Programmierprojekt zu „Einführung in die Informatik 2“

– Automatische Datenextraktion aus HTML-Dokumenten –

Sven Kosub

AG Algorithmik/Theorie komplexer Systeme
Universität Konstanz

E 202 | Sven.Kosub@uni-konstanz.de | Sprechstunde: Mittwoch, 14:00-15:00 Uhr, o.n.V.

Sommersemester 2008

2. Aufgabenblock

- Entwerfen Sie einen exakten Wrapper „bestseller.wrp“ in WDL (Wrapper Description Language), der für die Bücher-Bestseller bei Amazon.de folgende Daten in einer CSV-Datei ausgibt:
  - Titel des Buches
  - erster Autor des Buches (ohne „(Autor)“-Angabe)
  - Neupreis des Buches (ohne Währungsangabe)

  (Referenz ist die Datei „bestseller.html“ von der Vorlesungs-Webseite)
- Erweitern Sie die WDL um
  - Kommentare
  - Groß- und Kleinschreibung
  - erlaubte Leerzeichen
  - Reihenfolge der Klauselbeschreibungen
- Implementieren Sie einen Ein-Pass-Algorithmus zur Anwendung des geladenen Wrappers auf das HTML-Dokument
- Verbessern Sie die Stabilität (Korrektheit) Ihrer Implementierung

Organisatorisches

Ablauf:

- zweiter Aufgabenblock bis 23.07.2008, um 14:00 Uhr
- Dotierung 30 Punkte
- ZIP-Pools G 228-230

Hilfsmittel:

- http://java.sun.com/j2se/1.5.0/docs/api/
- http://www.eclipse.org/
- http://www.inf.uni-konstanz.de/algo/lehre/ss08/info2/