Univerza *v Ljubljani*
Fakulteta *za računalništvo*
in informatiko

Web information extraction and retrieval

Project 3
Inverted index

Author:
Erik Kristian Janežič

Ljubljana, 22.5.2020

Contents

1 Introduction

In this project we implemented an inverted index and query system against it for a set of 1416 locally stored websites. Goal of the project was also to analyze performance difference between querying an inverted index and searching the documents incrementally. For this reason we also implemented a simple incremental query search algorithm.

2 Implementation details

2.1 Building the index

Below are described steps of the index building algorithm:

- Get path to html document
- Use `beautifulsoup4` to parse the html document
- Remove the html elements with the following tags

  `['style', 'script', '[document]', 'head', 'title']` from the html tree. Children of these elements are removed as well. Elements with these tags dont contain text that we are interested in querying against.
- Extract only human readable textual content from remaining html elements
- Tokenize the text with `word_tokenize` tokenizer from `nltk.tokenize` package
- Enumerate the tokens before removing any of undesired tokens. This way we will later be able to extract correct snipets of surrounding text
- Clean out undesired tokens: non alphabetic tokens and stop words . Convert remaining tokens to lowercase
- Use remaining tokens and its indices to build a pandas dataframe of words for this document
- Aggregate the dataframe on words and use appropriate aggregation functions to calculate word frequencies and generate string of indices they appear at

- This dataframe already has unique words in column so it can be appended to Sqlite database using the `pandas` interface
- Execute above procedure for all html documents of interest
- To create dictionary of words we just execute the query

  `SELECT DISTINCT(word) FROM Posting` and create new table from it
- inverted index generation is finished

2.2 Querying

Below are described steps of querying algorithm for querying the inverted index:

- Preprocess query string in the same way as we processed the html in inverted index creation: tokenize, exclude non alphabetic tokens, exclude stop words, convert to lowercase
- Dynamically create sql query of the following form:

```
SELECT sum(frequency) as frequency,documentName,
group_concat(indexes) as indexes FROM Posting
WHERE word IN (?,?,?,...)
GROUP BY documentName
```

  where `?,?,?,...` represents variable number of "?" based on number of query arguments we receive
- Read sql results and extract indices of query words in the document
- Open found documents and tkoenize them (without removing any tokens) and use the indices obtained from previous step to generate short snipets around each query word
- Apply desired formatting to generate human readable query results

For basic incremental querying the algorithm is basically the same as in inverted index generation. The only difference is that we also exclude the tokens that are not in our query from the token list, and that we output the frequencies and snipets on the fly while we traverse individual documents.
Speed comparison of both qerry approaches when querying for the word "šola":

- Basic: 53.15174198150635 sec
- Inverted index:

  Found documents 0.0015423297882080078 sec

  Found snipets: 5.439236402511597 sec

With sniped generation included we see that inverted index querying is about 10 times faster. If we included snipets in the database while generating the index the speed up would be much greater (about 5000 faster for this example).

3 Database description

- Number of words found: 31952
- Some documents with highest word frequencies:

	word	document name	frequency
	proizvodnja	../site_data/evem.gov.si/evem.gov.si.371.html	2266
	gl	../site_data/evem.gov.si/evem.gov.si.371.html	1668
	spada	../site_data/evem.gov.si/evem.gov.si.371.html	1338
	dejavnosti	../site_data/evem.gov.si/evem.gov.si.371.html	1287
	xsd	../site_data/e-prostor.gov.si/e-prostor.gov.si.147.html	927
	skupnost	../site_data/podatki.gov.si/podatki.gov.si.340.html	809
	krajevna	../site_data/podatki.gov.si/podatki.gov.si.340.html	754
	ministrstvo	../site_data/evem.gov.si/evem.gov.si.371.html	589
	šola	../site_data/podatki.gov.si/podatki.gov.si.340.html	582
	dejavnost	../site_data/evem.gov.si/evem.gov.si.371.html	545

- Some words with highest frequencies in all documents:

	word	frequency
	podatkov	11048
	slovenije	9928
	republike	8572
	dejavnosti	5564
	podatki	4940
	portalu	4701
	navigation	4474
	krepko	4277
	navadno	4242
	pogoji	4180

- Words that occured in most documents:

	word	in num documents
	uporabe	1399
	pogoji	1398
	domov	1384
	portalu	1367
	prostor	1349
	sistem	1294
	pomoč	1269
	davki	1257
	zdravje	1257
	državni	1238

4 Example query results

Below are first few results of example queries with links to the full results on the Github.Here we present only first two snipets, and smaller set of found files. For each example we also

show how fast and how many relevant documents we found (e.g. `Found 754 documents in 0.0043`). At the end of each example we also report how long did it take to find the snipets around queried words (e.g. `Snippets found in: 23.596418142318726s`). Sniept generation takes more time because in our implementation we to reopen the document and tokenize it again to get original sequence of tokens in the document.

