Institut für Informationssysteme
Technische Universität Braunschweig

Prof. Dr. Wolf-Tilo Balke and Kinda El Maarry

Homework Assignment 9

Due date: 27th of January 2014

Please note that even though the homework assignments are optional, you're still highly encouraged to answer them, as they will help you prepare for your final exam. You can work in a group of two or alone. Solutions can be dropped off at the institute's homework mailbox located on the 2nd floor, next to room 238. In that case, please make sure both your name and matriculation number is noted down. If your answers span more than one sheet, kindly staple them together. Another alternative is to send your homework via email to: elmaarry@ifis.cs.tu-bs.de

---

LECTURE 11: WEB CRAWLING

EXERCISE 11.1

What's the purpose of large-scale web crawlers? And how do they differ from focused crawlers?

EXERCISE 11.2

Why do large-scale crawlers maintain a local DNS component?

EXERCISE 11.3

Fingerprinting is applied to check for possible duplication of URIs. Can we also use fingerprinting to check for possible duplication of content? Why?

EXERCISE 11.4

How does shingling's basic idea help in detecting near duplicates? (Avoid merely describing how shingling works, and justify why or how this approach can capture the duplication of content)

EXERCISE 11.5

In near-duplicate content detection, the process of shingling and computing the corresponding Jaccard coefficient is inefficient. How is this problem handled?

EXERCISE 11.6

Explain in your own words, and as per your understanding, why each of the following holds:

- $J(S(d), S(d')) \approx J(H(d), H(d'))$
- $J(H(d), H(d')) = Pr(\min(\Pi(d)) = \min(\Pi(d')))$