Source coding

1 Huffman code

Definition 1 (weighted code tree) *The* weighted code tree *for a prefix code* $B : \mathbb{X} \to \{0,1\}^*$ *and a probability distribution* $p : \mathbb{X} \to [0,1]$ *is the code tree for code* B *where the nodes are enhanced with the following weights: (1) for a leaf node with symbol* a*, we add weight* $p(a)$*, (2) to other (internal) nodes, we ascribe weights equal to the sum of weights of their children.*

Definition 2 (Huffman code) *The* Huffman code *for a probability distribution* $p : \mathbb{X} \to [0,1]$ *is a code whose weighted code tree is constructed by the following algorithm:*

1. *Create a leaf node for each symbol and add them to a list.*
2. *While there is more than one node in the list:*
   (a) *Remove two nodes of the lowest weight from the list.*
   (b) *Create a new internal node with these two nodes as children and with weight equal to the sum of the two nodes' weights.*
   (c) *Add the new node to the list.*
3. *The remaining node is the root node and the tree is complete.*

2 Task

1. Download some texts, DNA sequences, or other discrete symbolic sequences (e.g., music in an appropriate format) in a sufficient amount (say, about 1MB) from the internet.
2. Compute the Huffman code for the empirical distribution of characters.
3. Compare the expected length of the Huffman code with the entropy.
4. Describe what you have obtained in a report, attach the used scripts, and send it to me (`ldebowsk@ipipan.waw.pl`).