Web Information Extraction and Retrieval Programming Assignment 3

Rok Cej, Metodija Bucevski

1. Data processing and indexing

Implemented in `run-index.py`.

1.1. HTML parsing

The library BeautifulSoup is used for parsing HTML. Once the HTML tree is built, we remove the nodes labelled `"script", "style", "head", "meta", "[document]"` to eliminate content that does not contain visible text. Afterwards, the function `BeautifulSoup.get_text()` is called to extract text content from the modified HTML tree.

1.2. Text preprocessing

After text is extracted from a page, we process it in multiple steps. In the tokenization step, the library nltk is used. The steps are as follows:

1. Split text by whitespace
   - This is to enforce consistency, as the tokenizer can behave differently when encountering spaces and newline characters
   - For example, `"a' b"` → `["a", "'", "b"]`, but `"a'\nb"` → `["a'", "b"]`
2. Tokenize words using `nltk.tokenize.word_tokenize()`
3. Convert tokens to lower case
4. Remove stopwords and punctuation marks
   - Stopwords are defined in `stopwords.STOPWORDS_SLOVENE` (the provided stopwords were used with some slight modifications)
   - Punctuation marks are defined in `preprocessing.PUNCTUATION_MARKS`
5. Remove dot sequences (`".."`, `"..."`, `"...."`, etc.)
   - Tokenizer splits groups of symbols (`"!!!"` → `["!", "!", "!"]`), however, dots always stay grouped (`"..."` → `["..."]`), so they need to be handled separately
6. Remove leading apostrophe (`"'word"` → `"word"`)
   - Tokenizer separates trailing apostrophes, but not leading apostrophes, which can cause issues with quoted words (`"'quote'"` → `["'quote", "'"]`)
7. Remove trailing dot (`"word."` → `"word"`)
   - Tokenizer can sometimes leave a dot at the end of a word (`"gov.si.,"` → `"gov.si."`), so it needs to be manually removed.
8. Map tokens to their normalized version
   - This allows one word to match multiple words with the same meaning
   - The following word mappings are defined in `preprocessing.WORD_MAP`:
     - `["eur", "euro", "eurov", "evro", "evrov", "€"]` → `"eur"`

1.3. Indexing

After processing the text, words are inserted into the SQLite database along with their positions in the original text.

9. Data retrieval with inverted index (SQLite search)

Implemented in `run-sqlite-search.py`.

1. First, the same text preprocessing function from earlier is applied on the search query.
2. Next, we use the SQLite index to find which documents contain matches and where they are located within the document. Those documents are then parsed and snippets that containt a match are extracted. Each snippet includes 3 neighboring words on each side of the match. Overlapping snippets are merged and the rest are joined by putting "..." before and after each snippet. If a snippet includes the beggining or the end of the document, the "..." is omitted.
3. Finally, search results are printed to the standard output. These include the number of matches (frequency) within each document, the name of the document, and snippets containing the matches. By default, only top 5 results are displayed, with a maximum of 4 snippets per result (both parameters are configurable).

10. Data retrieval without inverted index (basic search)

Implemented in `run-basic-search.py`.

1. Similar to indexed search, the first step is to preprocess the search query.
2. Next, every single page is parsed and searched for exact matches. Snippets are extracted in the same way as previously described. In order to speed up this process, multiprocessing is used. The list of input pages is evenly divided among each process. Results of each process are then merged into a single list, which is then sorted by descending frequeny.
3. Search results are displayed just like in indexed search.

For comparison, a sequential (single-process) version of basic search is implemented in `run-basic-search-sequential.py`.

As a side note, multithreading in Python is executed in a sequential manner (although thread execution is interleaved). This makes multithreading appropriate for situations where threads spend a lot of time waiting. However, for heavy computation, multithreading is equally, if not less, efficient than single-threaded execution. Therefore, multiprocessing was used for this task instead, as processes in Python are actually executed in parallel.

11. Database results

- Number of indexed words: **47062**
- Words with the highest frequencies:

Word	Frequency sum
podatkov	11048
slovenije	9928
republike	8572

- Documents with highest frequencies:

Document	Frequency sum
evem.gov.si/evem.gov.si.371.html	82693
podatki.gov.si/podatki.gov.si.340.html	27423
e-prostor.gov.si/e-prostor.gov.si.166.html	11163

- Words that appear in the highest number of documents:

Word	Document count
uporabe	1399
pogoji	1398
domov	1384

12. Query results

Each query was executed using SQLite search (invered index) and basic search (no inderted index). Basic search was tested both on the parallel version (8 processes) and the sequential version. The following table compares search times for each query.

