Übung IR: Tika

May 24, 2009

Installation von Apache Tika

- Installieren Sie zuerst maven 2 und setzten Sie die Pfadvariable (z.B. in .profile)

  ```
  export PATH=$PATH:/path/to/maven/bin/
  ```

- Installieren Sie Apache Tika; Wechseln ins Tika-Verzeichnis und führen sie folgendes Kommando aus:

```
mvn install
```

Beuth-Dokumente und einfache Logfile-Analyse

- Sie finden auf der Vorlesungsseite eine Datei `beuth-dokumente.tar.gz`. Der Inhalt wurde mit folgendem Programm und Parametern erzeugt:

```
wget -r -S -A.html -l2 -o messages.log
      http://beuth-hochschule.de/
```

- Schauen sie sich die *man page* von `wget` an und erklären Sie die Parameter
- Logfileanalyse
  - Im Hautpverzeichnis `beuth-dokumente` finden Sie das Logfile `messages.log`:
  - Zählen Sie mit Betriebssystemmittel auf der Kommandozeile die Anzahl der heruntergeladenen `index.html`- und `print.html`- Dateien (Schreiben Sie hierzu kein Programm oder Skript!)

Apache Tika

- Starten Sie die GUI von Tika über

  ```
  java -jar lib/tika-0.3-standalone.jar --gui
  ```

- Starten Sie die Kommandozeilenversion von Tika über

  ```
  cat .../index.html | java -jar
              tika-0.3-standalone.jar -
  ```

- probieren Sie Variationen; siehe `--help`
- Schauen Sie sich die Dokumentation von Tika an
- Schauen Sie sich die Benutzung des `org.xml.sax.ContentHandler` an

Indizierung der Dokumente

- Schreiben Sie ein Programm, das den Text aus den HTML-Seiten indiziert.
- Implementieren Sie auch eine entsprechende Suche auf dem Text.

Anforderungen

- Aufbauend auf der Übung soll ein Programm für eine Suche auf den Webseiten der Beuth-Hochschule entstehen. Dabei soll später eine Dokumentenverarbeitung zwischen Tika und der Indizierung über Apache UIMA erfolgen.
- Daher ist es wichtig, dass sie sich ein gutes Software-Design unter folgenen Gesichtspunkten überlagen:
  - Modularität
  - erweiterbar und flexibel
  - *separation of concern*
  - Es ist anzunehmen, dass sich die Schnittstellen von Tika ändern werden.
- Schreiben Sie sauberen und dokumentierten Code;

Arbeitsergebnisse und Abgabe

- später