O oszczędnym dziennikarzu, czyli czym jest informacja i jak ją mierzymy?

Adam Doliwa
doliwa@matman.uwm.edu.pl

WYKŁAD DLA MŁODZIEŻY
WYDZIAŁ MATEMATYKI I INFORMATYKI UWM

Olsztyn, 9 lutego 2016 r.

Kod Morse'a

A	•–	J	•–––	S	•••	2	••–––
B	–•••	K	–•–	T	–	3	•••––
C	–•–•	L	•–••	U	••–	4	••••–
D	–••	M	––	V	•••–	5	•••••
E	•	N	–•	W	•––	6	–••••
F	••–•	O	–––	X	–••–	7	––•••
G	––•	P	•––•	Y	–•––	8	–––••
H	••••	Q	––•–	Z	––••	9	––––•
I	••	R	•–•	1	•––––	0	–––––

Stworzony w 1832 przez Samuela Morse'a i Alfreda Vaila sposób reprezentacji alfabetu, cyfr i znaków specjalnych za pomocą dźwięków, błysków światła, impulsów elektrycznych lub znaków popularnie zwanych kreską i kropką. Alfabetem źródłowym jest alfabet łaciński z cyframi i znakami specjalnymi, a alfabetem kodowym $M = \{\cdot, -, \ \}$ (trzecim elementem zbioru M jest spacja). Spacja oddziela słowa kodowe, grupy znaków oddzielamy trzema spacjami.

ASCII – *American Standard Code for Information Interchange*

Dec	Bin	Char	Dec	Bin	Char	Dec	Bin	Char	Dec	Bin	Char
0	0000000	NUL	32	0100000	space	64	1000000	@	96	1100000	‘
1	0000001	SOH	33	0100001	!	65	1000001	A	97	1100001	a
2	0000010	STX	34	0100010	”	66	1000010	B	98	1100010	b
3	0000011	ETX	35	0100011	#	67	1000011	C	99	1100011	c
4	0000100	EOT	36	0100100	$	68	1000100	D	100	1100100	d
5	0000101	ENQ	37	0100101	%	69	1000101	E	101	1100101	e
6	0000110	ACK	38	0100110	&	70	1000110	F	102	1100110	f
7	0000111	BEL	39	0100111	’	71	1000111	G	103	1100111	g
8	0001000	BS	40	0101000	(	72	1001000	H	104	1101000	h
9	0001001	TAB	41	0101001	)	73	1001001	I	105	1101001	i
10	0001010	LF	42	0100010	*	74	1001010	J	106	1101010	j
11	0001011	VT	43	0100011	+	75	1001011	K	107	1101011	k
12	0001100	FF	44	0100100	,	76	1001100	L	108	1101100	l
13	0001101	CR	45	0100101	-	77	1001101	M	109	1101101	m
14	0001110	SO	46	0100110	.	78	1001110	N	110	1101110	n
15	0001111	SI	47	0100111	/	79	1001111	O	111	1101111	o
16	0010000	DLE	48	0101000	0	80	1010000	P	112	1110000	p
17	0010001	DC1	49	0101001	1	81	1010001	Q	113	1110001	q
18	0010010	DC2	50	0101010	2	82	1010010	R	114	1110010	r
19	0010011	DC3	51	0101011	3	83	1010011	S	115	1110011	s
20	0010100	DC4	52	0101100	4	84	1010100	T	116	1110100	t
21	0010101	NAK	53	0101101	5	85	1010101	U	117	1110101	u
22	0010110	SYN	54	0101110	6	86	1010110	V	118	1110110	v
23	0010111	ETB	55	0101111	7	87	1010111	W	119	1110111	w
24	0011000	CAN	56	0110000	8	88	1011000	X	120	1111000	x
25	0011001	EM	57	0110001	9	89	1011001	Y	121	1111001	y
26	0011010	SUB	58	0110010	:	90	1011010	Z	122	1111010	z
27	0011011	ESC	59	0110011	;	91	1011011	[	123	1111011	{
28	0011100	FS	60	0110100	¡	92	1011100	\	124	1111100	—
29	0011101	GS	61	0110101	=	93	1011101	]	125	1111101	}
30	0011110	RS	62	0110110	¿	94	1011110	ˆ	126	1111110	˜
31	0011111	US	63	0110111	?	95	1011111	_	127	1111111	DEL

BIT – binary digit – cyfra dwójkowa

Problem

Ile bitów potrzeba do zapisania wyniku pięciu rzutów monetą?

