ifis
Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines
Summer Semester 2010

Prof. Dr. Wolf-Tilo Balke and Joachim Selke

Homework Assignment 1

Due to April 22, 2010
(36 points in total)

Please note: To be admitted to the final exam, you need **50% of all homework points.** Your homework has to be turned in until the **due date** indicated above, **before the lecture** starts. Please work together in **groups of two** students, larger groups are not permitted. Write your **names** and **matriculation numbers** on each page of your homework. If you have multiple pages, **staple** them together! Please drop your solutions into the **mailbox** at our institute (Informatikzentrum, second floor, next to room 238) or hand them over to us right before the lecture begins.

Exercise 1.1 (Homework Management System)

Please log in to our Homework Management System (HMS) at

`https://www.ifis.cs.tu-bs.de:8181/hms`

using your y-number and password, and **sign in for this lecture.** This will make grading and managing your homework submissions a lot easier, both for you and us! (2 points)

Exercise 1.2 (Classification Schemes)

The Association for Computing Machinery (ACM) maintains an own classification system for categorizing scientific documents from the srea 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 (such as Google) be classified according to this system? Briefly explain your answer. (2 points)

Exercise 1.3 (Installing MATLAB)

Install MATLAB R2010a on your computer. In order to do this, please proceed as follows:

- Create a user account at the Mathworks website (`http://www.mathworks.com`) using your TU email address (yourname@tu-bs.de). Other email addresses won't work!
- Login at the Mathworks website and open the *My Account* page (`http://www.mathworks.com/accesslogin/myAccount.do`). Download MATLAB R2010a for your operating system.
- Install MATLAB R2010a on your computer. The necessary license file (matlab.lic) can be found at the page `http://gitz-dl.tu-bs.de/software/matlab.pl` (only accessible from within

the TU's network or via VPN). Please ignore the rest of this page since it is mostly outdated. A detailed installation guide (which is particularly interesting for Linux and Mac users) can be found at the page `http://www.mathworks.com/access/helpdesk/help/base/install`.

If you have any problems which cannot be solved with this guide, feel free to ask Joachim Selke, who will provide assistance.

If you are finished, please run the command `ver` in MATLAB and print out the result, and hand it in with your homework. (5 points)

Exercise 1.4 (Getting to Know MATLAB)

Make yourself familiar with MATLAB by going through one or more of the MATLAB tutorials listed on the page `http://www.mathworks.com/academia/student_center/tutorials/launchpad.html`. Please work on this exercise very carefully. If you don't know how to use MATLAB, you won't be able to solve many of the following homework assignments.

Please answer the following questions regarding MATLAB:

a) How do you define a variable `A` that stores the matrix

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}?$$

(1 point)

b) How can you determine the size of a matrix (that is, its number of rows and columns)? (1 point)

c) If `A` is matrix, how can you find out what number stands in row 3 of column 2? (1 point)

d) How do you transpose a matrix in MATLAB? (1 point)

e) What is a sparse matrix and how does it differ from full matrices? In which situations it is advantageous to use sparse matrices instead of full ones? (2 points)

f) How do you compute the scalar product (also called dot product) between two vectors? (1 point)

g) Assuming that `A` is a $m \times n$ matrix, how do you compute a vector `v` such that the i-th entry of `v` contains the mean of `A`'s i-th row? (2 points)

h) If `A` is matrix, how can you add the number 3 to each entry of `A`? (1 point)

i) How do for loops work in MATLAB? Give a short example. (1 point)

j) How do you create 10 random 2-dimensional points (uniformly distributed in $[0, 1]^2$) and draw a scatterplot of them? (2 points)

k) How do you output text on the MATLAB console? Write a small *Hello world* program. (1 point)

l) What is the difference between a matrix and a cell array? How can the elements of a cell array be accessed? (2 points)

Exercise 1.5 (Boolean Retrieval with MATLAB I)

Download the MATLAB data file reuters-21578.mat from `http://www.ifis.cs.tu-bs.de/webfm_send/382`, which contains the Reuters-21578 data set. Reuters-21578 is a collection of 21578 documents which appeared on the Reuters newswire in 1987.[1] The file reuters-21578.mat contains four MATLAB variables: `TD` is the 43880 $\times$ 21578 term–documentation matrix of the collection, that is, its entry (i, j) indicates how many times the i-th term occurs in document j. `docs` is a 1 $\times$ 21578 cell array, whose j-th entry contains the complete text of document j. `terms` is a 43880 $\times$ 1 cell array, whose i-th entry contains the i-th index term of the collection. Finally, `terms_map` is a so-called containers.Map object that maps terms to their respective index numbers (for example, you can get the index number of the term *cocoa* by typing `terms_map('cocoa')`).

Please import the data from reuters-21578.mat into your MATLAB workspace and find out which documents are returned when asking the following queries in the Boolean retrieval model (you may find the operators `and`, `or`, and `not` useful):

a) *cocoa* AND *gold* (2 points)

b) *microsoft* OR (*apple* AND *computer*) (2 points)

c) *rich* BUT NOT *famous* (2 points)

Exercise 1.6 (Boolean Retrieval with MATLAB II)

In this exercice, we again use the Reuters-21578 data set. For each of the following result sets, write a (short) Boolean query which returns *exactly* the given result set.

a) $\{\text{doc6789}\}$ (2 points)

b) $\{\text{doc666}, \text{doc6543}\}$ (2 points)

c) $\{\}$ (1 point)

[1] Please see `http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt` for additional details.