

Assignment 3: Web Indexing

Gal Lindič, Andrej Drofenik, Ožbej Golob

I. INTRODUCTION

Web indexing means creating indexes for individual Web sites, intranets, collections of HTML documents, or even collections of Web sites. Indexes are systematically arranged items, such as topics or names, that serve as entry points to go directly to desired information within a larger document or set of documents. In this paper, we built a simple index and implemented querying against it. We extracted textual context from the provided HTML files, performed preprocessing and stored data into index. After that we used the built index for querying. Our implementation consisted of the following parts: (i) Data processing and indexing and (ii) Data retrieval.

II. IMPLEMENTATION

A. Data processing and indexing

First, we created the required database tables (*IndexWord* and *Posting*) with the provided SQL statement. For indexing we firstly retrieved text data from all web pages. We iterated all web pages, created an instance of *Document* object for each web page, and added it to an array that contained instances of *Document* object for each web page.

The *Document* class has a function *_process_document* that processes the document and returns postings. The function retrieves text with the *Beautiful Soup* module (*soup.get_text()*). The text words are tokenized with the *word_tokenize* function from the *nltk* module. The function then iterates through the tokenized words, makes them lower-case, removes stopwords, and adds the word to the postings dictionary. The postings dictionary keys are lower-case words and values are another dictionary with frequencies of this word in current document and the indexes of the word in current document.

After processing all web pages, we iterated through the array that contained the instances of *Document* object for each web page and called the *insert_posting* function. The function iterated through all words and their postings for the current document. It inserted the word into the *IndexWord* table (if it didn't already exist), joined the indexes array by comma into a string, and inserted the word, document name, frequency, and indexes into the *Posting* table.

B. Data retrieval with inverted index

For the data retrieval we used the index built in the II-A section. We split the query string into words, tokenize them, make them lower-case, and remove the stopwords. The result is an array that contains the remaining words. Then we use SQL to select all rows from the *Posting* table where the attribute *word* is equal to one of the search query words.

We then iterate through all the postings that we just retrieved. We create a dictionary called *results_dict*, where the key is site name and the value is another dictionary with frequency of query words, their indexes, and site name as keys. That way we merge the results by document (for multi-word queries). We then sort the *results_dict* by frequency and name descending.

For the purpose of getting snippets we created another table *DocumentText*, where we saved all web page's document text (*soup.get_text()*).

```
CREATE TABLE DocumentText (
    documentName TEXT NOT NULL PRIMARY KEY,
    text TEXT,
)
```

For each item in the sorted results array we get the document text from the *DocumentText* table using SQL and tokenize it. We iterate the indexes of search query words and select the neighbourhood of three words to form the snippets for each index.

We then print out the results in the required format.

C. Data retrieval without inverted index

For processing the search query we used the same process as in the II-B section. We then iterate through all web pages and process them in the same way as in the II-A section (we create an array of *Document* object instances for each web page and create postings for each word in the document in the *_process_document* function).

We then iterate each instance of the *Document* object and refer to it as *doc*. We check if each search query word is in the *doc.postings.keys()* array that contains all words in the current *doc*. If it is, we increment the *frequency* variable (that stores the frequency of all search query words in the *doc*) for the amount of frequency of this word. We also extend the *indexes* array (that stores the indexes of all search query words in the current array) with indexes of this word.

If *frequency* value is larger than 0, we create a *document* dictionary that contains the frequency of all search query words in current document, their indexes, current document name, and tokenized words of the current document (needed for snippets). We append the *document* dictionary to the *results_dict* where the key is site name and the value is *document* dictionary. That way we merge the results by document (for multi-word queries). We then sort the *results_dict* by frequency and name descending.

For each item in the sorted results array we get the document text from the item's value dictionary *text* field. We iterate the indexes of search query words and select the neighbourhood of three words to form the snippets for each index.

We then print out the results in the required format.