ORROR $\leftrightarrow$ 01101

Tyle samo, ile trzeba zadać pytań mających za możliwą odpowiedź

TAK lub NIE

BIT – binary digit – cyfra dwójkowa

Problem

Ile bitów potrzeba do zapisania wyniku pięciu rzutów monetą?

ORROR $\leftrightarrow$ 01101

Tyle samo, ile trzeba zadać pytań mających za możliwą odpowiedź

TAK lub NIE

BIT – binary digit – cyfra dwójkowa

Problem

Ile bitów potrzeba do zapisania wyniku pięciu rzutów monetą?

ORROR $\leftrightarrow$ 01101

Tyle samo, ile trzeba zadać pytań mających za możliwą odpowiedź

TAK lub NIE

Informacja jako redukcja niepewności

Problem

Ile bitów informacji ma wiadomość, że w poprzednim doświadczeniu za każdym razem uzyskaliśmy ten sam wynik?

ALBO

Teraz do opisania wyniku doświadczenia wystarczy TYLKO JEDNO pytanie

$$5 - x = 1 \qquad \Rightarrow \qquad x = 4$$

Wiadomość ta dostarcza nam cztery bity informacji

Informacja jako redukcja niepewności

Problem

Ile bitów informacji ma wiadomość, że w poprzednim doświadczeniu za każdym razem uzyskaliśmy ten sam wynik?

ALBO

Teraz do opisania wyniku doświadczenia wystarczy TYLKO JEDNO pytanie

$$5 - x = 1 \qquad \Rightarrow \qquad x = 4$$

Wiadomość ta dostarcza nam cztery bity informacji

Informacja jako redukcja niepewności

Problem

Ile bitów informacji ma wiadomość, że w poprzednim doświadczeniu za każdym razem uzyskaliśmy ten sam wynik?

ALBO

Teraz do opisania wyniku doświadczenia wystarczy TYLKO JEDNO pytanie

$$5 - x = 1 \qquad \Rightarrow \qquad x = 4$$

Wiadomość ta dostarcza nam cztery bity informacji

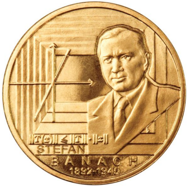
Konkurs

Zadanie 1

Ile bitów informacji ma wiadomość, że rzucając osiem razy monetą uzyskaliśmy raz orła i siedem razy reszkę?

Pierwsza osoba, która nadeśle prawidłowe odpowiedzi na wszystkie (dwa) zadania na adres doliwa@matman.uwm.edu.pl

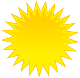
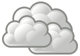
otrzyma w nagrodę kolorowy pendrive z konferencji w Pekinie lub (do wyboru) monetę z najwybitniejszym matematykiem polskim XX wieku



Uwaga: W liście swoje imię i nazwisko należy zakodować używając alfebetu Morse'a

Opowieść o dziennikarzu (część pierwsza)

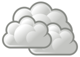
Z Centusiowa do Słoneczkowa został wysłany młody dziennikarz, którego zadaniem było przesyłanie do centrali raz na tydzień wiadomości zbiorczej o pogodzie jaka była każdego dnia.

słonecznie	zachmurzenie małe	zachmurzenie duże	deszcz
			
00	01	10	11

Przykładowa wiadomość: 01001101001001

Opowieść o dziennikarzu (część druga)

Młody dziennikarz zauważył, że w Słoneczkowie (średnio rzecz biorąc) raz na dwa dni jest zachmurzenie małe, raz na cztery dni słonecznie, raz na osiem dni jest zachmurzenie duże oraz raz na osiem dni pada deszcz. Postanowił to wykorzystać do skonstruowania lepszego kodowania wiadomości.

