Aufgabenblatt 5: Thesauren, Vektorraummodell

Aufgabe 1 Thesaurus Beispiele

Suchen Sie die folgenden Beispiele in den angegebenen Thesauren. Beschreiben Sie die Probleme, auf die Sie dabei gestoßen sind, sowie Vor- und Nachteile der jeweiligen Thesauren. Wie könnten die Thesauren in einer Literaturdatenbank genutzt werden? Wie in einem anderen Informationssystem?

a) Suchen sie im "Getty Thesaurus of Geographic Names"
http://shiva.pub.getty.edu/cgi-bin/tgn_browser/tgn.spl?key=1000000&searchtype=hier&form_file=/tgn_browser/index.html&lang=vern
in der hierarchischen Darstellung die Einträge für Darmstadt, Frankfurt (Main), Frankfurt (Oder) und Andlau.

b) Suchen Sie die Wörter aus dem Beispiel 1 (vertebrates, invertebrates, animal, cat, dog, pidgeon, monkey, rat, mouse, rabbit) in der online Version des Rodget's Thesaurus (http://www.thesaurus.com).

c) Versuchen Sie diese Wörter oder entsprechende Deskriptoren auch im OECD Macrothesaurus (http://info.uibk.ac.at/info/oecd-macroth/) zu finden.

Aufgabe 2 Zipf'sches Gesetz

a) Erläutern Sie das "Zipf'sche Gesetz".

b) Welche Folgerungen können daraus für das Information Retrieval gezogen werden?

Aufgabe 3 Das Vektorraummodell mit einer Inverted Document Frequency Gewichtung

a) Wie unterscheiden sich globale und lokale Einflussfaktoren bei der Gewichtung von Termen?

b) Beschreiben Sie einige globale und lokale Einflussfaktoren.

c) Berechnen Sie für die ersten vier Titel aus Aufgabe 1 Blatt 3 die gewichtete Indexierung mit der globalen Gewichtsformel

$$\frac{m}{d(i)}$$

wobei m die Anzahl der Dokumente in der Sammlung (in diesem Fall also die Anzahl der Titel) bezeichne und $d(i)$ die Anzahl der Dokumente, in denen der Term t_i vorkommt. Ignorieren Sie bei der Bestimmung der Terme Gross- und Kleinschreibung und behandeln Sie alle Nicht-Buchstaben als Worttrenner. (Eine Datei mit den Titeln in der vorgesehenen Form finden Sie unter http://www.darmstadt.gmd.de/~ferber/ubung/docs.txt)

d) Welche lokalen Einflussfaktoren könnten in diesem Fall zur Gewichtung verwendet werden?

e) Berechnen Sie mit dem Skalarprodukt die Ähnlichkeit der vier Titel zur Anfrage

```
indexing structure for a speech database
```

Aufgabe 4 Relevance Feedback

Gegeben seien die folgenden Dokumente mit gewichteten Termen:

```
{ROT:0.4, GRÜN:0.6, BLAU:0.3, GELB:0.5, SCHWARZ:0.7}
{ORANGE:0.6, LILA:0.8, ROSA:0.7, BEIGE:0.6, BLAU:0.4, GELB:0.2}
{BLAU:0.2, PINK:0.6, GRAU:0.3, BEIGE:0.5, SCHWARZ:0.4}
{ROT:0.4, LILA:0.2, WEISS:0.7, SCHWARZ:0.4}
{GELB:0.4, LILA:0.5, GRAU:0.2, ORANGE:0.7, SCHWARZ:0.6}
{BLAU:0.6, GRÜN:0.4}
```

In dieser Menge werde mit der Anfrage

```
Gelb:1.0, Lila:0.2
```

nach dem Vektorraummodell gesucht. Dabei wird als Ähnlichkeitsmaß das Skalarprodukt verwendet. Relevant seien die Dokumente 2, 3 und 4.

Berechnen sie zwei Relevance Feedback Durchgänge mit der Methode von Rocchio mit Parametern $\alpha = 0.7$, $\beta = 0.2$ und $\gamma = 0.1$