III. DATABASE

The *IndexWord* contains 49,082 indexed words and the table *Posting* contains 402,988 postings. The words that appeared in at least 1000 documents are: ('*pogoji*', 1398), ('*domov*', 1384), ('*prostor*', 1349), ('*sistem*', 1285), ('*pomoč*', 1269), ('*davki*', 1257), ('*zdravje*', 1257), ('*državni*', 1238), ('*varnost*', 1210), ('*zaposlovanje*', 1206), ('*republike*', 1165), ('*slovenije*', 1054). The words that appeared at least 1000 times in a single document are: ('*proizvodnja*', 2266), ('*gl*', 1668), ('*spada*', 1338), ('*dejavnosti*', 1284). The documents with the highest frequency of all words are: ('*evem.gov.si.371.html*', 103206), ('*podatki.gov.si.340.html*', 32154), ('*e-prostor.gov.si.166.html*', 11202), ('*e-prostor.gov.si.147.html*', 10375), ('*podatki.gov.si.511.html*', 6996).

IV. RESULTS

Table I shows the execution times of search queries for queries with index and without index.

Table I
EXECUTION TIMES OF SEARCH QUERIES.

Query	With index (s)	Without index (s)
predelovalne dejavnosti	10.378	65.946
trgovina	6.843	68.398
social services	1.393	64.291
velenje	1.853	89.588
prometne nesreče	6.355	65.267
študentski servis	4.380	75.534
Average	5.200	71.504

We can see that the average time for our 6 queries with index is 5.200 seconds and without index it's 71.504 seconds. That means that we can find a query more than 13-times faster by using index.

Figures 1-6 show the output of data retrieval with inverted index and Figures 7-12 show the output of data retrieval without inverted index. The full output is also available on *github repository* in the *printouts* folder.

V. CONCLUSION

In this paper, we built a simple index and implemented querying against it. We extracted textual context from the provided HTML files, performed preprocessing and stored data into index. After that we used the built index for querying. We measured times for 6 queries with and without index. The average query time with index was 5.200 seconds while without index it was 71.504 seconds. That means that by using index we can find a query more than 13-times faster.

Figure 1. Search query with index on "predelovalne dejavnosti".

Figure 2. Search query with index on "trgovina".

Frequencies	Document	Snippet
5	e-sprava.gov.si/e-uprava.gov.si-5.html	... , retirement Social services , health , ... etc. ? Social services , health , ... Labour , retirement Social services , health ... relationship etc. ? Social services , health ... I obtain financial social assistance ? How ...
5	e-uprava.gov.si/e-uprava.gov.si-45.html	... , retirement Social services , health , ... etc. ? Social services , health , ... labour , retirement Social services , health ... relationship etc. ? Social services , health ... I obtain financial social assistance ? How ...
1	podatki.si/podatki.gov.si-348.html	... recreation and spa Services Ltd. TERME MARIBOR ...
1	even.gov.si/even.gov.si-681.html	... Records and Related Services (AJPES) ...

Figure 4. Search query with index on "velenje".

Frequencies	Document	Snippet
Results found in 6355ms.		
9	enev.gov.si/enev.gov.si_371.html	... 26,510 proizvodnja električne prometne signalizacije ; semaforjev ... razsvetljavo (razen prometne signalizacije) proizvodnja ... razsvetljivo (razen prometne signalizacije) proizvodnja ... cigaret Izdelava električne prometne signalizacijske opreme 28 ...
9	podatki.gov.si/podatki.gov.si_261.html	... In infrastruktura Cestoprometne nesreč v delovanju v ... Cestoprometne nesreč ... Nadaljujte z ... in infrastruktura Cestoprometne nesreč v delovanju v ... naslovom " Cestoprometne nesreč ... Nadaljujte z ...
4	podatki.gov.si/podatki.gov.si_465.html	... in infrastruktura Cestoprometne nesreč v delovanju v ... naslovom " Cestoprometne nesreč ... Nadaljujte z ... in infrastruktura Cestoprometne nesreč v delovanju v ... naslovom " Cestoprometne nesreč ... Nadaljujte z ...
3	podatki.gov.si/podatki.gov.si_443.html	... priznanja na področju prometne preventivne in prizadevanju ... na področju zagotavljanja prometne varnosti 29 ogledov ... skrbijo za izvajanje prometne preventivne in koordinacijo
3	podatki.gov.si/podatki.gov.si_476.html	... In druge spremljajoče prometne dejavnosti (52,290 ... In druge spremljajoče prometne dejavnosti (52,290 ... In druge spremljajoče prometne dejavnosti (52,290 ...
3	enev.gov.si/enev.gov.si_123.html	... starkev, hujše nesreč) skupaj z ... ; inženirske, prometne , komunikacijske ...
2	podatki.gov.si/podatki.gov.si_389.html	... ali pa tudi vseh prepotencialnih lokalno vreme , prometne nesreč ali sterilno ...
1	enev.gov.si/enev.gov.si_81.html	
1	enev.gov.si/enev.gov.si_239.html	... ; inženirske , prometne , komunikacijske ...
1	enev.gov.si/enev.gov.si_177.html	... Postavite zadanes prometne signalizacije (52,290 ...
1	enev.gov.si/enev.gov.si_177.html	... Postavite zadanes prometne signalizacije (52,290 ...
1	e-uprava.gov.si/e-uprava.gov.si_56.html	tek stoj za stojanje prometne preklicke lahko pridobite ...
1	e-uprava.gov.si/e-uprava.gov.si_55.html	... e-mail: info@e-uprava.gov.si ...
1	e-pristop.gov.si/e-pristop.gov.si_11.html	... uporabljaju z običajni prometne in komunalne infrastrukture ...

