WEIR - Report on the second assignment

Matej Bevec
Nejc Hirci
Rok Nikolič

I. Introduction

In this assignment we explored different ways to index and retrieve data. We implemented a simple scrip for indexing and saving data to a SQLite database and a scripts for searching this index. We also made a script that uses simple word processing to just query our raw data.

II. Indexing

To build an index we used all the steps outlined in the assignment. We used beautiful soup for parsing the raw text and then the NLTK tokenizer to split up our found strings. We then compared our list of words with a list of stop words and removed the matches. Lastly we created a set out of this list to make it unique. For the words in this set we calculated the frequency of their appearance in our full word list as well as their indexes. We fed all this information into a predefined SQLite table structure to get our reverse index.

III. SQLite search

For the SQLite search we first process the query the same way we processed our document for building the index. For this processed query we make a query to the database and save the raw result. This is then processed to extract the relevant information of document name, frequency and indexes. We use some light regex work to extract and split out indexes and also extract the snippets of text around the indexes. This is all packaged up in a result along with the time it took to query this data. The information processing time not included.

IV. Basic search

The basic search preforms the same final processing to extract the relevant data but gets the raw data differently. It preforms simple operations to search our entire prepossessed raw data. It checks if a file contains all the words in our query. If it does we check the frequency buy summing all the time the word appears and we get the indexes buy enumerating over the raw data and checking for our query words. This is a much slower process.

V. Results

The database contained 49097 indexed words, the highest frequency words were: ”podatkov” with 10502 finds, ”slovenije” with 9856, ”republike” 8583, ”podatki” 5562, ”dejavnosti” 5560. And the top few documents: ”../data\evem.gov.si\evem.gov.si.371.html” with 102250 finds, ”../data\podatki.gov.si\podatki.gov.si.340.html” with 32138 and ”../data\e-prostor.gov.si\e-prostor.gov.si.166.html” with 11202. Bellow are the results for these queries: “predelovalne dejavnosti”, “trgovina”, “social services”, ”tuja”, ”sp”, ”vozniško dovoljenje”. The times for SQLite queries were a lot lower mostly because of the bottleneck of opening files in the basic search.

VI. Conclusion

In conclusion, the assignment demonstrated the effectiveness of SQLite search in achieving significantly faster query times compared to the basic search method. We observed a speed up of about 1500x to 3000x with the database for SQLite containing a total of 49097 indexed words. Overall, the results highlighted the efficiency and performance benefits of utilizing SQLite for data indexing and retrieval.

```
Results for query: "predelovalne dejavnosti"

    Results found in 34629.5ms.

    Frequencies Document                                    Snippet
    ----------- ------------------------------------------- ----------------------------------
    1288        ..\data\evem.gov.si\evem.gov.si.371.html iskanje ustrezne šifre dejavnosti /st
    19          ..\data\evem.gov.si\evem.gov.si.460.html drugje nerazvrščene predelovalne deja
    9           ..\data\evem.gov.si\evem.gov.si.15.html  pogojih za opravljanje dejavnosti in
```

Figure 1: Basic search 1.

```
Results for query: "predelovalne dejavnosti"

    Results found in 19.1ms.

    Frequencies Document                                      Snippet
    ----------- ----------------------------------------------- ----------------------------------
    1288        ../data\evem.gov.si\evem.gov.si.371.html          za infrastrukturo c predelovalne dejavno
    75          ../data\evem.gov.si\evem.gov.si.377.html          defektolog v zdravstveni dejavnosti deka
    40          ../data\podatki.gov.si\podatki.gov.si.340.html    - nosilec dopolnilne dejavnosti na kmeti
    39          ../data\evem.gov.si\evem.gov.si.452.html          druge storitvene dejavnosti , drugje...9
    31          ../data\evem.gov.si\evem.gov.si.653.html          dovoljenje za opravljanje dejavnosti spe
    29          ../data\evem.gov.si\evem.gov.si.72.html           od dohodka iz dejavnosti republika slove
    29          ../data\evem.gov.si\evem.gov.si.398.html          usmerjene na opravljanje dejavnosti ( np
    23          ../data\evem.gov.si\evem.gov.si.442.html          dejavnosti za nego...96.040 ) / dejavnos
    19          ../data\evem.gov.si\evem.gov.si.460.html          drugje nerazvrščene predelovalne dejavno
    18          ../data\evem.gov.si\evem.gov.si.28.html           za opravljanje gospodarske dejavnosti .
```

