

## Schematische Darstellung eines invertierten Dateisystems und Übung

In einer Datenbank sind die folgenden beiden Dokumente enthalten:

Dokument 1

Er konnte nie über etwas lachen,  
wie kann ein Mensch so tief verflachen !

Dokument 2

Wozu, so fragt man sich, Reichtum, Wohlstand, Macht,  
wenn alles dies die Menschen nur verflacht?

(aus: Christian Morgenstern, Sprüche)

### Arbeitsschritte der Volltextinvertierung:

Erster Schritt: Erstellung eines Index in der Reihenfolge des Vorkommens im Text.  
Wort erhält beim ersten Vorkommen eine feste Adresse (Nr.)

Zweiter Schritt: alphabetische Sortierung

numerisch sortierter Index		alphabetisch sortierter Index	
Adresse	Wort	Wort	Adresse
1	er	alles	22
2	konnte	die	24
3	nie	dies	23
4	über	ein	9
5	etwas	er	1
6	lachen	etwas	5
7	wie	fragt	15
8	kann	kann	8
9	ein	konnte	2
10	Mensch	lachen	6
11	so	Macht	20
12	tief	man	16
13	verflachen	Mensch	10
14	wozu	Menschen	25
15	fragt	nie	3
16	man	nur	26
17	sich	Reichtum	18
18	Reichtum	sich	17
19	Wohlstand	so	11
20	Macht	tief	12
21	wenn	über	4
22	alles	verflachen	13

23 dies	verflacht	27
24 die	wenn	21
25 Menschen	wie	7
26 nur	Wohlstand	19
27 verflacht	wozu	14

Dritter Schritt: Jetzt wird das invertierte Dateisystem erstellt, in dem für jedes Wort eine invertierte Liste erstellt wird, in der angegeben wird:

- Identifikationsnr. des Dokumentes
- Position im Text
- Häufigkeit
- Zusatzinformationen, wie Gewicht des Wortes im Text

### Erstellung des Invertierten Dateisystems

Nr. Wort	DocID	Position	Frequenz	Zusatzinfo
10	1	10	1	
11	1	11	1	
	2	2	1	
....	....	...	....	....
27	2	15	1	

Ein Zeile ist eine **invertierte Liste/Datei** im **invertierten Dateisystem**. Für jedes Wort wird eine eigenständige invertierte Liste in der invertierten Datei erstellt.

Achtung: die Begriffe invertierte Liste und invertiert Datei werden häufig verwechselt, bzw. synonym verwendet. Es gibt keine einheitliche Sprachregelung geben.

(Meine Quellen:

Einführung in Information Retrieval

[<http://pi3.informatik.uni-mannheim.de/lehre/vorlesungen/introIRSkript/IR162-209.pdf>]

C.J. van Rijsbergen : Information Retrieval, Kapitel 4

[<http://www.dcs.gla.ac.uk/~iain/keith/data/pages/72.htm>] )

**Übung:**

Die Datenbank wird durch folgendes Dokument ergänzt.

Dokument 3

Ein Mensch sitzt stumm und liebeskrank  
Mit einem Weib auf einer Bank;

Er nimmt die bittere Wahrheit hin,  
Daß sie zwar liebe, doch nicht ihn.

Ein andrer Mensch geht still vorbei  
Und denkt, wie glücklich sind die zwei,

(aus: Eugen Roth : Optische Täuschung)

1. Ergänzen Sie das invertierte Dateisystem um die invertierte Liste für Mensch.
2. Stellen Sie schematisch dar, wie im System nach Mensch gesucht wird.