Data processing, indexing and querying

Web Information Extraction and Retrieval 2018/19, Faculty of Computer and Information Science, University of Ljubljana

Matej Klemen, Andraž Povše, Jaka Stavanja

***Abstract*—In this work we present our implementation of a simple reverse index and queries against it. First we perform some data preprocessing, tokenization and indexing. With the gathered data and a structured index, we execute two different data retrieval techniques by testing multiple queries and displaying the results. We test the search using the created inverted index and using naive approach and compare the performance of both.**

I. Introduction

Querying the world wide web is a task we perform every day. In the background, popular search engines, such as Google, use databases that contain indexed website data along with some other secrets, to display most relevant results for the query at that specific moment. Therefore, in order to display relevant results, we must first process the information on the website, build an index based on it, and then use this information to display most relevant results. First step is processing the data and is described in chapter II. Next comes the indexing, which is presented in chapter III. Final step is the data retrieval we perform after we are faced with a query. The implementation is described in chapter V. The details about the printing of the output of our system (printing snippets from the files, where our query words are found) are presented in chapter VI. The times needed for performing queries using 2 different search approaches are presented in VII. In chapter VIII we summarize our work and provide some ideas for possible improvement.

II. Data processing

We used the same data processing function for both inverted-index-based and naive querying. First we removed style and script tags from the HTML document. Then we turned the entire content to lowercase letters so we do not have to deal with different representations of the same word. Next step was to tokenize the text using a Slovene tokenizer, available in the NLTK library. We also removed stopwords (again, specific for Slovenian language) because they usually do not hold critical meaning behind a query. Another positive aspect of stopword removal is that we are able to implement our system more efficiently, because we do not have to store (index) them.

III. Data indexing

A. Inverted index

When we have the processed and tokenized text from the websites available, we can then proceed by remembering the offsets of each word in the list of all tokens. For example, if the query word "trgovina" is at the n-th position in the array of tokens, we will save n into the database as the position of the word in the text. Since we will later be outputting snippets based on the text without the HTML elements in it, we can safely just remember the token position. As will be described in the following chapters, the snippet construction will take the input file in which the query word occurs, preprocess it and find the query word at the position that we have saved when creating the index. We only save the offsets in the arrays of tokens for tokens that are in a pre-specified vocabulary. The vocabulary consists of 50000 most frequent (non-stopword) words from the *ccGigafida* corpus, which is a publicly available corpus, containing Slovenian texts of various genres. The use of a limited vocabulary speeds up the building of the index, but specializes the built index somewhat. Since *ccGigafida* is a Slovenian corpus, our vocabulary only contains Slovenian words, so non-Slovenian queries see a drop in quality of results. When we are searching for the occurence of the word in a file, we can go to the Posting table and find the offsets of the word in each file we have found the word in and count them. That contributes to the frequency and will rank a document higher in the search.

B. Naive approach

With the naive approach it is kind of hard to split the process into different sections, because everything happens within a single query.

The entire process starts after we execute a query, when each file is opened and preprocessed. We then calculate the offsets and frequencies of words which appear in the query. Finally, we use the calculated information to display top results based on sum of query word frequencies.

IV. Database

The information about the entries in the database can be seen in Table I. Indexed words are the ones that are present in our vocabulary, which is made from a Slovenian corpus called *ccGigafida*. We used the custom vocabulary because it contains a vast majority of meaningful words, but still reduces the time spent for building the index. The performance of foreign (non-Slovenian) queries is worse because of that, but Slovenian queries work great. There are about 4.5 times as many words as there are indexed words that we are using.

Table I
Database information

Indexed words	50,000
Words	281,553
Maximum frequency	2,266
Most frequent word	"proizvodnja"

V. Data retrieval

A. Inverted index

With the inverted index approach, we used the index we had previously built in order to perform data retrieval. We first split the query into single words and transform each one to lowercase. Then, we search our index for words (SQL query, returning the correct words) which are included in the query and take note of frequencies and offsets of each. Lastly, we sort the results based on sum of query word frequencies and display the top 5 documents.

B. Naive approach

As already written in III, everything in naive approach happens in a single iteration. Data retrieval was performed by splitting the query into words and searching the HTML documents for the occurences of these words using naive string search. Afterwards, we display top results based on the sum of query word frequencies in our documents.

VI. Printing the results

In order to get a clear picture of the results we receive with a specific query, we used the following method for printing.

Upon getting the offset of the query, that we have saved in the database, we tokenize the original text that the word was found in and we search for the three tokens (not counting punctuation as a separate token) in its neighborhood and construct a snippet for the output that is comprised of an ellipsis, followed by the three words with the punctuation symbols to the left of the query word, the query itself and then the three words along with punctuation to the right of the query word. If we have to append a punctuation symbol to the output snippet string, we simply concatenate it to the string, but if we are adding a token (word) to the string, we add a space before it when concatenating it.

