# Web Information Extraction and Retrieval Programming Assignment 3

Martin Arsovski, Maja Nikoloska, Emil Batakliev

May 2021

## 1 Introduction

After the first two projects, we continue implementing a new web crawler for our third programming assignment. For this project we were given a file of 1416 web pages which were crawled from four different domains. The task consisted of two parts:

1. Data processing and indexing
2. Data retrieval

In the following chapters, we will take a look at the methods used and the results obtained from those methods.

## 2 Data Processing and Indexing

Since we have four sites, the data processing begins by extracting the text page by page, from each site. We first use tokenization, then we have to check whether those words, converted to lowercase, are the same as one of our stopwords and if not, add them to our database. The stopwords that need to be used for this assignment have been predetermined. What we did is add some stopwords or rather some punctuation marks we felt needed to be added to the list. The stopwords added are:

1. "."
2. "-"
3. "–"
4. "/"
5. "("
6. ")"

7. "?"
8. "!"
9. ","
10. ":"
11. ";"

Within our database we fill out four parameters:

1. The word added.
2. The site where the word was found.
3. The frequency of the word.
4. The index where the word was found.

The results show that after filling our database we have 47305 words and 391767 postings.

# 3 Data Retrieval

For the data retrieval part we used two methods in order to retrieve our data. The first method is the "SQLite search" method or the "Inverted index" method. Here, we use the database which we previously filled with the first part (Data processing and indexin) and simply retrieve its data.

This task is created in order for us to learn that using a database is much faster than simply searching for the words each time through each page. In order for us to notice that difference, we also had to implement a "Basic search" method, which goes page to page and searches for the words required, then finally printing out the results we wanted.

Upon testing we came up with some results which are shown in the chapters below, and discussed in the conclusion.

# 4 Results with SQLite Search

In our code we printed the top 10 pages with the biggest frequencies for some queries. For the purpose of this report we will only show the results of the top 3 pages with the biggest frequencies.

## 4.1 predelovalne dejavnosti

Word: predelovalne dejavnosti
Results found in 11.967 ms.
Biggest frequencies: 1273, 75, 35
Some of the documents: evem.gov.si.371.html, evem.gov.si.377.html, evem.gov.si.452.html
Some of the snippets:

- PREDELOVALNE DEJAVNOSTI 10 Proizvodnja ..., dejavnosti /storitve in informacij ..., dejavnosti poslovni subjekti po ..., dejavnosti na kmetiji in ... (on site evem.gov.si.371.html)
- dejavnosti - brušenje in ..., dejavnosti - električna popravila ..., dejavnosti - gradnja temeljev ...(on site evem.gov.si.377.html )
- dejavnosti , drugje nerazvrščene ..., dejavnosti poslovni subjekti po ..., dejavnosti potrebuje tudi pravna ..., dejavnosti higienske nege in ... (on site evem.gov.si.452.html)

## 4.2 trgovina

Word: trgovina
Results found in 3.99 ms.
Biggest frequencies: 244, 86, 76
Documents: evem.gov.si.371.html, evem.gov.si.21.html, podatki.gov.si.340.html
Some of the snippets:

- trgovina na debelo s ..., trgovina na drobno s ..., trgovina na debelo ali ..., trgovina z njihovimi deli ... (on site evem.gov.si.371.html)
- Trgovina z zdravili na ..., Trgovina na debelo s ..., Trgovina na debelo z ..., Trgovina na drobno v ... (on site evem.gov.si.21.html)
- trgovina in storitve , ..., trgovina in storitve , ..., trgovina in proizvodnja , ..., trgovina d.o.o . ALMA ..., trgovina in posredovanje materialov ... (on site podatki.gov.si.340.html)

## 4.3 social services

Word: social services
Results found in 0.9980 ms.
Biggest frequencies: 4, 4, 1
Documents: e-uprava.gov.si.9.html, e-uprava.gov.si.45.html, podatki.gov.si.340.html
Some of the snippets:

- Social services , health ... Social services , health ... services , health , ... services , health , ... (on site e-uprava.gov.si.9.html)
- Social services , health ... Social services , health ... services , health , ... services , health , ... (on site e-uprava.gov.si.45.html)

