Internet and Web Applications, Übung 4

Henry Haustein

Aufgabe 1: Web Crawling

(a) `/WORLD` → `http://www.cnn.com/WORLD`
`http://www.cnn.com:80` → `http://www.cnn.com`
`http://www.cnn.com/1.html#3` → `http://www.cnn.com/1.html`

(b) BFS: zuerst Breite, dann Tiefe
DFS: zuerst Tiefe, dann Breite

Aufgabe 2: TFIDF

(a) Anzahl aller Wörter N_j im Dokument j

- T1: 9 Wörter
- T2: 9 Wörter
- T3: 12 Wörter
- T4: 3 Wörter

Wie oft taucht der Begriff t (*computer* oder *science*) im Dokument j auf?

- T1: $n_{t,1} = 2 \Rightarrow TF_1 = \frac{2}{9}$
- T2: $n_{t,2} = 3 \Rightarrow TF_2 = \frac{3}{9}$
- T3: $n_{t,3} = 3 \Rightarrow TF_3 = \frac{3}{12}$
- T4: $n_{t,4} = 1 \Rightarrow TF_4 = \frac{1}{3}$

$D = 4$ und $D_t = 4$, weil alle Dokumente den Term enthalten $\Rightarrow IDF_t = 1$ und damit

- T1: $TFIDF_t = \frac{2}{9}$
- T2: $TFIDF_t = \frac{3}{9}$
- T3: $TFIDF_t = \frac{3}{12}$
- T4: $TFIDF_t = \frac{1}{3}$

$\Rightarrow$ T2 ist das beste Dokument.

(b) Aufblähen des Scores mit sinnlosen Begriffen möglich, ohne Mehrwert zu bieten

Aufgabe 3: PageRank

PageRank is a way to prioritize results of Web keyword search based on evaluation of the link structure. Invented by Google founders Brin and Page and applied as part of Google's ranking algorithm. Simplified assumption: A hyperlink from page A to page B is a recommendation of the content of page B by the author of A $\Rightarrow$ Quality of a page is related to its in-degree (number of incoming links). Recursion: Quality of a page is related to

- its in-degree and to
- the quality of pages linking to it