo obrtni način dejavnosti uvrščene listo
    17          evem.gov.si/evem.gov.si.28.html         zavodov opravljanje gospodarske dejavnosti lastnosti zasebnega ... pravne osebe
    posamezne dejavnosti posamezne vrste ... zadoščajo ustanovitev opravljanje dejavnosti zavoda ime ... dobiček opravljanja nepridobitne
    dejavnosti obdavči slabosti ... zasebnega zavoda število dejavnosti ustanovi zavod
    15          evem.gov.si/evem.gov.si.460.html            prijava evemdejavnostidrugje nerazvrščene predelovalne dejavnosti 32990 ...
    32990 drugje nerazvrščene predelovalne dejavnosti 32990 ... evemdejavnostidrugje nerazvrščene predelovalne dejavnosti 32990 drugje ...
```

B. *trgovina*

Result of query 'trgovina'.

```
Results for a query: "trgovina"

    Results found in 219ms.

    Frequencies Document                                Snippet
    ----------- ---------------------------------------- ----------------------------------------------------------------
    364         evem.gov.si/evem.gov.si.371.html            organizacij gl 46110 trgovina debelo kmetijskimi ... juh gl 10890 trgovina
    debelo mesnimi ... ipd gl 10890 trgovina debelo pripravljenimi ... jedmi gl 46380 trgovina drobno pripravljenimi ... skladiščenje
    nevarnih kemikalij trgovina debelo nevarnimi
    94          evem.gov.si/evem.gov.si.651.html            trgu dozimetrija govedoreja trgovina drobno specializiranih ... drobno
    specializiranih prodajalnah trgovina drobno nespecializiranih ... drobno nespecializiranih prodajalnah trgovina drobno specializiranih
    ... specializiranih prodajalnah živili trgovina drobno prodajaln ... nepremičninsko posredovanje nespecializirana trgovina debelo
    nespecializirana
    92          evem.gov.si/evem.gov.si.21.html         sklad prijava evempodročja trgovina našli informacije ... razvija seznam dejavnosti
    trgovina drobno nespecializiranih ... drobno nespecializiranih prodajalnah trgovina drobno prodajaln ... tržnic 47990 nespecializirana
    trgovina debelo trgovina ... nespecializirana trgovina debelo trgovina drobno stojnicah
    82          podatki.gov.si/podatki.gov.si.340.html          storitve doo dent trgovina storitve doo ... doo adria investicije trgovina
    posredništvo storitve ... storitve doo ahatservis trgovina storitve doo ... vzdrževanje doo alba trgovina proizvodnja doo ... almaja
    proizvodnja storitve trgovina doo alma
    12          evem.gov.si/evem.gov.si.623.html            izdelki široke porabe trgovina debelo izdelki ... široke porabe spada trgovina
    debelo lesenimi ... plutovinastimi izdelki ipd trgovina debelo kolesi ... kolesi deli zanja trgovina debelo pisarniškimi ...
    potrebščinami knjigami časopisi trgovina debelo usnjenimi
    11          evem.gov.si/evem.gov.si.329.html            materialom sanitarno opremo trgovina debelo lesom ... sanitarno opremo spada
    trgovina debelo neobdelanim ... debelo neobdelanim lesom trgovina debelo proizvodi ... primarne obdelave lesa trgovina debelo
    premaznimi ... sredstvi laki barvami trgovina debelo tapetami
    11          evem.gov.si/evem.gov.si.630.html            nerazvrščenimi predmeti gospodinjstvo trgovina drobno specializiranih ...
    gospodinjstvo spada specializirana trgovina drobno pohištvom ... drobno pohištvom specializirana trgovina drobno svetili ... opremo
    razsvetljavo specializirana trgovina drobno gospodinjsko ... porcelana keramike specializirana trgovina drobno izdelki
    9           evem.gov.si/evem.gov.si.320.html            materialom napravami ogrevanje trgovina debelo kovinskimi ... napravami
    ogrevanje spada trgovina debelo kovinskimi ... kovinskimi izdelki ključavnicami trgovina debelo izdelki ... debelo izdelki pritrjevanje
    trgovina debelo parnimi ... debelo parnimi kotli trgovina debelo sanitarno
    9           evem.gov.si/evem.gov.si.327.html            debelo napravami opremo trgovina debelo napravami ... napravami opremo spada
    trgovina debelo transportno ... motornih koles koles trgovina debelo industrijskimi ... debelo industrijskimi roboti trgovina debelo
    žico ... opremo industrijsko rabo trgovina debelo električnimi
    9           evem.gov.si/evem.gov.si.622.html            električnimi gospodinjskimi napravami trgovina debelo električnimi ...
```

C. *social services*

Result of query 'social services'.

```
Results for a query: "social services"

    Results found in 94ms.

    Frequencies Document                                Snippet
    ----------- ---------------------------------------- ----------------------------------------------------------------
    5           e-uprava.gov.si/e-uprava.gov.si.45.html         culture labour retirement social services health ... employment
    relationship etc social services health ... can obtain financial social assistance how ... labour retirement social services health
    death ... relationship etc social services health death
    5           e-uprava.gov.si/e-uprava.gov.si.9.html          culture labour retirement social services health ... employment
    relationship etc social services health ... can obtain financial social assistance how ... labour retirement social services health
    death ... relationship etc social services health death
    1           evem.gov.si/evem.gov.si.661.html             records and related services ajpes and
    1           podatki.gov.si/podatki.gov.si.340.html          recreation and spa services ltd terme
```

D. *robot*

Result of query 'robot'.

