Aufgabenblatt 4: Abgabe Donnerstag 21. 06. 12.00 Uhr.

Aufgabe 1 Termgewichtung:

Diese Grundformel zur Termgewichtung liegt zahlreichen Verfahren der automatischen Indexieurng zu Grunde:

$$\frac{\text{Häufigkeit, mit der t im Dokument vorkommt}}{\text{Anzahl der Dokumente in Kollektion, in denen t vorkommt}}$$

Mit Hilfe dieser Formel können Sie das Gewicht einzelner Indexterme für die Dokumente in einer Datenbank ermitteln.

Stellen Sie sich folgende Situation vor:
Wir betrachten drei Dokumente in einer medizinischen Datenbank. Die Datenbank enthält insgesamt 10.000 Dokumente. Um die Aufgabe einfach zu halten, bestehen unsere Dokumente jeweils nur aus dem Titel.

Ti1 = Anwendung von Tenormin bei Krankeiten des Herzens
Ti2 = Zusammenhang zwischen Beschwerden der Lunge und Beschwerden des Herzens
Ti3 = Tenormin bei Lungenkrankheiten

Folgende Verteilung der Indexterme liegt in unserer Datenbank vor (Grundform, auf Kompositazerlegung verzichten wir).

Termverteilung in medizinischer Datenbank

Term t	Anzahl der Dokumente die t enthalten
Anwendung	3000
Beschwerde	2000
Herz	500
Krankheit	4000
Lunge	600
Lungenkrankheit	200
Tenormin	40
Zusammenhang	1500

Aufgaben:

1. Errechnen Sie nach der obenstehenden Formel die Termgewichte der Indexterme in den drei Dokumenten Ti1-Ti3, für die vorliegende Dokumentenkollektion. Indexiert werden alle Substantive.
2. Stellen Sie sich jetzt vor, Sie erhalten folgende Anfrage: "Tenormin für

Beschwerden des Herzens". Errechnen Sie die Relevanz der einzelnen Dokumente fu□r die Anfrage und erstellen Sie ein Ranking der Dokumente nach Relevanz.
3. Welches Dokument würde Ihnen auf die Anfrage "Tenormin für Beschwerden des Herzens" angezeigt, wenn es in der Suchmaschine kein Verfahren zur Termgewichtung gäbe und die Recherche nach der Methode des "exact match" durchgeführt würde?
Quelle: Knorz 1994,http://www.bui.haw-hamburg.de/pers/ulrike.spree/medo3ueb2.html

Aufgabe 2:

(Übung 6.10 „Introduction to Information Retrieval")
Benutzen Sie folgende Tabelle von Termfrequenzen für 3 Dokumente bezeichnet als Doc1, Doc2, Doc3 . Berechnen Sie die tf-idf Gewichte für die Terme car, auto, insurance, best, für jedes Dokument. Benutzen Sie die idf Werte aus Tabelle 6.8.

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

▶ **Figure 6.9** Table of tf values for Exercise 6.10.

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

▶ Figure 6.8 Example of idf values. Here we give the idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

Aufgabe 3 Termgewichtung
Die Wörter w1,w2,w3 sollen in einer Kollektion mit 10 Dokumenten d1.....d10 die nachfolgend genannten Vorkommenshäufigkeiten haben. Nehmen Sie einfachheitshalber an, dass alle anderen Wörter in den Dokumenten Stoppwörter sind. Berechnen Sie jeweils die sich ergebenden tf_idf Werte der Wörter in den Dokumenten.

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
w1	3	0	0	5	12	0	0	2	8	1
w2	8	6	0	12	0	0	9	1	3	10
w3	0	1	7	0	1	5	12	0	2	0