Technische Universität München
Fakultät für Informatik
Lehrstuhl für Effiziente Algorithmen
Sandeep Sadanandan

Sommer Semester 2009
Problem Sheet 7
Jun 19, 2009

Python For Fine Programmers

Deadline: July 2, 2009

From the previous problem set, we have implemented a program to generate a graph of URLs from starting from an initial URL. In this exercise sheet, we are going to develop it further.

Problem 1 (3 Points)

Implement an HTMLParser, so that for every node (URL node) in the graph, the parser could filter out the text-contents of the URL-page.

Problem 2 (3 Points)

Once the HTMLParse is in place, use the details given in the lecture, to generate the `tf` values of the words in the document.

For further use, the `tf` values are to be stored in a `shelve` object.

Problem 3 (4 Points)

Once the `tf` values of all words and documents are in place, then using the information from the lecture, generate the `tf-idf` values for the word-document pair.

Note/Bonus: Design the whole exercise as a class object called Crawler or Spider. The Crawler class should be able to update itself in case of events like a change in the contents of a file or addition/deletion of a file.