We select the first three offsets found in the reverse index and construct the snippets in the way described above and concatenate them into an output for one file under one query phrase. We repeat the process for the top 5 results according to frequency.

For the pretty output used in the final implementation, we use the *tabulate* library for Python to construct a table showing the frequencies and snippets.

VII. Results

The results (outputs of our program) for the given (3) and additionally selected (3) queries can be seen in Listing A.

A. Naive approach VS Inverted index

The results we have shown above were fetched both using an inverted index and naive approach. In Table II, we can see the difference in time elapsed for each approach. Main issue with the naive approach is the file processing we are doing after each query, which takes up a lot of processing time.

Table II
Time for query results using inverted index and naive approach.

Query	**Inverted index**	**Naive approach**
“predelovalne dejavnosti”	6.00 s	60.77 s
“trgovina”	5.99 s	69.94 s
“social services”	1.93 s	67.59 s
“Sistem SPOT”	0.52 s	68.73 s
“davek in dajatve”	0.25 s	72.83 s
“poravnava”	4.18 s	60.45 s

VIII. Conclusion

We presented 2 different approaches to handle a query. First one used inverted index and was substantially faster than the other, naive implementation where we processed each file at time and had no built index. Building the index (SQLite database) did not take that much time compared to 0 with naive approach, but the end result (when measuring time elapsed for each query) was much better. Our index was built on a closed vocabulary of 50000 Slovenian words, which brought along its upsides and downsides, that were already discussed in the report. One possible improvement would be the removal of this closed vocabulary and building the index on a possibly unlimited set of words, that would be encountered in processed websites. This would increase the time that our index would take to be built, but would make search results better for general queries.

Another obvious improvement of our system could be the use of multithreadding to process the documents as a big bottleneck of our system are input-output operations.

The implementation of the indexer and retriever can be found on https://github.com/matejklemen/pa3.

