Teoria informacji i kodowania: Lista 12

Zadanie 1. Pokaż, jak policzyć parsing LZ78 w czasie liniowym (lub prawie liniowym).

Zadanie 2. Niech $c(n)$ oznacza liczbę fraz w parsingu P dla $x = x_1x_2\cdots x_n \in \{0,1\}^n$, przy czym każda fraza w P jest inna.

Pokaż, że

$$c(n) \leq \frac{n}{(1-\epsilon_n)\log n}$$

gdzie $\lim_{n\to\infty} \epsilon_n = 0$

Zadanie 3 (Trochę inny dowód uniwersalności LZ78)**.** Rozpatrzmy parsing generowany przez LZ78, jako ciąg symboli $f_1 \cdots f_c$, zauważmy, że są one parami różne (poza być może ostatnim, zignorujmy ten problem). Koszt zakodowania to z grubsza empiryczna entropia $H_e(f_1 \cdots f_c) = c\log(c)$. Dowód optymalności kodowania Huffmana działa też dla entropii empirycznej — kod Huffmana jest optymalny i daje najwyżej $c\log c + c$ bitów. Żeby ograniczyć $c\log c$ wystarczy więc podać pewne kodowanie, dla którego potrafimy policzyć rozmiar. By zakodować fragment, podajemy:

- długość frazy f
- kontekst
- numer frazy po uwzględnieniu powyższych.

Jaki jest rozmiar tego kodowania (zauważ, że do potrzeb kodowania nie musimy podawać „słownika", wystarczy, że powyższa funkcja jest jednoznacznie wyznaczona).

Wykorzystaj fragmenty dowodu z wykładu.

Zadanie 4. Pokaż, że problem stopu dla uniwersalnej maszyny bezprefiksowej jest nierozstrzygalny.

Zadanie 5 (Ograniczenie dolne na kompresję gramatykową)**.** W kompresji gramatykowej reprezentujemy słowo w jak gramatykę bezkontekstową, przy czym każda reguła jest postaci

$$X \to YZ$$

i Y, Z są albo nieterminalami albo literami. Rozmiarem gramatyki jest suma długości produkcji.

Używając złożoności Kołmogorowa udowodnij, że dla każdego n istnieje słowo binarne w długości n, którego minimalna gramatyka ma rozmiar $\Omega(n/\log n)$.

Zadanie 6. Pokaż, że ograniczenie z poprzedniego zadania jest ścisłe, tzn. każde słowo binarne ma gramatykę rozmiaru $\mathcal{O}(n/\log n)$.

Użyj podobnego podejścia jak w metodzie czterech Rosjan (podziel słowo na krótkie podsłowa, długości $\log n/c$. Ile jest takich słów? Ile może mieć w sumie gramatyka dla nich?

Zadanie 7 (Ograniczenie dolne na LZ77)**.** Możliwym uogólnieniem algorytmu LZ77 jest wersja bez ograniczenia na długość okna, taką zwykle rozpatruje się w algorytmach tekstowych.

Parsing LZ77 $f_1, \ldots, f_c$ słowa w jest generowany w następujący sposób: jeśli mamy już parsing $f_1, \ldots, f_m$ (początkowo $m = 0$) prefiksu w (początkowo w' jest puste), gdzie $w = f_1 \cdots f_m w'$ to następna fraza f_{m+1} to najdłuższa prefiks w' który jest podsłowem $f_1 \cdots f_m$ lub pojedyncza litera, jeśli taki najdłuższy prefiks ma długość 0.

Używając złożoności Kołmogorowa udowodnij, że dla każdego n istnieje słowo binarne w długości n, którego parsing LZ77 ma rozmiar $\Omega(n/\log n)$.