Seminar 3

Algorithm 1 (Variable byte code)
A number n is encoded in variable byte code in the following procedure:

1. *Take a binary representation of n with padding to the length of a multiple of 7.*
2. *Split into of 7 bit blocks right-to-left.*
3. *Add 1 to the beginning of the last block and 0 to the beginning of all previous blocks.*

Example: $VB(824) = 00000110\,10111000$

Definition 1 (α code)
Unary code, also referred to as α code, is a coding type where a number n is represented by a sequence of n 1s (or 0s) and terminated with one 0 (or 1). That is, 6 in unary code is 1111110 (or 0000001). The alternative representation in parentheses is equivalent but for this course we use the default representation.

Definition 2 (γ code)
γ code is a coding type, that consists of an offset and its length: $\gamma(n) = \alpha\big(\textit{length of offset}(n)\big), \textit{offset}(n)$. Offset is a binary representation of a number n without the highest bit (1). The length of this offset encoded in the unary (α) code. Then the number 60 is encoded in γ as 111110,11100.

Definition 3 (δ code)
A number n is encoded in δ code in the following way: $\delta(n) = \gamma\big(\textit{length of offset}(n)\big), \textit{offset}(n)$. Analogously, 600 is encoded in δ as 1110,001,001011000.

Definition 4 (Zipf's law)
Zipf's law says that the i-th most frequent term has the frequency $\frac{1}{i}$. In this exercise we use the dependence of the Zipf's law $cf_i \propto \frac{1}{i} = ci^k$ where cf_i is the number of terms t_i in a given collection with $k = -1$.

Definition 5 (Heaps' law)
Heaps' law expresses an empiric dependency of collection size (number of all words) T and vocabulary size (number of distinct words) M by $M = kT^b$ where $30 \leq k \leq 100$ and $b \approx \frac{1}{2}$.

Exercise 3/1

Count variable byte code for the postings list $\langle 777, 17\,743, 294\,068, 31\,251\,336\rangle$. Bear in mind that the gaps are encoded. Write in 8-bit blocks.

Exercise 3/2

Count γ and δ codes for the numbers 63 and 1023.

Exercise 3/3

Calculate the variable byte code, γ code and δ code of the postings list $P = [32, 160, 162]$. Note that gaps are encoded. Include intermediate results (offsets, lengths).

Exercise 3/4

Consider a posting list with the following list of gaps

$$G = [4, 6, 1, 2048, 64, 248, 2, 130].$$

Using variable byte encoding,

- What is the largest gap you can encode in 1 byte?
- What is the largest gap you can encode in 2 bytes?
- How many bytes will the above gaps list require under this encoding?

Exercise 3/5

From the following sequence of γ-encoded gaps, reconstruct first the gaps list and then the original postings list. Recall that the α code encodes a number n with n 1s followed by one 0.

1110001110101011111101101111011

Exercise 3/6

What does the Zipf's law say?

Exercise 3/7

What does the Heaps' law say?

Exercise 3/8

A collection of documents contains 4 words: *one*, *two*, *three*, *four* of decreasing word frequencies f_1, f_2, f_3 and f_4. The total number of tokens in the collection is 5000. Assume that the Zipf's law holds for this collection perfectly. What are the word frequencies?

Exercise 3/9

How many distinct terms are expected in a document of 1,000,000 tokens? Use the Heaps' law with parameters $k = 44$ and $b = 0.5$
