Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Patrick Brosi

Information Retrieval
WS 2016/2017

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

UNI FREIBURG

Exercise Sheet 4

Submit until Tuesday, November 22 at **2:00pm**

Exercise 1 (6 points)

This exercise is about proving one direction of Shannon's famous source coding theorem.

In the lecture, we proved that for any prefix-free code it always holds that $\sum_x 2^{-L_x} \leq 1$, where L_x is the length of the encoding for integer x. Use this to show that for a random variable X drawn from $\{1, \ldots, m\}$, the expected code length $E(L_X)$ is always at least the entropy $H(X)$.

Hint: this is best done with Lagrangian optimization, as explained in Lecture 3.

Exercise 2 (8 points)

This exercise is about the optimal encoding for the gaps of the lists from an inverted index.

In the lecture we have shown that it is reasonable to assume that a fixed gap of an inverted list is distributed like a random variable X with X with $\Pr(X = x) = (1-p)^{x-1} \cdot p$ for some $p < 1$. Show that under this assumption, Golomb encoding with modulus $M = \lceil 1/p \cdot \ln 2 \rceil$ is an entropy-optimal encoding for the gaps, that is, $E(L_X) \leq H(X) + O(1)$.

Hint: you can use without proof that $e^x \geq 1 + x$ for all real numbers x.

Exercise 3 (6 points)

This exercise is about calculating the space usage of an optimally gap-encoded inverted index.

Assume we have a document collection with a total of N words from a vocabulary of m words. An inverted index for that collection hence has m inverted lists $L_1, \ldots, L_m$ with a total of N postings. We assume that the list lengths are Zipf-distributed, that is, l_j is proportional to $1/j$. By Exercise 2, this implies that the average code length for the gaps of L_j is $\Theta(\log j)$; you can use this without proof. Show that under these assumptions the expected total number of bits required to gap-encode all the inverted lists is $\Theta(N \cdot \log m)$.

Hint: you can use without proof that $\sum_{j=1}^{m} 1/j = \Theta(\log m)$ and $\sum_{j=1}^{m} (\log j)/j = \Theta(\log^2 m)$.

[turn over without increasing the entropy]

Commit your solutions in a single PDF in a new sub-directory *sheet-04* of your folder in the course SVN, and commit it. We recommend that you typeset your solution using LaTeX. With nice handwriting (if you are not sure if your handwriting is nice, it is not), you may also hand in a scan. In that case, take care that the scan has sufficient resolution and that the file is not too large (< 1 MB). Also commit the usual *experiences.txt*.

What is the entropy of the distribution of A, T, G, C in your DNA? And why, in DNA replication, is one half of the strand copied "forward" and the other one "backward"?