Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Björn Buchhold

Information Retrieval
WS 2013/2014

http://ad-wiki.informatik.uni-freiburg.de/teaching

Exercise Sheet 4

Submit until Tuesday, November 19 at 4:00pm

Exercise 1 (5 points)

Assume that an infinite inverted list (with document ids: $0, 1, 2, \ldots$) is generated as follows. Each document id is included in the list with a given probability p, independent of the inclusion of all other document ids. Let G be the (random) gap between two document ids in this list. Determine the probability distribution for G, that is, the value of $\Pr(G = i)$, for all $i \in \mathbb{N}$.

Exercise 2 (5 points)

Consider a list generated as described in Exercise 1, for a given probability p. Prove that Golomb-encoding is entropy-optimal for gap-encoding this list, and for which value of M. Hint: Try $M = c \cdot 1/p$, for a suitable c, and use that $1 - p \leq e^{-p}$.

Exercise 3 (10 points)

Implement Golomb-encoding (compression and decompression) and compare it to Variable-Byte (VB) encoding. Compare the two on the inverted list for *american* and the one for *freiburg*. Use your result from Exercise 2 for the value of M. Take $p = m/N$, where m is the size of the inverted list and N is the total number of documents. Put your results in the result table on the Wiki, following the instructions given there. Briefly discuss your results in your *experiences.txt* for this exercise sheet.

You find a full implementation of VB-encoding (together with the parsing code from Exercise Sheet 1) on the Wiki, in both C++ and Java. Note that you can re-use much of this code for your implementation of Golomb-encoding. Do not forget the unit tests (one for compression and one for decompression), they will be invaluable for debugging.

Add your code to a new sub-directory *exercise-sheet-04* of your folder in the course SVN, and commit it. Make sure that *compile*, *test*, and *checkstyle* run through without errors on Jenkins. As usual, also commit a text file *experiences.txt* where you briefly describe your experiences with this exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where. Don't forget to include the short discussion asked for in Exercise 3.