WYKŁAD 1

PRZEDMIOT INFORMATYKI

Informatyka - dziedzina wiedzy, która zajmuje się przetwarzaniem informacji - w szczególności za pomocą urządzeń technicznych.

Informacja - [def. wg. Mazurkiewicza] Wielkość abstrakcyjna, która może być przechowywana w pewnych obiektach, przetwarzana w pewnych obiektach, przesyłana pomiędzy pewnymi obiektami i stosowana do sterowania pewnymi obiektami. Przez obiekty rozumie się tutaj organizmy żywe, urządzenia techniczne, systemy oraz grupy takich obiektów.

Z informacją są związane 3 podstawowe pojęcia:

A. Komunikat - warstwa fizyczna informacji

B. Kodowanie - zespół ogólnie przyjętych reguł

C. Wiadomość - warstwa abstrakcyjna

dwadzieścia pięć $\xrightarrow{\text{j. polski}}$ liczba dwadzieścia pięć

XXV $\xrightarrow{\text{system rzymski}}$ –//–

11001 $\xrightarrow{\text{s. dwójkowy}}$ –//–

twenty five $\xrightarrow{\text{j. angielski}}$ –//–

25 $\xrightarrow{\text{s. dziesiętny}}$ –//–

2. Elementy ilościowej teorii informacji:

Problem jednoznacznego kodowania i rozkodywania informacji przy użyciu kodu binarnego (ciągów 01 (zero jedynkowych)) prowadzi do pojęcia kodu prefiksowego. Pod pojęciem kodu prefiksowego rozumiemy zbiór słów binarnych, w których każdy z tych ciągów binarnych nie jest początkiem innego ciągu.

WYKŁAD 2

Niech X oznacza zmienną losową przyjmującą wartość w skończonym zbiorze wiadomości $\{x_1, x_2, \ldots, x_n\}$ i rozkładzie prawdopodobieństwa

$$P[X = X_k] = p_k \text{ dla } k = 1, 2, \ldots, n.$$

pojawienia się X_k-tej wiadomości. Niech S oznacza funkcję prefiksowego kodowania binarnego na tym zbiorze wiadomości. Oznaczmy przez moduł $|S|$ funkcję długości kodu S, tzn. $|S(X_k)|$ jest długością kodu $S(X_k)$ reprezentującego wiadomość X_k dla k=1,2,...,n. Przez długość kodu rozumiemy liczbę znaków. Wówczas średnia długość kodu dla pary (X, S) jest równa $E(|S|) = \sum_{k=1}^{n} |S(X_k)| \times P[X = X_k] = \sum_{k=1}^{n} |S(X_k)| P_k$

Oczywiście system kodowania jest tym lepszy im mniejsza jest średnia długość kodu $E(|S|)$.

Można dowieść następującej ~~ti~~ nierówności dla dowolnego prefiksowego kodu binarnego systemu S.

$$E(|S|) \geqslant H(\underline{X}) := \sum_{k=1}^{n} -p_k \log_2 p_k = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \qquad (2.1)$$

Przyjmujemy tutaj, że $0 \cdot \log_2 0 := 0$ i $0 \cdot \log_2 \frac{1}{0} := 0$, co jest naturalne z uwagi na równości

$$\lim_{t \to 0^+} t \log_2 t = \lim_{t \to 0^+} t \log_2 \frac{1}{t} = 0$$

Zauważmy, że $H(\underline{X})$ jest niezależna od systemu kodowania S. Zależy ona jedynie od rozkładu zmiennej losowej $\underline{X}$, czyli od rozkładu prawdopodobieństwa pojawienia się wiadomości x_k-tej, dla $k = 1, 2, \dots, n$.

Liczbę $H(\underline{X})$ nazywamy entropią informacyjną zmiennej losowej $\underline{X}$ lub entropią informacyjną źródła informacji opisywanego przez zmienną losową $\underline{X}$. Z uwagi na nierówność (2.1) entropia informacyjna $H(\underline{X})$ może być uważana jako teoretycznie możliwie najmniejsza średnia długość kodu dla wiadomości $x_1, x_2, \dots, x_n$.

Słowo teoretycznie pojawia się gdyż równość w nierówności (2.1) nie zawsze może być osiągnięta dla zadanego rozkładu zmiennej losowej $\underline{X}$.

