Web Information Extraction and Retrieval
Programming Assignment 3

Jure Bevc, Luka Tavčer, Grega Dvoršak
University of Ljubljana, Faculty for Computer and Information Science
bevc.jure@gmail.com, lt7339@student.uni-lj.si, gd4667@student.uni-lj.si

May 22, 2020

1 Introduction

This report summarizes our work on the third assignment for Web Information Extraction and Retrieval. As per the assignment's instructions, we extracted textual information from 1416 crawled web pages, performed preprocessing, stored data into index and used the index for querying. Our implementation consists of two parts: data processing with indexing and data retrieval. Data retrieval was done in two ways, with and without usage of inverted index.

2 Data Processing with Indexing

The data first needs to be retrieved from web pages. As the pages were provided in a compressed file as part of the instructions, we can set the HTML files to be read from a directory. The pages are grouped into four domains: ”e-prostor.gov.si”, ”e-uprava.gov.si”, ”evem.gov.si” and ”podatki.gov.si”. We read the textual content from the files using the BeutifulSoup [1] library. To speed up the processing, we also remove scripts from the HTML files using the same library. The extracted text is then tokenized using the Natural Language Toolkit (NLTK) [2], is converted to lowercase, and stopword removal is performed. An index word is created for each non-stopword and is inserted into the database. Indexes are then obtained as appearances of the index word in the original documents and their frequencies are calculated. If a posting for the token does not exist yet, a new one is created. Duplicates of indexes are handled by combining a set of the existing indexes with the new ones, so only unique indexes remain. Finally, we check if there were any new indexes added and if so, the database is updated accordingly.

The database contains 400849 postings and 49175 words. The most frequent tokens are punctuation symbols and numbers. The performance could be improved if these tokens were filtered out. The most frequent word entries are: ”vod”, ”go”, ”st”, ”proizvodnja”, ”javno”, ”vnos” and ”javnost”.

3 Data Retrieval with Inverted Index

To retrieve data from the given web pages, we first tokenize, remove stopwords from, and set the query to lowercase so that it has the same format as the text we preprocessed in the previous section. Then we execute the query by getting all postings of the tokens in the database. We can then retrieve textual data using these postings in a similar way as described in the previous section, only using the document names form the postings. We obtain data for each query from the database and print it as shown in the Results section

4 Data Retrieval Without Inverted Index

As is done when using inverted index, the first step when retrieving data without inverted index is to tokenize, remove stopwords and set the query to lowercase. This time, the postings need to be generated without the use of previously stored data in the database. Text from the documents is obtained in a similar way as it was in the preprocessing step. The positions, indexes and frequencies of the tokens are obtained by using regular expression matching operations. The postings are then created from these values. Finally, we use the data from the created postings to obtain results similarly as was done with the inverted index. The results are shown in the Results section.

