Data Compression Techniques (Spring 2012)

Course Exam, 01 March 2012 at 16-19 Lecturer: Juha Kärkkäinen

Please write on each sheet: your name, student number or identity number, signature, course name, exam date and sheet number. You can answer in English, Finnish or Swedish.

1. [*2+2+2+2+2+2 points*] What is the *main difference* between the concepts in the following pairs:

   (a) prefix codes versus arithmetic coding

   (b) gamma code versus delta code

   (c) entropy versus zeroth order empirical entropy

   Define the following concepts:

   (d) zero frequency problem

   (e) coarse optimality

   (f) searchable prefix sums

   A few lines for each part is sufficient.

2. [*6+6 points*] Compress the text $T =$ `senselessness` using zeroth order semiadaptive Huffman coding.

   (a) Construct the Huffman code using the symbol frequencies in the text. Encode T using the code.

   (b) The encoded text must be stored with additional information that is needed for decoding the text. Describe what information must be stored, and how to store it using as few bits as possible.

   You may assume that the alphabet is fixed to $\{\texttt{e}, \texttt{l}, \texttt{n}, \texttt{s}\}$ and the alphabetical order is $\texttt{e} < \texttt{l} < \texttt{n} < \texttt{s}$.

3. [*6+6 points*] Let $L =$ `rttrraa$ii` be the Burrows–Wheeler transform for a text T. The order of the symbols is $\texttt{\$} < \texttt{a} < \texttt{i} < \texttt{r} < \texttt{t}$. The last character of T is `$`.

   (a) What is T? Explain step-by-step the inverse Burrows–Wheeler transform, i.e., how T can be recovered from L.

   (b) Give the wavelet tree of L. Explain how the wavelet tree can be used in the inverse Burrows–Wheeler transform.

4. [*7+7 points*] Compare LZ77 (without distance or length limits) and LZ78 parsings.

   (a) Give an example of a text, where the number of phrases in the LZ78 parsing is much larger than in the LZ77 parsing.

   (b) Prove that the number of phrases in the LZ78 parsing is never smaller than in the LZ77 parsing.