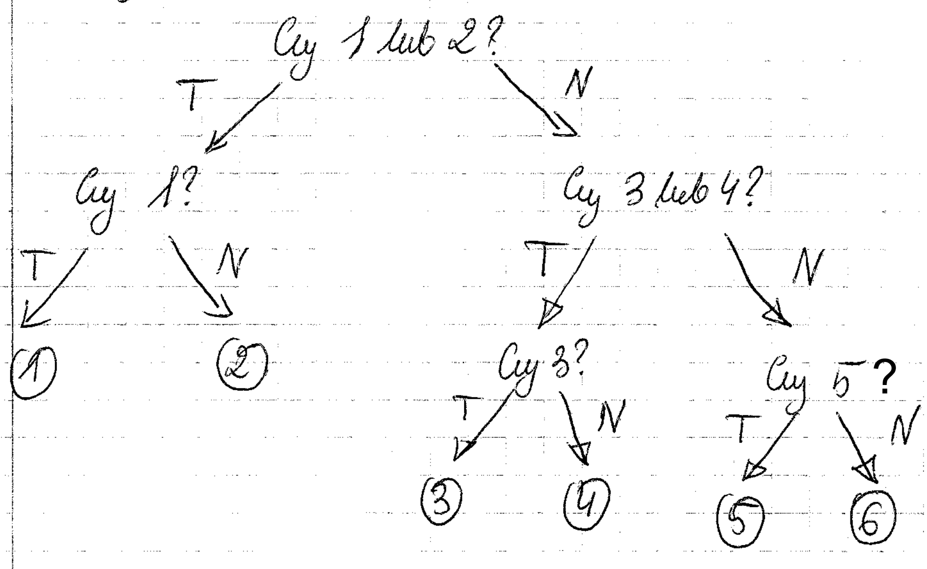
Przykład:

Rzucamy kostkę do gry. Niech $x_k := k$ oznacza liczbę wyrzuconych oczek dla $k = 1, 2, 3, 4, 5, 6$. Mamy zatem n=6 wiadomości $x_1, x_2, \dots, x_6$.

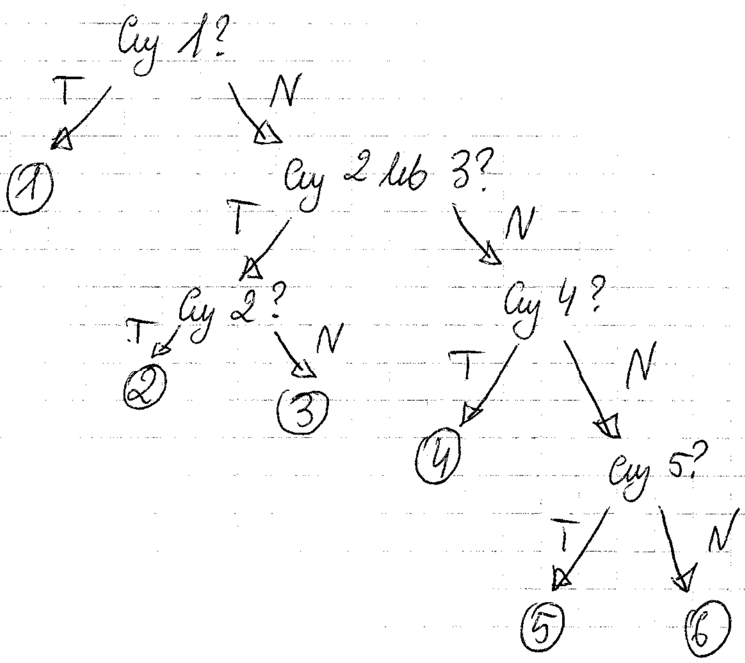
Dowolną funkcję prefiksowego kodowania binarnego S dla zbioru wiadomości można określić za pomocą tzw. binarnego systemu pytań, tzn. systemu prowadzącego do identyfikacji każdej z tych wiadomości w wyniku zadania skończonej ilości pytań na które można odpowiedzieć tak lub nie.