Figure 5. Search query with index on "prometne nesreće".

Figure 6. Search query with index on "študentski servis"

Figure 7. Search query without index on "predelovalne dejavnosti".

Figure 8. Search query without index on "trgovina".

Results for a query: "social services"
Results found in 1398ms.

Frequencies	Document	Snippet
1	...e-pravva.gov.si/e-pravva.gov.si.9.html	... retirement Social services ; health , ... etc ? Social services , health ; ... Labour , retirement Social services , health ... relationship etc. ? Social services , health ... I obtain financial social assistance ? How ...
1	e-pravva.gov.si/e-pravva.gov.si.45.html	... , retirement Social services , health , ... etc ? Social services , health ; ... Labour , retirement Social services , health ... relationship etc. ? Social services , health ... I obtain financial social assistance ? How ...
1	podatki.gov.si/podatki.gov.si.1340.html	... recreation and spa services Ltd. TERME MARIBOR ...

Figure 9. Search query without index on "social services"

Figure 10 Search query without index on "velenie"

Results for a query: "promete nesreće"
Results found in 65267ms.

Frequencies	Document	Snippet
9	even.gov.si/even.gov.si.371.html	... 26.510 ponuja elektrinske premostne signalizacije , semaforje , ... razsvetljivo (ravn premostne signalizacije) premostaja , ... razsvetljivo (ravn premostne signalizacije) premostaja ... cigaret Izdelava elektrinske premostne signalizacijske opreme 28 ... in infrastrukturi Cestnoprmetne nesreće in udeleženc v ... mativno ... Cestnoprmetne nesreće ... Nadaljujte z ... in infrastrukturi Cestnoprmetne nesreće in udeleženc v ... mativno ... Cestnoprmetne nesreće ... Nadaljujte z ... in infrastrukturi Cestnoprmetne nesreće in udeleženc v ... mativno ... Cestnoprmetne nesreće ... Nadaljujte z ...
9	even.gov.si/even.gov.si.392.html	... podatki.gov.si/podatki.gov.si.446.html
9	even.gov.si/even.gov.si.446.html	... podatki.gov.si/podatki.gov.si.446.html
9	even.gov.si/even.gov.si.476.html	... podatki.gov.si/podatki.gov.si.476.html
9	even.gov.si/even.gov.si.479.html	... podatki.gov.si/podatki.gov.si.479.html
9	even.gov.si/even.gov.si.399.html	... vrem , ... premostne nesreće ali Strelivo brezposlovnih ... lokacije vrem , ... premostne nesreće ali Strelivo ...
1	even.gov.si/even.gov.si.81.html	starkev , ... hujše nesreće) skupaj do ...
1	even.gov.si/even.gov.si.10.html	... vrem , ... premostne nesreće ali Strelivo ...
1	even.gov.si/even.gov.si.18.html	... in druge spremljajo premostne dejavnosti (26.290 ...
1	even.gov.si/even.gov.si.77.html	... Postavite začasne prometne signalizacije oz ...
1	e-prosten.gov.si/e-prosten.gov.si.56.html	... ne vozilca , ... prometne površine. (€ 80 ...
1	e-prosten.gov.si/e-prosten.gov.si.125.html	... upravljanju z objekti prometa in komunalni infrastrukture ...
1	e-prosten.gov.si/e-prosten.gov.si.111.html	