$p_1 = \frac{1}{4}$	$p_2 = \frac{1}{2}$	$p_3 = \frac{1}{8}$	$p_4 = \frac{1}{8}$
			
0	1	10	11

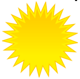
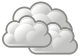
$$p_1 + p_2 + p_3 + p_4 = \frac{2}{8} + \frac{4}{8} + \frac{1}{8} + \frac{1}{8} = 1$$

Odpowiadająca  wiadomość: 101110101

może być odczytana (rozkodowana) jako: 

lub nawet jako: 

W poszukiwaniu optymalnego kodowania

$p_1 = \frac{1}{4}$	$p_2 = \frac{1}{2}$	$p_3 = \frac{1}{8}$	$p_4 = \frac{1}{8}$
			
01	1	001	000

Jest to kod jednoznaczny i natychmiastowy (słowo kodowe nie jest początkiem innego słowa kodowego)

Odpowiadająca pogodzie        wiadomość ma postać: 1010001010011

Czy nowy kod jest lepszy od pierwszego?

Ile zaoszczędzi dziennikarz używając przez 40 tygodni nowy kod zamiast pierwszego, jeśli przesłanie bitu kosztuje 1 PLN ?

Użytkowanie starego kodu kosztuje: $40 \times 7 \times 2 = 560$ [PLN]

Rozwiązanie zadania

Koszty związane z używaniem nowego kodu w ciągo 40 tygodni:

- średnio $20 \times 7 = 140$ dni z małym zachmurzeniem: $1 \times 140 = 140$
- średnio $10 \times 7 = 70$ dni słonecznych: $2 \times 70 = 140$
- średnio $5 \times 7 = 35$ dni z dużym zachmurzeniem: $3 \times 35 = 105$
- średnio $5 \times 7 = 35$ dni deszczowych: $3 \times 35 = 105$

Razem przesłano (średnio rzecz biorąc) $140 + 140 + 105 + 105 = 490$ bitów za sumę 490 PLN.

Dziennikarz oszczędził 560 [PLN] - 490 [PLN] = 70 [PLN]

Uwaga: Średnia długość słowa starego kodu wynosi $E(K_S) = 2$ bity, a nowego kodu wynosi $E(K_N) = 490/280 = 7/4$ bitu

Pytanie

Czy jest to najlepsze (optymalne) kodowanie?

Logarytmy informatyczne i entropia

$\ldots$	$\ldots$
$1 = 2^0$	$\lg 1 = 0$
$2 = 2^1$	$\lg 2 = 1$
$4 = 2^2$	$\lg 4 = 2$
$8 = 2^3$	$\lg 8 = 3$
$16 = 2^4$	$\lg 16 = 4$
$\ldots$	$\ldots$

Entropia źródła

$$\mathcal{H}(p_1, p_2, \ldots, p_n) = p_1 \lg \frac{1}{p_1} + p_2 \lg \frac{1}{p_2} + \ldots p_n \lg \frac{1}{p_n}$$

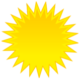
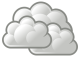
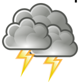
Przykład: Entropia pogody w Słoneczkowie

$$\mathcal{H}\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) = \frac{1}{4} \lg 4 + \frac{1}{2} \lg 2 + \frac{1}{8} \lg 8 + \frac{1}{8} \lg 8 = \frac{7}{4}$$

Twierdzenie Shannona

Dla dowolnego źródła i dowolnego odpowiadającego mu binarnego kodu natychmiastowego średnia długość $E(K)$ słowa kodowego jest nie mniejsza niż entropia tego źródła

$$E(K) \geq \mathcal{H}$$

$p_1 = \frac{1}{16}$	$p_2 = \frac{1}{16}$	$p_3 = \frac{1}{8}$	$p_4 = \frac{1}{2}$	$p_5 = \frac{1}{4}$
				

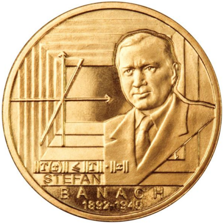
Zadanie 2

Średnie dane pogody w Deszczykowie przedstawione są w tabeli powyżej. Wyznacz odpowiedni optymalny binarny i natychmiastowy kod do przesyłania informacji o pogodzie.

Konkurs

Pierwsza osoba, która nadeśle prawidłowe odpowiedzi na oba zadania na adres doliwa@matman.uwm.edu.pl

otrzyma w nagrodę kolorowy pendrive z konferencji w Pekinie lub (do wyboru) monetę z najwybitniejszym matematykiem polskim XX wieku



Uwaga: W liście swoje imię i nazwisko należy zakodować używając alfebetu Morse'a