Rozważmy dwa systemy pytań:

I System



II System



$S_1(x_1) := 11$ $\qquad$ $S_2(x_1) := 1$

$S_1(x_2) := 10$ $\qquad$ $S_2(x_2) := 011$

$S_1(x_3) := 011$ $\qquad$ $S_2(x_3) := 010$

$S_1(x_4) := 010$ $\qquad$ $S_2(x_4) := 001$

$S_1(x_5) := 001$ $\qquad$ $S_2(x_5) := 0001$

$S_1(x_6) := 000$ $\qquad$ $S_2(x_6) := 0000$

Rozważmy sytuację gdy kostka jest symetryczna, tzn. $P[X = x_k] = p_k = \frac{1}{6}$ dla $k = 1,2,3 \dots 6$, wówczas

$$E(|S_1|) = \sum_{k=1}^{6} |S_1(x_k)|\, p_k = \frac{1}{6}(2+2+3+3+3+3) = \frac{2}{3} = 2\frac{2}{3}$$

$$E(|S_2|) = \sum_{k=1}^{6} |S_2(x_k)|\, p_k = \frac{1}{6}(1+3\cdot 3+2\cdot 4) = \frac{18}{6} = 3$$

$$E(|S_1|) = 2\frac{2}{3} < 3 = E(|S_2|)$$

Rozważmy teraz sytuację, gdy kostka nie jest symetryczna i rozkład jest następujący: $p_1 = 0{,}85$, $p_k = 0{,}01$ dla $k = 2,3,4,5,6$, wówczas: $E(|S_1|) = 2\cdot 0{,}85 + (2+4\cdot 3)\cdot 0{,}01 = 1{,}8 + 0{,}14 = 2{,}04$

$$E(|S_2|) = 1\cdot 0{,}85 + (3\cdot 3 + 2\cdot 4)\cdot 0{,}01 = 0{,}85 + 0{,}17 = 1{,}12$$

Zatem $E(|S_1|) = 2{,}04 > 1{,}12 = E(|S_2|)$

System S_1 jest gorszy od systemu S_2, bo daje średnio dłuższy kod.

W obu przypadkach entropia $\quad H(X_1) = \sum_{k=6}^{6} p_k \log_2 \frac{1}{p_k} = \log_2 6$

$$H(X_2) = 0{,}85 \cdot \log_2 \frac{100}{85} + 0{,}05 \cdot \log_2 100$$

Pojęcie entropii inf. jest podstawowe dla teorii informacji. Przyjmuje się że mierzy ona ilość inf. zawartej w zbiorze wiadomości opisywanym przez zmienną losową X. Liczbę $p_k \cdot \log_2 \frac{1}{p_k}$ nazywamy ilością informacji zawartej w wiadomości x_k-tej. Ilość inf. wyrażamy w jednostkach zwanych bitami oraz w jednostkach pochodnych do bitów.

1 bit → (1 b) tak się pisze

1 bit → odpowiada ilości informacji zawartej w dwóch jednakowo prawdopodobnych informacjach / wiadomościach.

$P[\underline{X} = x_k] = \frac{1}{2}$, dla $k = 1, 2, \dots$, to $H[\underline{X}] = \frac{1}{2} \cdot \log_2 2 + \frac{1}{2} \log_2 2 =$
$= \log_2 2 = 1$

Jednostki pochodne

1 B (1 bajt) = 8 b

1 KB (1 kilobajt) = 2^{10} B = 1024 B

1 MB (1 Megabajt) = 2^{10} KB = 2^{20} B

1 GB (1 Gigabajt) = 2^{10} MB = 2^{30} B

1 TB (1 Terabajt) = 2^{10} GB = 2^{40} B

Pojemność informacyjna układu pamięciowego:

Załóżmy, że dany jest układ pamięciowy V_N, w którym można zakodować N możliwości $x_1, x_2, \dots, x_N$. Przyjmując, że zmienna losowa $\underline{X}$ ma rozkład jednostajny, tzn. $P[\underline{X} = x_k] = \frac{1}{N}$, dla $k = 1, 2, \dots, N$ otrzymujemy równość

$$H(\underline{X}) = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k} = \sum_{k=1}^{N} \frac{1}{N} \log_2 \frac{1}{1/N} = N \cdot \frac{1}{N} \log_2 N = \log_2 N$$

Liczbę $\log_2 N$ nazywamy pojemnością informacyjną układu pamięciowego V. Wyrażamy ją w jednostkach ilości informacji (b, B, kB, MB, GB, TB).

1 b → $\log_2 N = 1$ → $N = 2^1 = 2$

1 B → $\log_2 N = 8$ → $N = 2^8 = 256$