



ifis

Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval & Web Search Engines

HW: Web Crawling

Please note: even though the homework assignments are optional, you are encouraged to answer them, as they will help you prepare for your final exam. Solutions can be dropped off at the institute's homework mailbox located on Informatikzentrum 2nd floor, next to room 238. Please do not forget your **Matrikelnummer** and your **full name** on your solutions. Another alternative is to send your homework by email to: pinto@ifis.cs.tu-bs.de. Please write your solutions in English. **Due date:** before next lecture (06.07.2017).

- 1.-What's the purpose of large scale web crawlers? And how do they differ from focused crawlers?
- 2.-Why do crawlers need to check for duplicate URIs and how do they do it?
- 3.-How does shingling's basic idea help in detecting near duplicates?