Figure 2: SQLite search 1.

```
Results for query: "trgovina"

    Results found in 34547.8ms.

    Frequencies Document                                    Snippet
    ----------- ------------------------------------------- ----------------------------------
    364         ..\data\evem.gov.si\evem.gov.si.371.html      gl . 46.110 trgovina na debelo...gl . 10.8
    94          ..\data\evem.gov.si\evem.gov.si.651.html      druga govedoreja druga trgovina na drobno.
    92          ..\data\evem.gov.si\evem.gov.si.21.html       moj e-vem evem›področja trgovina tu boste.
    82          ..\data\podatki.gov.si\podatki.gov.si.340.html a dent , trgovina in storitve.... adria in
    13          ..\data\evem.gov.si\evem.gov.si.623.html      trgovina na debelo...izdelki široke porabe
    12          ..\data\evem.gov.si\evem.gov.si.329.html      trgovina na debelo...in sanitarno opremo t
    12          ..\data\evem.gov.si\evem.gov.si.630.html      trgovina na drobno...predmeti za gospodinj
    10          ..\data\evem.gov.si\evem.gov.si.320.html      trgovina na debelo...napravami za ogrevanj
    10          ..\data\evem.gov.si\evem.gov.si.327.html      trgovina na debelo...napravami in opremo t
    10          ..\data\evem.gov.si\evem.gov.si.622.html      trgovina na debelo...električnimi gospodin
```

Figure 3: Basic search 2.

```
Results for query: "trgovina"

    Results found in 11.7ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ---------------------------------------
    364         ../data\evem.gov.si\evem.gov.si.371.html          gl . 46.110 trgovina na debelo...gl . 10.
    94          ../data\evem.gov.si\evem.gov.si.651.html          druga govedoreja druga trgovina na drobno
    92          ../data\evem.gov.si\evem.gov.si.21.html           moj e-vem evem›področja trgovina tu boste
    82          ../data\podatki.gov.si\podatki.gov.si.340.html a dent , trgovina in storitve.... adria i
    13          ../data\evem.gov.si\evem.gov.si.623.html          trgovina na debelo...izdelki široke porab
    12          ../data\evem.gov.si\evem.gov.si.329.html          trgovina na debelo...in sanitarno opremo
    12          ../data\evem.gov.si\evem.gov.si.630.html          trgovina na drobno...predmeti za gospodin
    10          ../data\evem.gov.si\evem.gov.si.320.html          trgovina na debelo...napravami za ogrevan
```

Figure 4: SQLite search 2.

```
Results for query: "social services"

    Results found in 34823.6ms.

    Frequencies Document                                        Snippet
    ----------- ------------------------------------------------ ----------------------------------------------------
    5           ..\data\e-uprava.gov.si\e-uprava.gov.si.45.html labour , retirement social services ,..., retirement soci
    5           ..\data\e-uprava.gov.si\e-uprava.gov.si.9.html  labour , retirement social services ,..., retirement soci
```

Figure 5: Basic search 3.

```
Results for query: "social services"

    Results found in 12.1ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ----------------------------------------------------
    5           ../data\e-uprava.gov.si\e-uprava.gov.si.9.html  labour , retirement social services ,...relationshi
    5           ../data\e-uprava.gov.si\e-uprava.gov.si.45.html labour , retirement social services ,...relationshi
    1           ../data\podatki.gov.si\podatki.gov.si.340.html  recreation and spa services ltd .
```

Figure 6: SQLite search 3.

```
Results for query: "tuja"

    Results found in 33845.2ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ----------------------------------------------------
    3           ..\data\evem.gov.si\evem.gov.si.371.html          tuji državljan oziroma tuja državljanka s...dom domača a
    2           ..\data\e-uprava.gov.si\e-uprava.gov.si.25.html kazniva dejanja obsodila tuja sodišča ,...vzgojne ukrepe
    2           ..\data\evem.gov.si\evem.gov.si.27.html           ustanovi domača ali tuja pravna oseba...tuje pravne osebe
    2           ..\data\evem.gov.si\evem.gov.si.398.html          ustanovi domača ali tuja fizična ali...register prebival
    1           ..\data\evem.gov.si\evem.gov.si.22.html           ustanovi domača ali tuja fizična ali
    1           ..\data\evem.gov.si\evem.gov.si.252.html          tuji državljan oziroma tuja državljanka s
```

Figure 7: Basic search 4.

```
Results for query: "tuja"

    Results found in 10.4ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ------------------------------------
    3           ../data\evem.gov.si\evem.gov.si.371.html          tuji državljan oziroma tuja državljanka s
    2           ../data\e-uprava.gov.si\e-uprava.gov.si.25.html kazniva dejanja obsodila tuja sodišča ,..
    2           ../data\evem.gov.si\evem.gov.si.27.html           ustanovi domača ali tuja pravna oseba...t
    2           ../data\evem.gov.si\evem.gov.si.398.html          ustanovi domača ali tuja fizična ali...re
    1           ../data\evem.gov.si\evem.gov.si.22.html           ustanovi domača ali tuja fizična ali
    1           ../data\evem.gov.si\evem.gov.si.252.html          tuji državljan oziroma tuja državljanka s
    1           ../data\evem.gov.si\evem.gov.si.652.html          ustanovi domača ali tuja fizična ali
```