5 Results

We ran and compared both types of searches for the queries: ”predelovalne dejavnosti”, ”trgovina”, ”social services”, ”Sistem SPOT”, ”eVem” and ”podnebnih sprememb”. The results are shown below. Note, that the snippets are shortened from their original size in order to fit to paper.

```
With inverted index:
Results for a query: predelovalne dejavnosti
Results found in 4ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
1570           evem.gov.si.371.html      anje ustrezne ifre dejavnosti /storitve...ojih za opravljanje dejavnosti. V iskal...
78             evem.gov.si.377.html      tolog v zdravstveni dejavnosti Dekan oz...tetik v zdravstveni dejavnosti Dimnikar...
47             evem.gov.si.452.html      oj e-VEMeVEMDejavnostiDruge sto...tiDruge storitvene dejavnosti, drugje n...
40             podatki.gov.si.340.html    NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJ...  CENTER INTERESNIH DEJAVNOSTI PTUJ    ...
31             evem.gov.si.398.html      jene na opravljanje dejavnosti (npr.: pr... namene opravljanja dejavnosti ipd.V ...

Without inverted index:
Results for a query: predelovalne dejavnosti
Results found in 38185ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
1570           evem.gov.si.371.html      anje ustrezne ifre dejavnosti /storitve...ojih za opravljanje dejavnosti. V iskal...
78             evem.gov.si.377.html      tolog v zdravstveni dejavnosti Dekan oz...tetik v zdravstveni dejavnosti Dimnikar...
47             evem.gov.si.452.html      oj e-VEMeVEMDejavnostiDruge sto...tiDruge storitvene dejavnosti, drugje n...
40             podatki.gov.si.340.html    NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJ...  CENTER INTERESNIH DEJAVNOSTI PTUJ    ...
31             evem.gov.si.398.html      jene na opravljanje dejavnosti (npr.: pr... namene opravljanja dejavnosti ipd.V ...

With inverted index:
Results for a query: trgovina
Results found in 0ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
368            evem.gov.si.371.html      nizacij, gl. 46.110trgovina na debelo s... in juh, gl. 10.890trgovina na debelo z...
96             evem.gov.si.651.html      a govedoreja Druga trgovina na drobno v... prodajalnah Druga trgovina na drobno v...
92             evem.gov.si.21.html       eVEMPodrojaTrgovinaTu boste n...m dejavnostiDruga trgovina na drobno v...
82             podatki.gov.si.340.html              A DENT, trgovina in storitve...  ADRIA INVESTICIJE trgovina, posrednit...
13             evem.gov.si.623.html      eVEMDejavnostiTrgovina na debelo z...iroke porabeTrgovina na debelo z...

Without inverted index:
Results for a query: trgovina
Results found in 38130ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
368            evem.gov.si.371.html      nizacij, gl. 46.110trgovina na debelo s...arske       raziskave, gl. 71.121SURS... in juh, gl. 10.890trgovina na debelo z...
96             evem.gov.si.651.html      a govedoreja Druga trgovina na drobno v... prodajalnah Druga trgovina na drobno v...
92             evem.gov.si.21.html       eVEMPodrojaTrgovinaTu boste n...m dejavnostiDruga trgovina na drobno v...
82             podatki.gov.si.340.html              A DENT, trgovina in storitve...  ADRIA INVESTICIJE trgovina, posrednit...
13             evem.gov.si.623.html      ijavaDomovRazmiljamZaenjam...eVEMDejavnostiTrgovina na debelo z...iroke porabeTrgovina na debelo z...

With inverted index:
Results for a query: social services
Results found in 0ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
5              e-uprava.gov.si.45.html   abour, retirementSocial services, hea...ationship etc.? Social services, hea...
5              e-uprava.gov.si.9.html    abour, retirementSocial services, hea...ationship etc.? Social services, hea...
1              evem.gov.si.661.html      Records and Related Services (AJPES) and...
1              podatki.gov.si.340.html    recreation and spa services ltd.      ...

Without inverted index:
Results for a query: social services
Results found in 39291ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
118            evem.gov.si.371.html      ajejo kredite le za socialne in izobrae...adov, razen obvezne socialne varnosti 65...
95             podatki.gov.si.340.html   ura in port                            Sociala in zaposlova...ura in port...
58             evem.gov.si.29.html       vne oblike podjetijSocialno podjetje (S...o podjetje (So.p.)Socialno podjetje (S...
50             evem.gov.si.32.html        15MarPrispevki za socialno varnost za ...ovoljno vkljuene v socialno zavarovanje...
29             podatki.gov.si.414.html   ura in port                            Sociala in zaposlova...ura in port...
```

```
With inverted index:
Results for a query: Sistem SPOT
Results found in 11ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
134            evem.gov.si.371.html      epublika Slovenija SPOT, Slovenska posl...revanje namakalnih sistemov, naprav za ...