- predelovalne dejavnosti

```
python run_sqlite_search.py "predelovalne dejavnosti"
Querring for:['predelovalne', 'dejavnosti']
Found 754 documents in 0.0043942928314208984s
Frequency      Document name
---------------------------------------------------------------
1291           evem.gov.si/evem.gov.si.371.html
75             evem.gov.si/evem.gov.si.377.html
40             podatki.gov.si/podatki.gov.si.340.html
38             evem.gov.si/evem.gov.si.452.html
31             evem.gov.si/evem.gov.si.653.html
30             evem.gov.si/evem.gov.si.398.html
28             evem.gov.si/evem.gov.si.72.html
23             evem.gov.si/evem.gov.si.442.html
18             evem.gov.si/evem.gov.si.28.html

    Snipet 0
---------------------------------------------------------------
    ...ustrezne šifre dejavnosti /storitve in...
    ...v zdravstveni dejavnosti Dekan oziroma...
    ...NOSILEC DOPOLNILNE DEJAVNOSTI NA KMETIJI...
    ...eVEM › Dejavnosti › Druge...
    ...za opravljanje dejavnosti specializirane prodajalne...
    ...na opravljanje dejavnosti ( npr...
    ...dohodka iz dejavnosti Davek od...
    ...eVEM › Dejavnosti › Dejavnosti...
    ...opravljanje gospodarske dejavnosti . Lastnosti...


 Snipet 1
---------------------------------------------------------------
 ...za opravljanje dejavnosti . V...
 ...v zdravstveni dejavnosti Dimnikar Diplomirana...
 ...CENTER INTERESNIH DEJAVNOSTI PTUJ CENTER...
 ...Druge storitvene dejavnosti , drugje...
 ...ali televizijske dejavnosti Dovoljenje za...
 ...namene opravljanja dejavnosti ipd ....
 ...dohodka iz dejavnosti Ko začnete...
 ...Dejavnosti › Dejavnosti za nego...
 ...za posamezne dejavnosti ali posamezne...

 Snippets found in: 23.596418142318726s
```

- trgovina

```
python run_sqlite_search.py "trgovina"
Querring for:['trgovina']
Found 127 documents in 0.00194549560546875s
Frequency      Document name
-----------------------------------------------------------
364            evem.gov.si/evem.gov.si.371.html
96             evem.gov.si/evem.gov.si.651.html
92             evem.gov.si/evem.gov.si.21.html
82             podatki.gov.si/podatki.gov.si.340.html
13             evem.gov.si/evem.gov.si.623.html
12             evem.gov.si/evem.gov.si.630.html
12             evem.gov.si/evem.gov.si.329.html
10             evem.gov.si/evem.gov.si.622.html
10             evem.gov.si/evem.gov.si.320.html
       Snipet 0
-----------------------------------------------------------
       .... 46.110 trgovina na debelo...
       ...govedoreja Druga trgovina na drobno...
       ...› Področja Trgovina Tu boste...
       ...DENT , trgovina in storitve...
       ...Dejavnosti › Trgovina na debelo...
       ...Dejavnosti › Trgovina na drobno...
       ...Dejavnosti › Trgovina na debelo...
       ...Dejavnosti › Trgovina
       ...Dejavnosti › Trgovina na debelo...
          Snipet 1
-----------------------------------------------------------
          .... 10.890 trgovina na debelo...
          ...prodajalnah Druga trgovina na drobno...
          ...dejavnosti Druga trgovina na drobno...
          ...ADRIA INVESTICIJE trgovina , posredništvo...
          ...široke porabe Trgovina na debelo...
          ...za gospodinjstvo Trgovina na drobno...
          ...sanitarno opremo Trgovina na debelo...
          ...gospodinjskimi napravami Trgovina na debelo...
          ...za ogrevanje Trgovina na debelo...
Snippets found in: 7.938992023468018s
```