Query	SQLite search time	Basic search time (parallel)	Basic search time (sequential)
“predelovalne dejavnosti”	5.1s	46.8s	136.2s
“trgovina”	4.4s	44.8s	128.9s
“social services”	1.2s	43.0s	125.8s
“eur”	0.2s	44.4s	134.3s
“osnovna srednja šola univerza”	4.3s	38.9s	124.3s
“podjetje d.o.o. s.p.”	1.4s	55.0s	146.2s

On average, SQLite search was **16.2** times faster than the parallel basic search, and **47.3** times faster than the sequential basic search.

12.1. Query “predelovalne dejavnosti”

This query had a lot of inaccurate matches due to the frequency of the word “dejavnosti” being much higher than that of “predelovalne”. Consequently, no exact matches of “predelovalne dejavnosti” are shown in the shortened snippets.

```
Results for query: "predelovalne dejavnosti"

  Frequency  Document                                  Snippet
  ---------  ----------------------------------------  ----------------------------------------
  1291       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 dejavnosti A KMETIJSTVO IN ...
  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 v zdravstveni dejavnosti Laboratorijski tehnik
Ladijski ...
  40         podatki.gov.si/podatki.gov.si.340.html    ... - 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 ...
  38         evem.gov.si/evem.gov.si.452.html          ... e-VEM eVEM › Dejavnosti › Druge
storitvene dejavnosti, drugje nerazvrščene (96.090) Druge storitvene dejavnosti, drugje nerazvrščene
(96.090) ... SKD šifra zajema dejavnosti in storitve, za ...
  31         evem.gov.si/evem.gov.si.653.html          ... Dovoljenje za opravljanje dejavnosti
specializirane prodajalne z ... radijske ali televizijske dejavnosti Dovoljenje za izvajanje sevalne
dejavnosti Dovoljenje za izvajanje sevalne dejavnosti Dovoljenje za izvajanje ...
```

12.2. Query “trgovina”

This query found some advertisements for shops (“ADRIA INVESTICIJE”, “ALBA”) along with trade-related government resources (“trgovina na debelo”, “trgovina na drobno”).

```
Results for query: "trgovina"

  Frequency  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 ...
  96         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 ...
  92         evem.gov.si/evem.gov.si.21.html           ... 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 ...
  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. ...
  13         evem.gov.si/evem.gov.si.623.html          ... › Dejavnosti › 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 ...
```

12.3. Query “social services”

This query successfuly found resources for social services, with some additional matches for other types of services, including spa (“TERME MARIBOR”).

```
Results for query: "social services"

  Frequency  Document                                  Snippet
  ---------  ----------------------------------------  ----------------------------------------
  5          e-uprava.gov.si/e-uprava.gov.si.45.html   ... culture Labour, retirement Social
services, health, death Taxes ... employment relationship etc.? Social services, health, death How ...
  5          e-uprava.gov.si/e-uprava.gov.si.9.html    ... culture Labour, retirement Social
services, health, death Taxes ... employment relationship etc.? Social services, health, death How ...
  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, ...
```

12.4. Query “eur”

This query demonstrates the benefits of currency normalization. Only by searching for “eur”, we get various matches including “EUR”, “evrov”, and “€”.

```
Results for query: "eur"

  Frequency  Document                                  Snippet
  ---------  ----------------------------------------  ----------------------------------------
  43         evem.gov.si/evem.gov.si.398.html          ... mesecev presegla 50.000 EUR. Kaj se
zgodi, ... pa presežemo 50.000 EUR obdavčljivega prometa? Identifikacijsko ... presegel znesek 50.000
EUR obdavčljivega prometa. Zahtevek ... obdavčljivega prometa 50.000 EUR, mora že od ...
  10         evem.gov.si/evem.gov.si.72.html           ... akontacije presega 400 evrov), ali v
trimesečnih ... ne presega 400 evrov). Obroki predhodne akontacije ... več kot 40.000 evrov, ali
80.000 evrov, če je bila ...
  10         evem.gov.si/evem.gov.si.77.html           ... dohodnine: Prihodki 232,00 € Odhodki -
84,00 € Višina splošne olajšave - 181,00 € Osnova 497,00 € Ugotavljanje davčne osnove ...
  8          podatki.gov.si/podatki.gov.si.456.html    ... iz tujine (1000 EUR) po sektorju
izvedbe ... za RRD (1000 EUR) po sektorju izvedbe ... za RRD (1000 EUR) po sektorju izvedbe ... za RRD
(1000 EUR) po sektorju izvedbe ...
  7          evem.gov.si/evem.gov.si.78.html           ... v višini 9,06 EUR in sicer: 7,05 EUR za
pokojninsko in ... dobe) in 2,01 EUR za zdravstveno zavarovanje ... višji od 5.948,64 EUR, pri čemer
v ...
```

