Übung: Scoring mit dem Vektorspace-Modell

May 5, 2009

Prinzip des Relevance Ranking mittels Vektorspace-Modell

- Gegeben sind folgende Dokumente
  - Doc1: Durchbruch in der Behandlung von Krebs
  - Doc2: Durchbruch in der Therapie des Adenokarzinom
  - Doc3: Behandlung von Krebs Behandlung
  - Doc4: Behandlung von Krebs
  - Doc5 (2 mal der Inhalt von Doc1): Durchbruch in der Behandlung von Krebs Durchbruch in der Behandlung von Krebs

Prinzip des Relevance Ranking mittels Vektorspace-Modell

- Gegeben sind folgende Anfragen (*Queries*):
  - Q1: Krebs
  - Q2: Krebs ODER Adenokarzinom
  - Q3: Behandlung

Prinzip des Relevance Ranking mittels Vektorspace-Modell

- Berechen Sie den Score jedes Dokumentes zu den einzelnen Anfragen mittels: $score_{d,q} = (\vec{d} \cdot \vec{q})/(|\vec{d}| * |\vec{q}|)$
- Gehen Sie von folgenden Annahmen aus:
  - Token = Terme ; d.h. der Analyser ist ein reiner *Whitespace-Tokenizer*
  - Inverse Document Frequency $idf_{t_i} = \ln N/n_t$ mit N ist die Anzahl der Dokumente; n_t ist die Anzahl der Dokumente in denen der Term t vorkommt
  - Element i des Query-Vektors: $q_i = w_{t_i,q} * idf(t_i) = tf_{t_i,q} * idf_{t_i}$
  - Element i des Dokument-Vektors: $d_i = w_{t_i,q} * idf(t_i) = (1 + \ln tf_{t_1,d}) * idf_{t_i}$
  - Die Längen-Norm des Dokumentes ist die Wurzel der Anzahl der Token des Dokuments $|\vec{d}| = \sqrt{n_d}$
  - Längen-Norm der Query $|\vec{q}| = 1$ (neutrales Element der Multiplikation, da keinen Einfluss auf Ranking)
  - ln: natürlicher Logarithmus

Prinzip des Relevance Ranking mittels Vektorspace-Modell

- Ranken Sie die Dokumente für die Anfragen mittels des Score.
- Welche Faktoren bestimmen das Ranking für die einzelnen Anfragen?