Technologien für das Internet I

Übung vom 18.11.2010
Bearbeiten bis: 25.11.2010
Bei Fragen: demidova@l3s.de

TF/IDF Gewichtung und Retrieval im Vektorraum

Gegeben sei eine Dokumentsammlung mit insgesamt zwei Dokumenten:

Dokument 1:
Wenn Fliegen hinter Fliegen fliegen, fliegen Fliegen Fliegen nach!
Dokument 2:
Die Fliegen (Brachycera)bilden eine Unterordnung der Zweiflügler.

a. Erzeugen Sie aus den Dokumenten einen invertierten Index.
Tokenisierung-Regeln: Wort-weise (Satzzeichen ignorieren), alles klein geschrieben.
Geben Sie dabei jeweils an passender Stelle für jeden Term die entsprechenden Werte für TF und DF an.

b. Die Funktion für eine Suchanfrage-Dokument-Relevanz sei

$$w_{Q,d} = \sum_{q \in Q,d} \left(1 + \log\left(TF_{q,d}\right)\right) * log \frac{N}{DF_q}$$

Berechnen Sie die jeweilige Relevanz für die beiden Dokumente im Bezug auf die beiden folgenden Suchanfragen:

Q_1= *billig fliegen*
Q_2= *Zweiflügler*
Erläutern Sie die Ergebnisse!