Figure 11. Search query without index on "prometne nesreće"

Results for a query: "Studentski servis"		
Results found in 75544ms.		
Frequencies Document Snippet		
1	UJEDNOSTAVLJENA AGENCIJA M SERVIS, Kadevskie storitve ... , trgovina in servis, d.o.o. LJUBLJANA RIROTEHNIKA , servis in vzdrževanje računalniške DIMIKARSTVO , SERVIS IN MONTAŽA , NOTARKA E.D.T . SERVIS , vzdrževanje energetskih ,... ELEKTRONIK KRA 2 VODOMEROV BUDIANT PODPREDSEK ... d.o.o. STUDENTSKE SERVIS - posredovanje delovne ... d.o.o. STUDENTSKI SERVIS TRBOVLJE d.o.o. , ... d.o.o. TALIM SERVIS IN INŽENIRING , ... SISTEMI JOS , servis tehničic in ustoli ... pospeševanje turizma TUR SERVIS , turistična agencija ... 3 e-prosten.gov.si/e-prosten.gov.si.138.html ... sestavljanje Popravila , servis potopnih črpalk , hibridne opreme) Servisne tehničnih aparatorov in ... aparator v servis . Servis je popularno obiskovan na spletni strani : //www.e-prosten.gov.si/138.html . Vsebina: WPS Office , servis MTS motorov , plinogeneratorjev , ... biloprovod . Popravilo in se 4 e-prosten.gov.si/e-prosten.gov.si.139.html ... Izdelava je servis za kontrolo obstoja ... več Javni spletni servis (WMTS) ... povezava na spletni servis (MTS) ... OTH . Servis je na voljo ... DB4/GK in DB6/TM.Spleteti servis WMTS (Web ... Web Map Tile Servis) omogoča vpogled ... 5 e-prosten.gov.si/e-prosten.gov.si.140.html ... jo prikriboj Studentski servis . Napovedana je ... vrne na studentski servis , ta pa ... pooblaščene organizacije (Studentski servisi , Zavod WPS Office , servis na spletni servis (WMTS) ... OTH . Servis je na voljo ... DB4/GK in DB6/TM.Spleteti servis WMTS (Web ... Web Map Tile Servis) omogoča vpogled ... 6 e-prosten.gov.si/e-prosten.gov.si.141.html ... podatki.gov.si/podatki.gov.si.559.html ... posebno podlage ali servisi , ki bo ... 7 e-prosten.gov.si/e-prosten.gov.si.142.html ... podatki.gov.si/podatki.gov.si.559.html ... posebno podlage ali servisi , ki bo ... 8 e-prosten.gov.si/e-prosten.gov.si.143.html ... podatki.gov.si/podatki.gov.si.559.html ... podatki.gov.si/podatki.gov.si.559.html ... izbrati Kakovosten računalovski servis . Kdo pa ... 9 e-prosten.gov.si/e-prosten.gov.si.144.html ... izbrati Kakovosten računalovski servis . Kdo pa ... 10 e-prosten.gov.si/e-prosten.gov.si.145.html ... SLOVENE) . Servisna apacija QM ... 11 e-prosten.gov.si/e-prosten.gov.si.146.html ... izbrati Kakovosten računalovski dom lahko ustvarjuje ... 12 e-prosten.gov.si/e-prosten.gov.si.147.html ... ; Popravilo , servis potopnih črpalk ; ... 13 e-prosten.gov.si/e-prosten.gov.si.148.html ... ; Strošek : Servisna apacija za renging ... 14 e-prosten.gov.si/e-prosten.gov.si.149.html ... Izdelava tehnoloških sistemov za obstoja ... 15 e-prosten.gov.si/e-prosten.gov.si.124.html ... Web Map Tile Servis) , ki ...	

Figure 12. Search query without index on "študentski servis".