

## Homework Assignment I

### Exercise 1.1

The Association for Computing Machinery (ACM) maintains an own classification system for categorizing scientific documents from the area of computer science. This system is called the ACM Computing Classification System and was last revised in 1998. How would a book about Web search engines be classified according to this system?

### Exercise 1.2

What are the possible benefits of using sophisticated classification schemes such as MeSH? What are possible problems, and in what scenarios are such approaches worth the effort?

### Exercise 1.3

Text documents can easily be stored in a relational database, e.g., using the CLOB data type. Why shouldn't we just use established database technology for information retrieval tasks?

### Exercise 1.4

What is the meaning of the terms *information need* and *relevance* in information retrieval? What is the connection between both?

### Exercise 1.5

What is the difference between the bag of words model and the set of words model?

### Exercise 1.6

In what kind of scenarios can Boolean retrieval be useful? Why?

### Exercise 1.7

In practice, term–document matrices tend to be extremely sparse, i.e., most of its entries are zero (around 99.9%). Can you think of any better ways to store these matrices than using a two-dimensional array with #documents rows and #terms columns?