Appendix

Listing 1. Query search outputs

```
Results for query:  predelovalne dejavnosti

  Frequency  Document                 Snippet
-----------  -----------------------  ---------------------------------------------------------------------------
       1288  evem.gov.si.371.html     ...za infrastrukturo c [predelovalne] dejavnosti 10 proizvodnja...32 druge raznovrstne [predelovalne] dejavnosti 32.110
             kovanje...32.990 drugje nerazvr?ene [predelovalne] dejavnosti sem spada...
         75  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...
         40  podatki.gov.si.340.html  ...kalan - nosilec dopolnilne [dejavnosti] na kmetiji bregar... port center interesnih [dejavnosti] ptuj center judovske...
               olskih in ob olskih [dejavnosti] center urbane kulture...
         39  evem.gov.si.452.html     ...druge storitvene [dejavnosti] , drugje nerazvr?ene ( 96.090...drugje nerazvr?ene ( 96.090 ) / [dejavnosti] / evem
               republika slovenija...e-vem evemdejavnostidruge storitvene [dejavnosti] , drugje nerazvr?ene ( 96.090...
         31  evem.gov.si.653.html     ...dovoljenje za opravljanje [dejavnosti] specializirane prodajalne z...radijske ali televizijske [dejavnosti] dovoljenje
               za izvajanje...za izvajanje sevalne [dejavnosti] dovoljenje za izvajanje...

Results found in 6.00371s using inverted index and in 60.77384s using naive approach...
----------------------------------------
Results for query:  trgovina

  Frequency  Document                 Snippet
-----------  -----------------------  ---------------------------------------------------------------------------
        364  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...
         94  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...
         92  evem.gov.si.21.html      ...moj e-vem evempodro?ja [trgovina] tu boste na li...seznam dejavnosti druga [trgovina] na drobno v...nespecializiranih
               prodajalnah druga [trgovina] na drobno zunaj...
         82  podatki.gov.si.340.html  ...d.o.o . a dent , [trgovina] in storitve , d.o.o...d.o.o . adria investicije [trgovina] , posredni tvo , storitve in...
               storitve d.o.o . ahatservis [trgovina] in storitve , d.o.o...
         13  evem.gov.si.623.html     ... [trgovina] na debelo z...izdelki iroke porabe [trgovina] na debelo z...porabe sem spada : [trgovina] na debelo z...

Results found in 5.99337s using inverted index and in 69.93808s using naive approach...
----------------------------------------
Results for query:  social services

  Frequency  Document                 Snippet
-----------  -----------------------  ---------------------------------------------------------------------------
          5  e-uprava.gov.si.45.html  ...culture labour , retirement [social] services , health , death...employment relationship etc. ? [social] services ,
                health , death...i obtain financial [social] assistance ? how do...
          5  e-uprava.gov.si.9.html   ...culture labour , retirement [social] services , health , death...employment relationship etc. ? [social] services ,
                health , death...i obtain financial [social] assistance ? how do...
          1  evem.gov.si.661.html     ...records and related [services] ( ajpes ) and the...
          1  podatki.gov.si.340.html  ...recreation and spa [services] ltd . terme maribor...

Results found in 1.92837s using inverted index and in 67.59124s using naive approach...
----------------------------------------
Results for query:  Sistem SPOT

  Frequency  Document                  Snippet
-----------  ------------------------  ---------------------------------------------------------------------------
         69  evem.gov.si.68.html       ...je v pripravi . [sistem] spot , slovenska poslovna...to?ka nudi celovit [sistem] brezpla?nih storitev za...ima
               vlagatelj : vzpostavljen [sistem] naro?anja strank , prostor...
         38  evem.gov.si.63.html       ...uvaja nov nacionalni [sistem] spot , slovenska poslovna...prehaja tudi dosedanji [sistem] vem , tako to?ke...pod enotno
                znamko . [sistem] spot bo poslovnim...
         32  e-prostor.gov.si.18.html  ...dr avni prostorski koordinatni [sistem] / epsg kode za...datum in koordinatni [sistem]   ks , ki...geodetski datum 1996
                , [sistem] trirazse nih kartezi?nih koordinatepsg...
         24  e-prostor.gov.si.57.html  ...se prijaviti v [sistem] . uporabni ko ime ni...ker pa na  [sistem] preverja pravilnost serijske...za?etno stran in [
               sistem] bi moral delovati...
         23  e-prostor.gov.si.24.html  ...dr avni prostorski koordinatni [sistem] ta stran uporablja...informacijedr avni prostorski koordinatni [sistem] domov /
                informacije / vsa...dr avni prostorski koordinatni [sistem] poro?anje o prodajnih...

Results found in 0.52443s using inverted index and in 68.73898s using naive approach...
----------------------------------------
Results for query:  davek in dajatve

  Frequency  Document                 Snippet
-----------  -----------------------  ---------------------------------------------------------------------------
         19  evem.gov.si.7.html       ...dajatve , tro arine in [davek] na dodano vrednost...v sloveniji ddv ( [davek] na dodano vrednost...9,5 % - ni ja stopnja
                [davek] od dohodkov pravnih...
          5  e-uprava.gov.si.52.html  ...kdaj se pla?uje [davek] na promet nepremi?nin...kdaj moram pla?evati [davek] na vodna plovila...kdaj moram pla?evati [
                davek] na dedi ?ino , v...
          5  evem.gov.si.71.html      ...vodenje podjetja / davki / [davek] na dodano vrednost...dodano vrednost ( ddv ) [davek] na dodano vrednost...za d.o.o .
                za?ni [davek] na dodano vrednost...
          5  evem.gov.si.9.html       ...vodenje podjetja / davki / [davek] od dohodka pravnih...dohodka pravnih oseb [davek] od dohodka pravnih...je osnova za [
                davek] rezidenta in poslovne...
          4  evem.gov.si.72.html      ...vodenje podjetja / davki / [davek] od dohodka iz...dohodka iz dejavnosti [davek] od dohodka iz...postanete zavezanec za
                [davek] od dohodka iz...

Results found in 0.24758s using inverted index and in 72.82933s using naive approach...
----------------------------------------
Results for query:  poravnava

  Frequency  Document              Snippet
-----------  --------------------  ---------------------------------------------------------------------------
          5  evem.gov.si.660.html  ...gospodarske dru be / prisilna [poravnava] republika slovenija spot...evemzapiramzapiranje gospodarske dru beprisilna [
                poravnava] prisilna poravnava prisilna...dru beprisilna poravnava prisilna [poravnava] prisilna poravnava je...
          2  evem.gov.si.387.html  ...registra dru b . prisilna [poravnava] v postopku prisilne...likvidacija dru be prisilna [poravnava] ste?aj dru be
                zakonodajni...
          1  evem.gov.si.274.html  ...mre e proti to?i ; [poravnava] zemlji ?a , nalaganje in...
          1  evem.gov.si.371.html  ...mre e proti to?i [poravnava] zemlji ?a , nalaganje in...
          1  evem.gov.si.388.html  ...likvidacija dru be prisilna [poravnava] ste?aj dru be odjava...

Results found in 4.18491s using inverted index and in 60.45122s using naive approach...
----------------------------------------
```