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 2

Due to May 6, 2010
(36 points in total)

Exercise 2.1 (Index compression)

Let's assume that we want to derive posting lists from the term–document matrix of the Reuters collection and store them on disk.

a) How many bits of storage space do we need if we store document IDs as 16-bit integers and term frequencies as 8-bit integers? (2 points)

b) How many bits of storage space do we need at least if we store document IDs and term frequencies in unary encoding? (Please keep in mind that to minimize storage space you must assign new IDs to the documents based on to their frequency; the most frequent document gets ID 0, the second-most frequent gets ID 1, and so on.) (2 points)

c) How many bits of storage space do we need at least if we store document IDs and term frequencies in gamma encoding? (2 points)

Exercise 2.2 (Tokenization on the Reuters data set)

Take another look at the Reuters data set. To create the term–document matrix (and the corresponding list of terms), we already tokenized the document collection. Please describe how the typical problems of tokenization (as discussed in the lecture) have been handled by our tokenizer. What could have been improved here? (5 points)

Exercise 2.3 (Filtration and stemming)

Perform filtration and stemming on the Reuters data set! For filtration, please use the list of stopwords given on `http://www.textfixer.com/resources/common-english-words.txt`. For stemming, please use the implementation of the Porter stemmer provided by the MATLAB toolbox *TMG*, which can be downloaded from `http://scgroup20.ceid.upatras.gr:8000/tmg`. Make sure that you end up list of terms (and a corresponding term–document matrix) that does not contain any duplicate terms.

How is the number of terms affected by each step? Does it make a difference if you perform filtration first and then stemming, or if you do it the other way around? (15 points)

Exercise 2.4 (Vector space retrieval)

Take the term–document matrix you created in the previous exercise and replace the pure term frequencies by TF–IDF weights. Answer the query *taxes AND usa* both with the Boolean retrieval model and the vector space model, and compare the results. (Keep in mind that you need to apply stemming and filtration to the query first.) (5 points)

Exercise 2.5 (Term frequencies in the Reuters data set)

In Lecture 4 we will claim that in every document collection the term frequencies are distributed according to Zipf's law. Check it yourself on the Reuters data set! (You may want to use a log-log plot to get a clear picture.) Does Zipf's law hold after stemming and filtration have been performed? (5 points)