Figure 8: SQLite search 4.

```
Results for query: "sp"

    Results found in 33809.6ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ------------------------------------------------
    18          ..\data\e-prostor.gov.si\e-prostor.gov.si.166.html 498334,99 116654,39 1183,92 sp . savinjska...514913,
    2           ..\data\evem.gov.si\evem.gov.si.36.html           lastnika objekta za sp izjava prokurista/zastopnika.
    2           ..\data\evem.gov.si\evem.gov.si.59.html           lastnika objekta za sp izjava prokurista/zastopnika.
    1           ..\data\e-uprava.gov.si\e-uprava.gov.si.46.html   shibbolejev indikator na sp prijave uporabnika
```

Figure 9: Basic search 5.

```
Results for query: "sp"

    Results found in 10.4ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ------------------------------------------------
    18          ../data\e-prostor.gov.si\e-prostor.gov.si.166.html 498334,99 116654,39 1183,92 sp . savinjska...514913
    2           ../data\evem.gov.si\evem.gov.si.36.html           lastnika objekta za sp izjava prokurista/zastopnika
    2           ../data\evem.gov.si\evem.gov.si.59.html           lastnika objekta za sp izjava prokurista/zastopnika
    1           ../data\e-uprava.gov.si\e-uprava.gov.si.46.html   shibbolejev indikator na sp prijave uporabnika
```

Figure 10: SQLite search 5.

```
Results for query: "vozniško dovoljenje"

    Results found in 32790.8ms.

    Frequencies Document                                          Snippet
    ----------- ------------------------------------------------- ------------------------------------------------------------
    7           ..\data\e-uprava.gov.si\e-uprava.gov.si.56.html če se ... vozniško dovoljenje z...se ... vozniško dovoljen
    5           ..\data\e-uprava.gov.si\e-uprava.gov.si.33.html potni list , vozniško dovoljenje )...list , vozniško dovol
    2           ..\data\e-uprava.gov.si\e-uprava.gov.si.44.html potni list , vozniško dovoljenje )...list , vozniško dovol
    2           ..\data\evem.gov.si\evem.gov.si.361.html        potni list ali vozniško dovoljenje )...list ali vozniško d
    2           ..\data\evem.gov.si\evem.gov.si.647.html        potni list ali vozniško dovoljenje )...list ali vozniško d
```

Figure 11: Basic search 6.

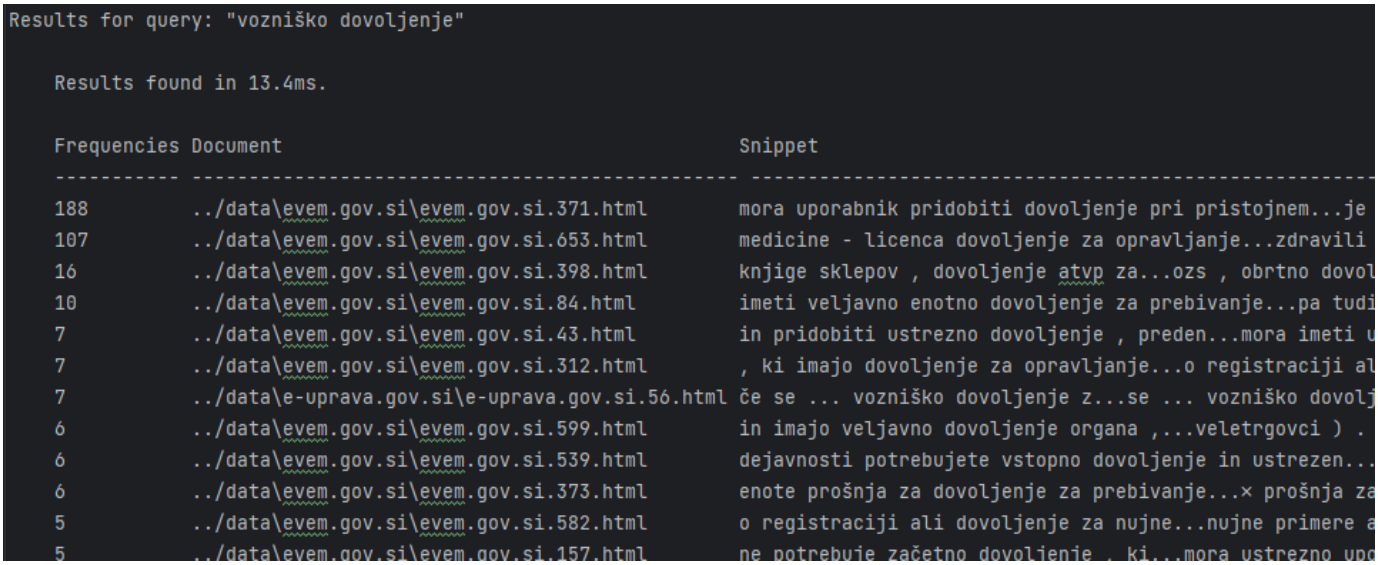

Figure 12: SQLite search 6.