

Znajdź (i opisz metodę znalezienia) kod optymalny do zakodowania fragmentu wiersza St. Barańczaka:

*Tu kaczka i tam kaczka.*

*Czy ta kaczka, czy ta czka?*

*Nie: acz klacz czka i pstra czka*

*Paczka perliczek – kaczka*

*Baczy raczej na kacze*

*Obyczaje i kwacze.*

Pomiń spacje i znaki interpunkcyjne i zaniedbaj różnicę między wielkimi i małymi literami. Podaj długość (liczbę bitów) zakodowanego tekstu.

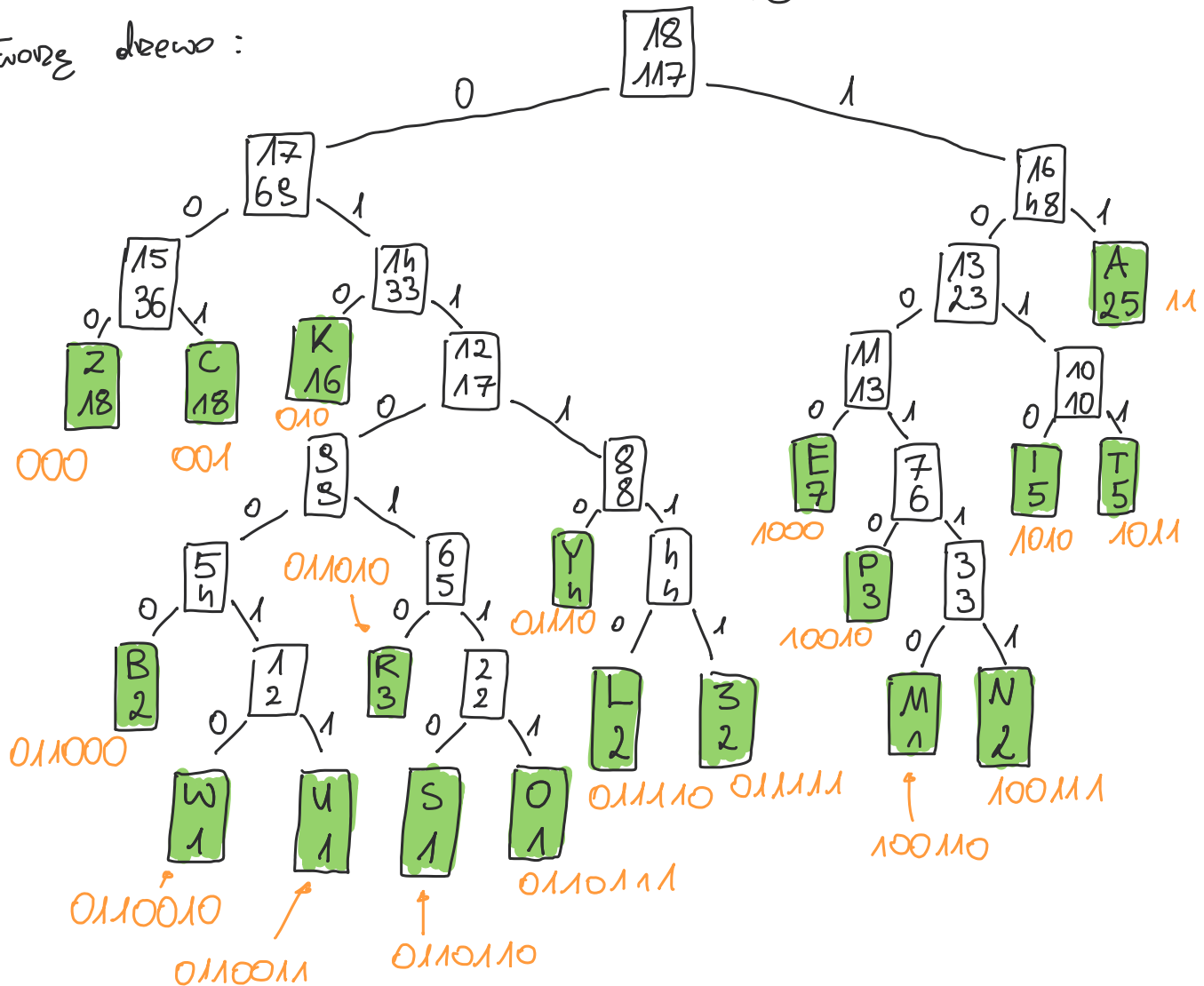
Wskazówka. W tekście występują tylko następujące litery: A B C E I J K L M N O P R S T U W Y Z

Po zliczeniu i uporządkowaniu liter względem liczby wystąpień otrzymujemy:

A	C	2	K	E	I	T	Y	P	R	B	S	L	N	M	O	S	U	W
25	18	18	16	7	5	5	4	3	3	2	2	2	2	1	1	1	1	1

Wszystkich liter jest 117. (nie będę pisał 117 jako mianownika w prawdopodobieństwie kolejnych liter)

Tworzę drzewo:



Ilość bitów potrzebne do zakodowania litery / głębokość na której leży dany węzeł:

A	C	2	K	E	I	T	Y	P	R	B	S	L	N	M	O	S	U	W
2	3	3	3	4	4	4	5	5	6	6	6	6	6	6	7	7	7	7

Zatem po ilość bitów potrzebne do zakodowania tekstu to:

$$\begin{aligned}
 &25 \cdot 2 + 18 \cdot 3 + 18 \cdot 3 + 16 \cdot 3 + 7 \cdot 4 + 5 \cdot 4 + 5 \cdot 4 + 4 \cdot 5 + 3 \cdot 5 + 3 \cdot 6 + 2 \cdot 6 + 2 \cdot 6 + 2 \cdot 6 + \\
 &+ 2 \cdot 6 + 1 \cdot 6 + 1 \cdot 7 + 1 \cdot 7 + 1 \cdot 7 + 1 \cdot 7 = 408
 \end{aligned}$$

Metoda konstrukcji drzew:

1) Utwórz listę drzew binarnych, które w węzłach przechowują parę: [literka, prawdopodobieństwo]. Na początku drzewa składają się wyłącznie z jednego wierzchołka - korzenia.

2) Dopóki na liście jest więcej niż jedno drzewo:

a) Zdejmij z listy dwa drzewa o najmniejszym prawdopodobieństwie zapisanym w ich korzeniach

b) scal drzewa w jedno poprzez podpięcie ich do wspólnego ojca, którego prawdopodobieństwo jest sumą wziętych drzew oraz nie przechowuje on żadnej literki.