Coding assignment 3

Web information extraction and retrieval

Jošt Gombač, Marko Krajinović, Rok Petrovčič Vižintin

1 Inverted index

For each page we do the following:

1.1 Text extraction

For text extraction we use *BeautifulSoup*. We also include content of meta tags with *name* = {*title, keywords, description*}.

1.2 Pre-processing

We first transform text to lowercase. We then extract tokens using *nltk.word_tokenize*. For each token we calculate its character based offset index. This is the index at which the token starts. Lastly, we remove all stopwords. We also remove all non-alphanumeric tokens (punctuations).

1.3 Appending to index

We implemented a simple interface for database communication. For each token, we check if it already exists in the table *IndexWord*. We do the same for table *Posting*, where we check for document as well. For each word occurrence in the document, we simply increment its frequency and append the relevant index to the end, keeping it sorted.

2 Database

We collected 32308 unique words. The following table contains 10 words with highest frequencies.

Frequncy	**Word**	**Document**
2266	proizvodnja	evem.gov.si/evem.gov.si.371.html
1668	gl	evem.gov.si/evem.gov.si.371.html
1338	spada	evem.gov.si/evem.gov.si.371.html
1290	dejavnosti	evem.gov.si/evem.gov.si.371.html
927	xsd	e-prostor.gov.si/e-prostor.gov.si.147.html
809	skupnost	podatki.gov.si/podatki.gov.si.340.html
754	krajevna	podatki.gov.si/podatki.gov.si.340.html
589	ministrstvo	evem.gov.si/evem.gov.si.371.html
582	šola	podatki.gov.si/podatki.gov.si.340.html
545	dejavnost	evem.gov.si/evem.gov.si.371.html

The following table contains 10 documents with most words.

Frequency	Document
77361	evem.gov.si/evem.gov.si.371.html
25423	podatki.gov.si/podatki.gov.si.340.html
3985	evem.gov.si/evem.gov.si.398.html
3397	e-prostor.gov.si/e-prostor.gov.si.57.html
3375	e-prostor.gov.si/e-prostor.gov.si.147.html
2975	podatki.gov.si/podatki.gov.si.511.html
2779	e-prostor.gov.si/e-prostor.gov.si.150.html
2606	evem.gov.si/evem.gov.si.651.html
2510	evem.gov.si/evem.gov.si.653.html
2247	e-uprava.gov.si/e-uprava.gov.si.56.html

3 Data retrieval

3.1 Querying the index

The data retrieval process takes a search query string. The given string is then pre-processed using the same methods from section 1.2. After the array of tokens is returned, they are used to query the inverted index. The *Posting* table is queried for postings which contain any of the tokens. It is then grouped by the document name and filtered to only include documents, which contain all the tokens in the search query. Lastly the results are ordered by the sum of the frequencies and include all indexes of every token. In this manner, the desired results are attained with a single query of the SQLite database.

3.2 Snippets

We used the token indexes to display snippets along with the search results. The corresponding document is opened and the text around the given character index is returned. The snippets stop after a certain number of words are displayed or when the beginning or and of a sentence is reached.

4 Data retrieval without an inverted index

This method uses similar logic to the indexed one. The main difference being, that we don't query a database. It searches for tokens by opening every HTML document sequentially. The document's text is then pre-processed into a list of words and their indices via the same method used when building the inverted index. Each of the search query's input tokens is then matched against all the words in the text. If a document contains at least 1 occurrence of each token, it is then saved in the list of results. The frequency of each token in a given document is summed up. Each result also holds the indices at which each token appears in the document. Snippets are then generated in the same way as in the previous method.

5 Results

Index building took approx. 3 hours. We used the same processing logic for both methods of data retrieval, so both gave us the same results. The only difference was the speed of execution. We will detail the results of each query, and compare the time taken between the two methods. Snippets will also be displayed, but they will be shortened, so the tables render properly. Outputs of both methods are saved in the file *results.txt* in the root directory of the project.

5.1 Queries

Query	**Basic time**	**Index time**
predelovalne dejavnosti	49928ms	12ms
Frequency	**Document**	**Snippet**
1294	evem.gov.si/evem.gov.si.371.html	... dejavnosti, Izpisanih, SKD, Šifra, ...
23	evem.gov.si/evem.gov.si.460.html	... proizvodnja, dejavnosti, ipd...
13	evem.gov.si/evem.gov.si.15.html	... motornih, vozil, kovin, dejavno...

Query	**Basic time**	**Index time**
trgovina	48159ms	2ms
Frequency	**Document**	**Snippet**
364	evem.gov.si/evem.gov.si.371.html	46.110 trgovina na debelo s ...
97	evem.gov.si/evem.gov.si.651.html	... Trgovina, debelo, drobno, Izvajanje, ...
94	evem.gov.si/evem.gov.si.21.html	... Trgovina, debelo, drobno, prodajalnah, ...
82	podatki.gov.si/podatki.gov.si.340.html	... trgovina in storitve, d.o.o.
19	evem.gov.si/evem.gov.si.623.html	... debelo, trgovina, dejavnosti, opravljanje ...

Query	**Basic time**	**Index time**
social services	47844ms	1ms
Frequency	**Document**	**Snippet**
5	e-uprava.gov.si/e-uprava.gov.si.45.html	... culture Labour, retirement Social ...
5	e-uprava.gov.si/e-uprava.gov.si.9.html	... culture Labour, retirement Social ...

Query	**Basic time**	**Index time**
sistem SPOT	48260ms	15ms
Frequency	**Document**	**Snippet**
75	evem.gov.si/evem.gov.si.68.html	Sistem SPOT, Slovenska poslovna ... SPOT, ...
42	evem.gov.si/evem.gov.si.63.html	... uvaja nov nacionalni sistem SPOT, ...
37	evem.gov.si/evem.gov.si.67.html	... in podpora I Sistem SPOT I Seznam ...
25	evem.gov.si/evem.gov.si.49.html	... sistema, sistem, VEM, strani, uporabe, ...
23	evem.gov.si/evem.gov.si.398.html	... ga sistem ne podpira. ...

Query	**Basic time**	**Index time**
EVEM	48095ms	9ms
Frequency	**Document**	**Snippet**
6	evem.gov.si/evem.gov.si.398.html	... da bo ... eVEM / Pomoč in ...
4	evem.gov.si/evem.gov.si.406.html	... , obrazec M-2, eVEM, ZZZS, rok za ...
4	evem.gov.si/evem.gov.si.48.html	... z osebnim premoženjem, eVEM, odjava, ...
4	evem.gov.si/evem.gov.si.658.html	... registra Slovenije, PRS, eVEM, obrazec ...
4	evem.gov.si/evem.gov.si.84.html	... eVEM / Vodenje podjetja ...

Query	**Basic time**	**Index time**
Registracija samostojnega podjetnika	48128ms	3ms
Frequency	**Document**	**Snippet**
18	evem.gov.si/evem.gov.si.23.html	... Registracijo samostojnega ...
11	evem.gov.si/evem.gov.si.373.html	... Registracija samostojnega ...
10	evem.gov.si/evem.gov.si.35.html	...Registracija samostojnega ...
10	evem.gov.si/evem.gov.si.37.html	... Vpis podjetnika ...
10	evem.gov.si/evem.gov.si.6.html	... Začni Registracija ...