- social services

```
python run_sqlite_search.py "social services"
Querring for:['social', 'services']
Found 4 documents in 0.0014863014221191406s
Frequency      Document name
-----------------------------------------------------------
5              e-uprava.gov.si/e-uprava.gov.si.45.html
```

```
5               e-uprava.gov.si/e-uprava.gov.si.9.html
1               evem.gov.si/evem.gov.si.661.html
1               podatki.gov.si/podatki.gov.si.340.html
      Snipet 0
-------------------------------------------------------
      ...retirement Social services , health...
      ...retirement Social services , health...
      ...and Related Services ( AJPES...
      ...and spa services ltd. TERME...
         Snipet 1
-------------------------------------------------------------------
         ...? Social services , health...
         ...? Social services , health...



Snippets found in: 1.5825088024139404s
```

- vodovod kanalizacija

```
python run_sqlite_search.py "vodovod kanalizacija"
Querring for:['vodovod', 'kanalizacija']
Found 9 documents in 0.0015397071838378906s
Frequency       Document name
-------------------------------------------------------------
11              podatki.gov.si/podatki.gov.si.340.html
3               e-prostor.gov.si/e-prostor.gov.si.54.html
2               e-prostor.gov.si/e-prostor.gov.si.107.html
2               evem.gov.si/evem.gov.si.177.html
1               evem.gov.si/evem.gov.si.371.html
1               podatki.gov.si/podatki.gov.si.185.html
1               podatki.gov.si/podatki.gov.si.189.html
1               podatki.gov.si/podatki.gov.si.431.html
1               podatki.gov.si/podatki.gov.si.434.html
     Snipet 0
-----------------------------------------------------
     ...Javno podjetje Kanalizacija in čistilna...
     ...vodovod , kanalizacija , odlagališča...
     ...gml ) kanalizacija ( shp...
     ...vodnjaki , kanalizacija in septične...
     ...zbiralniki vodnjaki kanalizacija in septične...
     ...za javni vodovod po vodnih...
     ...priključki na vodovod , kohezijske...
     ...za javni vodovod po vodnih...
     ...za javni vodovod po vodnih...
          Snipet 1
-----------------------------------------------------
          ...vodovod , kanalizacija , d.o.o....
```

```
          ...ukine kabelska kanalizacija pod šifro...
          ...gml ) vodovod ( shp...
          ...vodnjaki , kanalizacija in septične...
 Snippets found in: 5.539072751998901s
```

- Slovenija napoved padavin

```
python run_sqlite_search.py "slovanija napoved padavin"
Querring for:['slovanija', 'napoved', 'padavin']
Found 27 documents in 0.0016171932220458984s
Frequency      Document name
-------------------------------------------------------
10             podatki.gov.si/podatki.gov.si.437.html
4              podatki.gov.si/podatki.gov.si.146.html
4              podatki.gov.si/podatki.gov.si.172.html
4              podatki.gov.si/podatki.gov.si.433.html
4              podatki.gov.si/podatki.gov.si.430.html
4              podatki.gov.si/podatki.gov.si.40.html
4              podatki.gov.si/podatki.gov.si.338.html
4              podatki.gov.si/podatki.gov.si.332.html
4              podatki.gov.si/podatki.gov.si.19.html
          Snipet 0
-------------------------------------------------------
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
          ...padavin , napoved ravni onesnaževal...
              Snipet 1
-------------------------------------------------------
              ...padavin , napoved ravni onesnaževal...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
              ...zraka in padavin - PM10...
Snippets found in: 1.4332432746887207s
```