81             evem.gov.si.49.html       epublika Slovenija SPOT, Slovenska posl... postopkovnega dela sistema e-VEMA. UV...
74             evem.gov.si.68.html       epublika Slovenija SPOT, Slovenska posl...itev poloaja toke SPOT registracija...
72             e-prostor.gov.si.150.html skega koordinatnega sistema v Sloveniji...storski koordinatni sistem/Projekti dr...
52             e-prostor.gov.si.57.html  storski koordinatni sistemSplona vpraa...orem se prijaviti v sistem. Uporabniko ...

Without inverted index:
Results for a query: Sistem SPOT
Results found in 38296ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
135            evem.gov.si.371.html      epublika Slovenija SPOT, Slovenska posl...revanje namakalnih sistemov, naprav za ...
81             evem.gov.si.49.html       epublika Slovenija SPOT, Slovenska posl...ti, ki bi lahko uporabniku nastale zarad...
74             e-prostor.gov.si.150.html skega koordinatnega sistema v Sloveniji...storski koordinatni sistem/Projekti dr...
74             evem.gov.si.68.html       epublika Slovenija SPOT, Slovenska posl...itev poloaja toke SPOT registracija...
52             e-prostor.gov.si.57.html  storski koordinatni sistemSplona vpraa...orem se prijaviti v sistem. Uporabniko ...

With inverted index:
Results for a query: eVem
Results found in 0ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
5              evem.gov.si.398.html      ramMoj e-VEMeVEMPomo in podpor...slano preko portala eVEM oziroma preko e...
3              evem.gov.si.84.html       ramMoj e-VEMeVEMVodenje podjetj...preko portala SPOT (eVEM). Naseznamu e-...
1              evem.gov.si.36.html       ramMoj e-VEMeVEMTiskan...
1              evem.gov.si.362.html      ramMoj e-VEMeVEMNotarjiUra...
1              evem.gov.si.371.html      ramMoj e-VEMeVEMSKD Seznam...

Without inverted index:
Results for a query: eVem
Results found in 38686ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
6              evem.gov.si.390.html      etniki skladPrijava    ...r.           evemPrijava      ...nemogoiti pikotek evemPrijava...
5              evem.gov.si.398.html      ramMoj e-VEMeVEMPomo in podpor...ti identifikacijsko tevilko za DDV (naj...
5              evem.gov.si.375.html      ortal za poslovne subjekte in samostojne...ramMoj e-VEMeVEMPomo in podpor...
4              evem.gov.si.33.html       ramMoj e-VEMeVEMVmesnik za refu... elektronski naslov evem.mjuping@govpong...
3              evem.gov.si.13.html       ramMoj e-VEMeVEMKadrovski vmesn... elektronski naslov evem.mjuping@govpong...

With inverted index:
Results for a query: podnebnih sprememb
Results found in 0ms
Frequencies    Document                  Snippet
----------     ------------------------- -----------------
38             evem.gov.si.398.html      poslovanje. Zakon o spremembah in dopoln...sivloi ponovitev, spremembo ali razvel...
19             e-prostor.gov.si.54.html  trukture. Elaborati sprememb z vsemi pri...D48/GK). Elaborate sprememb s koordinat...
19             evem.gov.si.393.html      ke podjetijPrijava sprememb podatkov v ... SlovenijePrijava sprememb podatkov v ...
9              evem.gov.si.37.html       k e-Vem:prijavite spremembo podatkov v...lo Odvzem ali sprememba pooblastil...
9              evem.gov.si.373.html      avarovanje, prijava sprememb podatkov o ...     Prijava sprememb firme, posl...

Without inverted index:
Results for a query: podnebnih sprememb
Results found in 38382ms
Frequencies    Document                  Snippet
----------     -----------------------   -----------------
42             evem.gov.si.398.html      vezanec v obdobju zadnjih 12 mesecev ne ...poslovanje. Zakon o spremembah in dopoln...
23             e-prostor.gov.si.54.html  ega katastra je prikaz zasedenosti prost...trukture. Elaborati sprememb z vsemi pri...
22             evem.gov.si.393.html      ke podjetijPrijava sprememb podatkov v ... SlovenijePrijava sprememb podatkov v ...
12             evem.gov.si.373.html      anja s strani zakonitega zastopnika mon...avarovanje, prijava sprememb podatkov o ...
12             evem.gov.si.86.html       postopek zaposlitveSprememba podatkov v...alnih zavarovanjihSprememba podatkov v...
```

From the results we can see, that the most important aspect of working with inverted indexes is speed. The indexed queries take only a few milliseconds to return results while the basic search might take a few seconds. Other than the speed, some differences seem to appear in the results as the basic search might find more occurrences of a token. The difference however is not significant enough to choose the basic search as the better approach when considering the speed-accuracy trade off. Some queries, like "predelovalne dejavnost" and "trgovina" do return the same results in both cases.

References

[1] The beautiful soup library. https://www.crummy.com/software/BeautifulSoup. Accessed: 2020-05-21.

[2] The natural language toolkit. https://www.nltk.org. Accessed: 2020-05-01.