Programming Assignment 3

Implementing a data index and more

Jaka Kokošar, Danijel Maraž, and Toni Kocjan

Fakulteta za Računalništvo in Informatiko UL dm9929@student.uni-lj.si, jk0902@student.uni-lj.si, tk3152@student.uni-lj.si

Povzetek The article covers the work done in the scope of the third programming assignment as part of the subject web information extraction and retrieval.

Keywords: Data Processing Indexing Retrieval

1 Introduction

After having collected web pages in the first assignment and thoroughly stripped them down to the data we are interested in in the second we were ready to continue in the final step which is constructing a data index and implementing the function of querying.

2 Data Processing

Glavna funkcija *preprocess* prejme kot argument niz *raw_text* in mu:

– S funkcijo *remove_punctuation* odstrani ločila in več
– Z *nltk.tokenize.word_tokenize* pretvori v vrsto besednih značk
– Odstrani se značke, ki niso alfabetične
– Vse velike začetnice se pretvorijo v male
– Odstrani se značke, ki so štopword"
– S pomočjo pretvorbe v podatkovno strukturo množice se odstranijo duplikati

3 Indexing

4 Data Retrieval