- AVTO ŠOLA za izpit in delavnica

```
python run_sqlite_search.py "avto šola izpit delavnica"
```

```
Querring for:['avto', 'šola', 'izpit', 'delavnica']
Found 46 documents in 0.0016863346099853516s
Frequency      Document name
-----------------------------------------------------------
584            podatki.gov.si/podatki.gov.si.340.html
60             evem.gov.si/evem.gov.si.371.html
25             evem.gov.si/evem.gov.si.653.html
17             e-uprava.gov.si/e-uprava.gov.si.36.html
5              evem.gov.si/evem.gov.si.569.html
4              e-uprava.gov.si/e-uprava.gov.si.16.html
4              evem.gov.si/evem.gov.si.223.html
3              evem.gov.si/evem.gov.si.252.html
3              e-uprava.gov.si/e-uprava.gov.si.53.html
2              evem.gov.si/evem.gov.si.398.html
2              e-uprava.gov.si/e-uprava.gov.si.56.html
       Snipet 0
-----------------------------------------------------------
       .... ISTRA AVTO IN KALISTER...
       ...delov Tapeciranje avto sedežev Globinsko...
       ...B Bibliotekarski izpit C Certificiranje...
       ...1 Za izpit s tolmačem...
       ...ima opravljen izpit za strokovno...
       ...opravlja zaključni izpit ... Visoka...
       ...poseben strokovni izpit ali zaposluje...
       ...opravljen lovski izpit in veljavno...
       ...lahko opravim izpit za voditelja...
       ...imeti mojstrski izpit , da...
       ...je državni izpit , s...
            Snipet 1
-----------------------------------------------------------
            ...SISTEM ZA AVTO Trgovina na...
            ...vozil . Delavnica pridobi pooblastilo...
            ...iz vode Izpit za pooblaščenega...
            ...2 Za izpit s tolmačem...
            ...ima opravljen izpit za strokovno...
            ..... Osnovna šola Kdaj in...
            ...poseben strokovni izpit . Osebi...
            ...ima lovski izpit in veljavno...
            ...register , izpit za voditelja...
            ..., mojstrski izpit ni več...
            ...je državni izpit , s...
Snippets found in: 7.41966986656189s
```

5 Discussion

Results under 3 show us that additional preprocessing steps would improve the relevancy of our index. For example in table 1 we see that there are some words that are not actual

words listed as most frequent. We could use a dictionary of all Slovenian words to check if a token is a word in the preprocesing step of index generation algorithm. In table 3 we see that words that occurred in most documents most likely belong to navigation links and terms and condition boxes which are present on almost all pages. To fix this problem we could compute TF-IDF of words instead of frequencies and we would and get more descriptive results. We computed the TF-IDF measures to validate the assumptions and we see in the table below that the lowest TF-IDF values in fact belong to the aforementioned words. We could also do sequence alignment on html documents and remove html elements that occur in majority of documents in the preprocesing step before we extract the tokens.

	word	document name	TFIDF
	uporabe	../site_data/podatki.gov.si/podatki.gov.si.340.html	4.47393005423591e-07
	pogoji	../site_data/podatki.gov.si/podatki.gov.si.340.html	4.7552471805738e-07
	portalu	../site_data/evem.gov.si/evem.gov.si.371.html	8.92517478816262e-07
	davki	../site_data/evem.gov.si/evem.gov.si.371.html	1.53104199300554e-06
	uporabe	../site_data/evem.gov.si/evem.gov.si.398.html	2.87312668313715e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.57.html	3.56020243993593e-06
	domov	../site_data/evem.gov.si/evem.gov.si.371.html	4.01018225613731e-06
	domov	../site_data/podatki.gov.si/podatki.gov.si.340.html	4.3574777519323e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.150.html	4.36192251302147e-06
	uporabe	../site_data/evem.gov.si/evem.gov.si.371.html	4.55849817373422e-06
	uporabe	../site_data/evem.gov.si/evem.gov.si.377.html	5.53643398824579e-06
	domov	../site_data/evem.gov.si/evem.gov.si.398.html	5.59668365329031e-06
	uporabe	../site_data/evem.gov.si/evem.gov.si.32.html	6.11060473501174e-06
	domov	../site_data/e-prostor.gov.si/e-prostor.gov.si.57.html	6.52479349034552e-06
	upravo	../site_data/evem.gov.si/evem.gov.si.371.html	6.91370748884117e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.54.html	7.48877774695322e-06
	piškotkih	../site_data/evem.gov.si/evem.gov.si.371.html	7.58601083171613e-06
	piškotkov	../site_data/evem.gov.si/evem.gov.si.371.html	7.76802298981197e-06
	domov	../site_data/e-prostor.gov.si/e-prostor.gov.si.150.html	7.99410822797655e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.11.html	8.18340371264895e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.18.html	8.62732853931655e-06
	pogoji	../site_data/e-prostor.gov.si/e-prostor.gov.si.13.html	8.79686119620269e-06
	uporabe	../site_data/evem.gov.si/evem.gov.si.651.html	8.8187944256354e-06
	prijava	../site_data/evem.gov.si/evem.gov.si.371.html	8.89922130336071e-06
	pogoji	../site_data/e-uprava.gov.si/e-uprava.gov.si.31.html	8.97319026249628e-06
	arhiv	../site_data/evem.gov.si/evem.gov.si.371.html	9.04554165395134e-06