- This problem set has *three* questions. The programming problems may be harder, or require more time, than their point value suggests.
- Please type your solutions to the written component and send a pdf file to **REDACTED**. The pdf filename should be "EDS-A4-<your_user_name>.pdf"
- The deadline is **24.05.2014** anywhere on Earth.

(40) 1. **Wavelet Tree:** In class we showed how wavelet trees can support the three operations, `Rank`, `Select`, and `Access` in $\Theta(\log \sigma)$ time. Suppose that we wish, for a particular index i, to report the number of occurrences of character α up to position i, for *all* $\alpha \in [0, \sigma - 1]$. The simple approach is to just call $\texttt{Rank}(i, \alpha)$ for each $\alpha \in [0, \sigma - 1]$. This requires $\Theta(\sigma \log \sigma)$ time. Show how to do this in optimal $\Theta(\sigma)$ time.

(40) 2. **Higher Order Entropy:** In class we defined the *zeroth order empirical entropy* of a string S of length n as $H_0(S) = \sum_\alpha p_\alpha \log_2(1/p_\alpha)$ where p_α is the relative frequency of character α in S. For example, if $S = \texttt{0110110}$ then $p_0 = 3/7$ and $p_1 = 4/7$, so $H_0(S) \approx 0.9852$. *In this formula, if some* $p_\alpha = 0$*, then we take* $\log_2(1/p_\alpha) = 0$.

For any $k \geq 0$, there is an analogous concept called the *k-th order empirical entropy*. It is defined as follows: $H_k(S) = (1/n) \sum_{b \in [\sigma]^k} |S_b| H_0(S_b)$ where S_b is the concatenation of the of the characters that follow occurrences of substring b in S: examples will follow. Note that $[\sigma]^k$ is the set of all possible strings of length k drawn from the alphabet $[0, \sigma - 1]$. Furthermore, for technical reasons, we will consider the string to be cyclic for the purposes of the computation, so that the first character follows the last.

In what follows we restrict our attention to bit strings, so $\sigma = 2$. As an example of the definition, if $S = \texttt{0110110}$ and $b = \texttt{0}$, then we would concatenate the bits following `0`s in S: i.e., $\underline{\texttt{01}}\texttt{10}\underline{\texttt{1}}\texttt{10}$. Note that the first bit is underlined because it cyclically follows the last `0`. Therefore, $S_{\texttt{0}} = \texttt{011}$. If $b = \texttt{01}$, then we would mark $\texttt{01}\underline{\texttt{1}}\texttt{01}\underline{\texttt{1}}\texttt{0}$, so $S_{\texttt{01}} = \texttt{11}$.

For our example string if we set $k = 1$ we have:

$$H_1(S) = \frac{1}{7}\left(|S_0| H_0(S_0) + |S_1| H_0(S_1)\right) =$$

$$\frac{1}{7}\left(\left(|\texttt{011}| H_0(\texttt{011}) + |\texttt{1010}| H_0(\texttt{1010})\right) \approx 0.9650\right.$$

This is slightly smaller than the zeroth order entropy. In general, for each $k \geq 0$, we have $H_{k+1}(S) \leq H_k(S)$. For the example, we have $H_0(S) \approx 0.9852$, $H_1(S) \approx 0.9650$, $H_2(S) \approx 0.2857$, $H_3(S) \approx 0.2857$, etc.

Your task: For all integers $k \in [1, 4]$, construct a bit string S^k where $H_k(S^k) = 0$, but $H_0(S^k) = 1$ and $H_{k-1}(S^k) \geq 0.499$. Your answer need only consist of the bit strings. **Please ensure they are under 20 bits. I will specify how to submit these later in the week.**

(SPOJ:20) 3. `http://www.spoj.com/DS/problems/HENTROPY/`