Web Mining: Übung 3

Lösungsvorschlag

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 1:
Einteilung der Testdaten

TECHNISCHE UNIVERSITÄT DARMSTADT

- Verwendeter Testdatensatz: Co-training Experiments for COLT 98
  `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-51/www/co-training/data/`
- Testdaten sind in Unterordnern sortiert:
  Art/Kategorie/VolleURL
- Art kann „fulltext“ oder „inlinks“
- „fulltext“: HTML-Dateien
- „inlinks“: Link-Text(e), die auf die Seiten verwiesen haben.
  Beispiel: `<a href=VolleURL>Link-Text</a>`
- Zur einfacheren Verarbeitung soll die Ordnerhierarchie entfallen und die Informationen in den Dateinamen übernommen werden.

Aufgabe 1:
BASH: Neusortierung/Entfernung der Hierarchie
Teil 1: Die dunklen Zeichen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Mit einem Bash-Skript soll die Hierarchie flach gezogen werden
- Das Namensschema der Dateien soll lauten:
  fulltext|inlinks $\underbrace{---\text{-}--}_{\text{Trenner}}$ Kategorie $\underbrace{---\text{-}--}_{\text{Trenner}}$ VolleURL
- Die URL soll dabei außerdem angepasst werden:
  - Doppelpunkte werden entfernt, weil sich Windows daran stört
  - `http://www.` wird ersetzt, um die URLs zu verkürzen. Windows stört sich auch an langen Dateipfaden
- Am Ende sehen die Dateinamen so aus:
  $\underbrace{\textit{fulltext}}_{\textit{Art}}--\underbrace{\textit{course}}_{\textit{Kategorie}}--\underbrace{\textit{BEG}}_{\textit{http...}}--$ cs.washington.eduˆeducationˆcoursesˆ431ˆ
- Vorteil des Ersetzens gegenüber dem Wegwerfen einzelner Zeichen/Abschnitte: Es gehen keine Informationen verloren. Bei Bedarf kann der gesamte Link wiederhergestellt werden.

Aufgabe 1:
BASH: Neusortierung/Entfernung der Hierarchie
Teil 2: Angriff der Kategorisierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Die Daten müssen nun noch in Test- und Trainingsdaten eingeteilt werden
- Das Skript wirft für jede Datei die Münze und dokumentiert das Ergebnis ebenfalls im Dateinamen
- Beispiel:
  `test--fulltext--course--BEG--cs.wisc.edu^~dyer^cs766.html`
- Das Ergebnis der Verteilung scheint zufällig und damit brauchbar:

	Training	Test	Summe
course	233	227	460
non-course	802	840	1642
Summe	1035	1067	2102