```
Results for a query: "robot"

    Results found in 139ms.

    Frequencies Document                                  Snippet
    ----------- ----------------------------------------- ------------------------------------------------------------
    1           e-uprava.gov.si/e-uprava.gov.si.1.html          podrobnosti prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.10.html         podjetje prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.11.html         podjetje prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.12.html         podjetje prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.13.html         podjetje prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.14.html         podjetje prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.15.html         otroka prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.16.html         slovenije prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.17.html         podatke prosimo izpolnite robot aktualno novice
    1           e-uprava.gov.si/e-uprava.gov.si.18.html         postopkih prosimo izpolnite robot aktualno novice
```

E. *davčna olajšava*

Result of query 'davčna olajšava'.

```
Results for a query: "davčna olajšava"

    Results found in 282ms.

    Frequencies Document                                  Snippet
    ----------- ----------------------------------------- ------------------------------------------------------------
    7           evem.gov.si/evem.gov.si.77.html         podlagi normiranih odhodkov davčna obveznost prihodki ... izračun davčne obveznosti
    davčna obveznost ugotavlja ... ugotavlja podlagi lestvice davčna obveznost 49700 ... davčna obveznost 49700 davčna obveznost znaša ...
    20 davčne osnove davčna osnova 23200
    3           evem.gov.si/evem.gov.si.646.html            razliko kapitalskih družb davčna obveznost družbenike ... dejanskega izplačila
    dobička davčna obveznost družbenike ... naslov pravne osebe davčna številka tuje
    3           evem.gov.si/evem.gov.si.7.html          39 50 20 davčna osnova ugotavlja ... davčni register dodeljena davčna številka
    davčna ... dodeljena davčna številka davčna številka identifikacijska
    3           evem.gov.si/evem.gov.si.72.html         prihodkov dejanskih odhodkov davčna osnova dobiček ... prihodkov normiranih
    odhodkov davčna osnova ugotovi ... davčnem letu ugotavlja davčna osnova zavezancu
    2           evem.gov.si/evem.gov.si.398.html            ddv računu navedena davčna številka davčnega ... naslov matična številka davčna
    številka dejavnosti
    2           evem.gov.si/evem.gov.si.404.html            pomeni status normiranca davčna osnova dohodka ... davčnem letu ugotavlja
    davčna osnova zavezancu
    2           evem.gov.si/evem.gov.si.656.html            storitve navedbo obrnjena davčna obveznost primeru ... navesti klavzulo
    obrnjena davčna obveznost poenostavljeni
    2           evem.gov.si/evem.gov.si.8.html          davčni register dodeljena davčna številka davčna ... dodeljena davčna številka
    davčna številka identifikacijska
    2           evem.gov.si/evem.gov.si.9.html          davčnem letu ugotavlja davčna osnova zavezancu ... stroški davčne olajšave davčna
    izguba priznajo
    2           podatki.gov.si/podatki.gov.si.134.html          ime ime priimek davčna številka rojstva ... ime priimek emšo davčna
    številka spol
```

F. podatki

Result of query 'podatki'.

```
Results for a query: "podatki"

    Results found in 774ms.

    Frequencies Document                                  Snippet
    ----------- ------------------------------------------ ----------------------------------------------------------------
    27          e-prostor.gov.si/e-prostor.gov.si.57.html           občin zavihkom brezplačni podatki našli povezavo ... povezavo
    aplikacijo egeodetski podatki egp pomočjo ... dostopne geodetske podatke podatki občinah dostop ... nalog izobraževalnega procesa
    podatki brezplačni podatke ... dostopni zavihka brezplačni podatki potrebno izpolniti
    25          e-prostor.gov.si/e-prostor.gov.si.170.html          podatke zemljiškega katastra podatki lastnikih podatke ... podatke
    katastra stavb podatki upravljavcih lastnikih ... inpodatke registra nepremičnin podatki upravljavcih lastnikihvsi ... upravljavcih
    lastnikihvsi ostali podatki geodetskih evidenc ... podatkovstruktura veljavnost podatkov podatki geodetske uprave
    25          e-prostor.gov.si/e-prostor.gov.si.7.html            podatke zemljiškega katastra podatki lastnikih podatke ... podatke
    katastra stavb podatki upravljavcih lastnikih ... inpodatke registra nepremičnin podatki upravljavcih lastnikihvsi ... upravljavcih
    lastnikihvsi ostali podatki geodetskih evidenc ... podatkovstruktura veljavnost podatkov podatki geodetske uprave
    15          podatki.gov.si/podatki.gov.si.437.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.184.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.230.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.265.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.277.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.280.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
    strukturirani podatki lastniškem formatu
    14          podatki.gov.si/podatki.gov.si.283.html          povezava seznam novosti podatki katalog ijz ... oceno 3 povezani podatk
    vsebujejo uri ... npr rdf povezljivi podatki vsebujejo naslove ... npr rdf strukturirani podatki odprtem formatu ... npr csv
```

V. Conclusion

We were successful in implementing the invert index approach as well as the manual checking one. The results we got are in favor to the invert index one as it is clearly much faster and far less demanding as the manual checking one. With further improvements to database structure and better pre-processing steps we could improve the time needed for a search result execution.