ifis
Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines
Summer Semester 2011

Prof. Dr. Wolf-Tilo Balke and Joachim Selke

Homework Assignment 3

Exercise 3.1

What does the Probabilistic Ranking Principle state? Use your own words.

Exercise 3.2

Both Probabilistic Indexing (PI) and Binary Independence Retrieval (BIR) try to estimate $\Pr(d \in R_q)$, where q is a document, q is a query, and R_q is the set of all documents being relevant with respect to q. Also, both PI and BIR rely on Bayes' Theorem to interpret this probability. What are the two different approaches used here? Do not use any formulas!

Exercise 3.3

Both PI and BIR make an independency assumption. Explain each assumption in your own words (do not use any formulas!) and discuss whether it is reasonable or not.

Exercise 3.4

What do Zipf's Law and Heaps' Law state? Use your own words. What are the consequences of these laws for information retrieval systems?

Exercise 3.5

What are the advantages and disadvantages of stemming?

Exercise 3.6

Why do we need inverted indexes in information retrieval?

Exercise 3.7

As hard disks are extremely cheap these times, it doesn't seem to be a good idea to waste CPU resources by compressing and decompressing inverted indexes. Why are compressed indexes a good idea anyhow?

Exercise 3.8

LSI can be used to recognize synonyms, antonyms, and semantically related terms. Briefly explain how this works. Use your own words, no not use any formulas. In particular, explain what a latent space is and what its properties are.

Exercise 3.9

LSI also has interesting applications in automated document translation. Can you imagine why and how LSI can help here?