12.5. Query “osnovna srednja šola univerza”

This query looks for mentions of educational institutions. We can see it was mostly successful by finding phrases such as “Walfdorska osnovna šola”, “Srednja šola Kdaj lahko opravljam”, and “Univerza v Ljubljani”.

```
Results for query: "osnovna srednja šola univerza"

  Frequency  Document                                  Snippet
  ---------  ----------------------------------------  ----------------------------------------
  1140       podatki.gov.si/podatki.gov.si.340.html    ... mediji d.o.o. 2. OSNOVNA ŠOLA SLOVENSKA
BISTRICA 2TDK, ... ANDRAGOŠKI ZAVOD LJUDSKA UNIVERZA VELENJE ANDRAGOŠKI ZAVOD MARIBOR-LJUDSKA UNIVERZA
ANDREJC nizke gradnje, ...
  18         evem.gov.si/evem.gov.si.371.html          ... tkanin, katerih bistvena osnovna
sestavina je guma, ... tkanine, katere bistvena osnovna sestavina je guma, ... tkanin, katere bistvena
osnovna sestavina je plastika, ... načelih (npr. Waldorfska osnovna šola) · osnovnošolsko ...
  8          e-prostor.gov.si/e-prostor.gov.si.150.html  ... za uporabo programa, Univerza v
Ljubljani, Fakulteta ... uporabo spletne aplikacije, Univerza v Ljubljani, Fakulteta ... . Tehnično
poročilo, Univerza v Ljubljani, Fakulteta ... . Tehnično poročilo, Univerza v Ljubljani, Fakulteta ...
  5          e-uprava.gov.si/e-uprava.gov.si.16.html   ... višjo šolo ... Osnovna šola Kdaj in
kako ... osnovno šolo ... Srednja šola Kdaj lahko opravljam ...
  4          e-prostor.gov.si/e-prostor.gov.si.46.html  ... . Diplomska naloga. Univerza v Ljubljani,
Fakulteta ... . Diplomska naloga. Univerza v Ljubljani, Fakulteta ... . Diplomska naloga. Univerza v
Ljubljani, Fakulteta ... . Diplomska naloga. Univerza v Ljubljani, Fakulteta ...
```

12.6. Query “podjetje d.o.o. s.p.”

This query finds mentions of companies and businesses on provided pages. Some of the matches include dentists (“zobozdravstvena dejavnost d.o.o.”) and resources on how to start your own business (“Postopek ustanovitve s.p.”).

```
Results for query: "podjetje d.o.o. s.p."

  Frequency  Document                                  Snippet
  ---------  ----------------------------------------  ----------------------------------------
  1386       podatki.gov.si/podatki.gov.si.340.html    ... ZOBJE zobozdravstvena dejavnost d.o.o.
2leva dentalna medicina in mediji d.o.o. 2. OSNOVNA ŠOLA ... za razvoj projekta, d.o.o. 32BISEROV,
d.o.o., Zobna ambulanta Alenka ...
  51         evem.gov.si/evem.gov.si.23.html           ... › Želim ustanoviti podjetje › Poslovne
oblike ... › Postopek ustanovitve s.p. Postopek ustanovitve s.p. Registracija samostojnega podjetnika
(s.p.) Začni Status samostojnega ...
  30         evem.gov.si/evem.gov.si.22.html           ... › Želim ustanoviti podjetje › Poslovne
oblike ... kje lahko ustanovite podjetje? Ustanovitev samostojnega podjetnika ... notarju, ko
ustanavljate podjetje? Kdo ne more ... člena). Samostojni podjetnik (s.p.) Samostojni podjetnik je ...
  30         evem.gov.si/evem.gov.si.35.html           ... povezanih izdelkov za s.p. Začni
Pridobitev dvojnika ... z BiH za s.p. Začni Pridobitev dovoljenja ... Registracija enostavne eno-
osebne d.o.o. Začni Pridobitev obrtnega ... davčnih podatkov za d.o.o. Začni Predložitev zahtevka ...
  30         evem.gov.si/evem.gov.si.6.html            ... povezanih izdelkov za s.p. Začni
Pridobitev dvojnika ... z BiH za s.p. Začni Pridobitev dovoljenja ... Registracija enostavne eno-
osebne d.o.o. Začni Pridobitev obrtnega ... davčnih podatkov za d.o.o. Začni Predložitev zahtevka ...
```