Kodowanie i kompresja danych 2020

Laboratorium nr 1 i 2 (na zaliczenie)

Zadanie na laboratorium

Dla dyskretnych zmiennych losowych X i Y entropia Y warunkowana przez X jest określona wzorem

$$H(Y|X) = \sum_{x \in X} P(x) \cdot H(Y|x)$$

gdzie

$$H(Y|x) = \sum_{y \in Y} P(y|x) \cdot I(y|x)$$

a $P(z)$ oznacza prawdopodobieństwo z a $I(z)$ informację związaną z z.

Napisz program który dla podanego pliku traktowanego jako ciąg 8-bitowych symboli policzy częstość występowania tych symboli oraz częstość występowania symboli po danym symbolu (częstość występowania pod warunkiem, że poprzedni znak jest dany, dla pierwszego znaku przyjmij, że przed nim jest znak o kodzie 0). Dopisz funkcje które dla policzonych częstości traktowanych jako zmienne losowe policzy entropię i entropię warunkową (warunkowaną znajomością poprzedniego symbolu), oraz poda różnicę między nimi.

Program ma wypisywać wyniki w sposób czytelny i łatwy do dalszego przetwarzania.

Przeanalizuj wyniki działania swojego programu dla przykładowych plików tekstowych, doc, pdf, mp4 czy jpg (weź pliki o rozmiarze co najmniej 1MB).

Zadania przygotowawcze do kolokwium

Zadanie 1

Sprawdź, czy następujące kody są jednoznacznie dekodowalne:

a) $\{0, 01, 11, 111\}$;

b) $\{0, 01, 110, 111\}$;

c) $\{0, 10, 110, 111\}$;

d) $\{1, 10, 110, 111\}$.

Czy któryś z nich jest prefiksowy?

Zadanie 2

Uogólnij nierówność Krafta (tj. twierdzenie o nierówności i algorytm budowania kodów) do kodowania za pomocą k symboli (zamiast kodowania binarnego).

Zadanie 3

Pokaż, że entropia dla źródła z dwoma symbolami $H(p, 1-p)$ ma maksimum dla $p = \frac{1}{2}$.

Zadanie 4

Niech S oznacza określone źródło symboli, zaś S^k źródło, z którego otrzymujemy bloki k symboli z S, z tym że poszczególne elementy bloku są generowane niezależnie i każdy element bloku jest generowany zgodnie z rozkładem prawdopodobieństwa określonym dla S. Pokazać, że $H(S^k) = kH(S)$. Pokazać, że równość ta może być fałszywa, jeśli pominiemy założenie o niezależności.

Zadanie 5

Załóżmy, że wszystkim literom alfabetu odpowiada to samo prawdopodobieństwo. Jakie kodowanie powinno dać minimalną średnią długość kodu w tej sytuacji.