- services ltd . TERME ... (on site podatki.gov.si.340.html)

### 4.4 trga nepremičnin

Word: trga nepremičnin
Results found in 2.964 ms.
Biggest frequencies: 48, 48, 45
Documents: e-prostor.gov.si.1.html, e-prostor.gov.si.51.html, e-prostor.gov.si.12.html
Some of the snippets:

- trga nepremičnin ( ETN ..., trga nepremičnin in poročanju ..., trga nepremičnin je zakonsko ..., trga nepremičnin je dostopna ... (on site e-prostor.gov.si.1.html)
- nepremičnin je javna zbirka ..., nepremičnin in indeksih vrednosti ..., nepremičnin geodetske uprave ( ... (on site e-prostor.gov.si.51.html)
- nepremičnin v naravi . ..., nepremičnin v naravi . ..., nepremičnin so prevzeti podatki ..., nepremičnin , ki niso ... (on site e-prostor.gov.si.12.html)

### 4.5 Slovenija

Word: Slovenija
Results found in 10.9 ms.
Biggest frequencies: 64, 15, 14
Documents: podatki.gov.si.340.html, podatki.gov.si.414.html, podatki.gov.si.424.html
Some of the snippets:

- SLOVENIJA , MINISTRSTVO ZA ..., SLOVENIJA UPRAVNA ENOTA AJDOVŠČINA ..., SLOVENIJA UPRAVNA ENOTA BREŽICE ..., SLOVENIJA UPRAVNA ENOTA CELJE ... (on site podatki.gov.si.340.html)
- SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA (on site podatki.gov.si.414.html)
- Slovenija , letno 55 ... Slovenija , letno 40 ... Slovenija , letno 32 ... (on site podatki.gov.si.424.html)

### 4.6 Sistem SPOT

Word: Sistem SPOT
Results found in 15.619 ms.
Biggest frequencies: 68, 36, 34
Documents: evem.gov.si.68.html, evem.gov.si.63.html, e-prostor.gov.si.18.html
Some of the snippets:

- SPOT , Slovenska poslovna ..., SPOT registracija se pridobi ..., SPOT registracija morajo izpolnjevati ..., (on site evem.gov.si.68.html)

- SPOT bo poslovnim subjektom ..., SPOT , Slovenska poslovna ..., SPOT predstavlja 12 regijskih ...(on site evem.gov.si.63.html)
- sistem trirazsežnih kartezičnih koordinat ..., sistem dvorazsežnih geodetskih/elipsoidnih koordinat ..., sistem normalnih ortometričnih višin ...(on site e-prostor.gov.si.18.html)

# 5 Results with basic search

Same as before, we printed the top 10 pages with the biggest frequencies for some queries. For the purpose of this report we will only show the results of the top 3 pages with the biggest frequencies.

## 5.1 predelovalne dejavnosti

Word: predelovalne dejavnosti
Results found in 68009.164 ms.
Biggest frequencies: 1291, 75, 40
Some of the documents: evem.gov.si.371.html, evem.gov.si.377.html, evem.gov.si.452.html
Some of the snippets:

- PREDELOVALNE DEJAVNOSTI 10 Proizvodnja ..., dejavnosti /storitve in informacij ..., dejavnosti poslovni subjekti po ..., dejavnosti na kmetiji in ... (on site evem.gov.si.371.html)
- dejavnosti - brušenje in ..., dejavnosti - električna popravila ..., dejavnosti - gradnja temeljev ...(on site evem.gov.si.377.html )
- dejavnosti , drugje nerazvrščene ..., dejavnosti poslovni subjekti po ..., dejavnosti potrebuje tudi pravna ..., dejavnosti higienske nege in ... (on site evem.gov.si.452.html)

## 5.2 trgovina

Word: trgovina
Results found in 66918.601 ms.
Biggest frequencies: 364, 96, 92
Documents: evem.gov.si.371.html, evem.gov.si.21.html, podatki.gov.si.340.html
Some of the snippets:

- trgovina na debelo s ..., trgovina na drobno s ..., trgovina na debelo ali ..., trgovina z njihovimi deli ... (on site evem.gov.si.371.html)
- Trgovina z zdravili na ..., Trgovina na debelo s ..., Trgovina na debelo z ..., Trgovina na drobno v ... (on site evem.gov.si.21.html)
- trgovina in storitve , ..., trgovina in storitve , ..., trgovina in proizvodnja , ..., trgovina d.o.o . ALMA ..., trgovina in posredovanje materialov ... (on site podatki.gov.si.340.html)

### 5.3 social services

Word: social services
Results found in 66998.427 ms.
Biggest frequencies: 5, 5, 1
Documents: e-uprava.gov.si.9.html, e-uprava.gov.si.45.html, podatki.gov.si.340.html
Some of the snippets:

- Social services , health ... Social services , health ... services , health , ... services , health , ... (on site e-uprava.gov.si.9.html)
- Social services , health ... Social services , health ... services , health , ... services , health , ... (on site e-uprava.gov.si.45.html)
- services ltd . TERME ... (on site podatki.gov.si.340.html)

### 5.4 trga nepremičnin

Word: trga nepremičnin
Results found in 68821.034 ms.
Biggest frequencies: 49, 45, 45
Documents: e-prostor.gov.si.1.html, e-prostor.gov.si.51.html, e-prostor.gov.si.12.html
Some of the snippets:

- trga nepremičnin ( ETN ..., trga nepremičnin in poročanju ..., trga nepremičnin je zakonsko ..., trga nepremičnin je dostopna ... (on site e-prostor.gov.si.1.html)
- nepremičnin je javna zbirka ..., nepremičnin in indeksih vrednosti ..., nepremičnin geodetske uprave ( ... (on site e-prostor.gov.si.51.html)
- nepremičnin v naravi . ..., nepremičnin so prevzeti podatki ..., nepremičnin , ki niso ... (on site e-prostor.gov.si.12.html)

### 5.5 Slovenija

Word: Slovenija
Results found in 72552.651 ms.
Biggest frequencies: 65, 15, 14
Documents: podatki.gov.si.340.html, podatki.gov.si.414.html, podatki.gov.si.424.html
Some of the snippets:

- SLOVENIJA , MINISTRSTVO ZA ..., SLOVENIJA UPRAVNA ENOTA AJDOVŠČINA ..., SLOVENIJA UPRAVNA ENOTA BREŽICE ..., SLOVENIJA UPRAVNA ENOTA CELJE ... (on site podatki.gov.si.340.html)
- SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA , MINISTRSTVO ZA ... SLOVENIJA (on site podatki.gov.si.414.html)
- Slovenija , letno 55 ... Slovenija , letno 40 ... Slovenija , letno 32 ... (on site podatki.gov.si.424.html)

### 5.6 Sistem SPOT

Word: Sistem SPOT
Results found in 66748.875 ms.
Biggest frequencies: 70, 39, 34
Documents: evem.gov.si.68.html, evem.gov.si.63.html, e-prostor.gov.si.18.html
Some of the snippets:

- SPOT , Slovenska poslovna ..., SPOT registracija se pridobi ..., SPOT registracija morajo izpolnjevati ..., (on site evem.gov.si.68.html)
- SPOT bo poslovnim subjektom ..., SPOT , Slovenska poslovna ..., SPOT predstavlja 12 regijskih ...(on site evem.gov.si.63.html)
- sistem trirazsežnih kartezičnih koordinat ..., sistem dvorazsežnih geodetskih/elipsoidnih koordinat ..., sistem normalnih ortometričnih višin ...(on site e-prostor.gov.si.18.html)

## 6 Conclusion

What we figured out from our results is that the inverted index method is much faster than the basic search. While the inverted index method requires milliseconds to find our results, the basic search takes much more time depending on the computer which runs the code.

One problem we have noticed with our code is that the size of the frequencies with the basic search and the SQLite search are different in some cases. For example the last word "Sistem SPOT" with the SQLite search has the three biggest frequencies as "68, 36, 34", while the basic search has its three frequencies as "70, 39, 34".