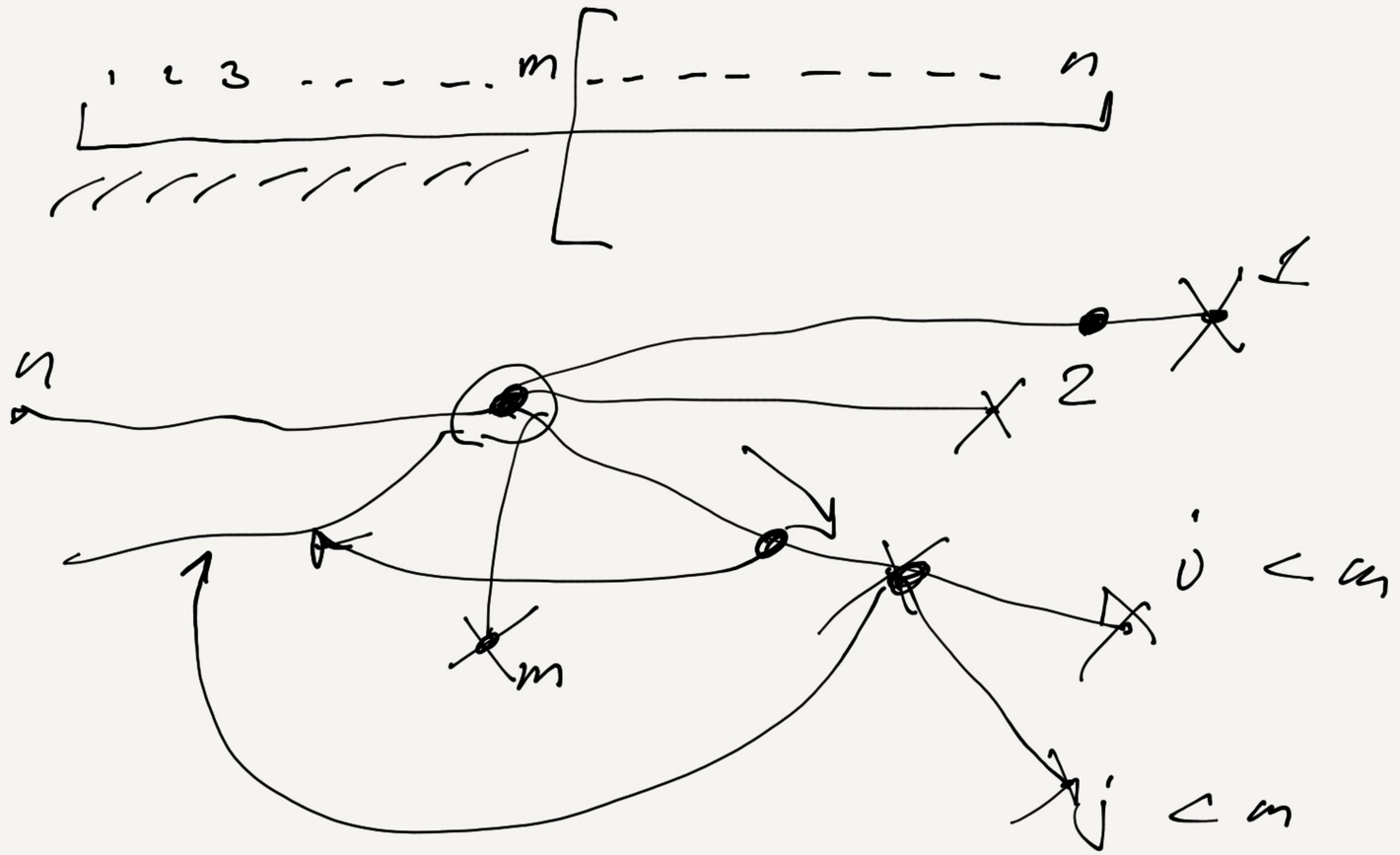


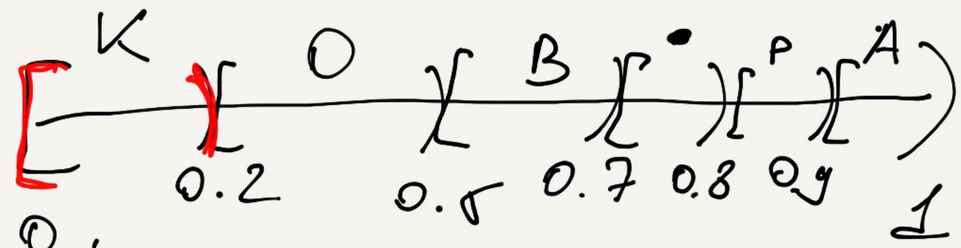
1. Арифметическое кодирование Ataliah
2. PPM MoFFat
3. BWT / MTF / RLE ИТМО

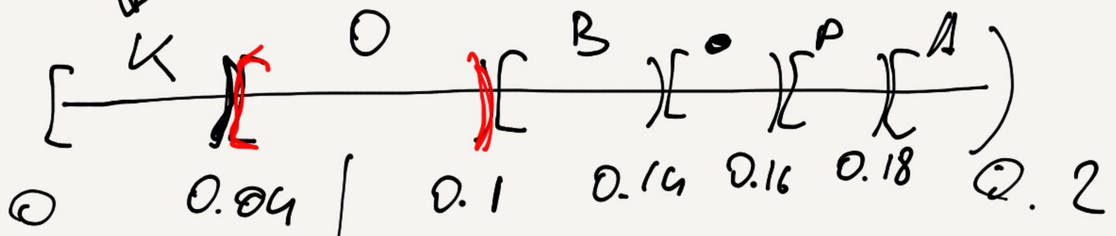
MoFFat

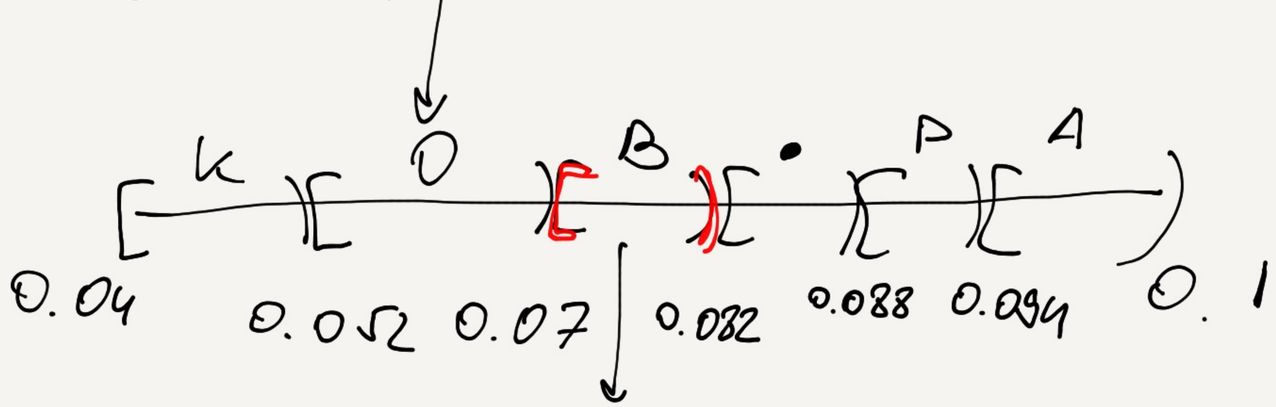
КОВ • КОРОВА

10

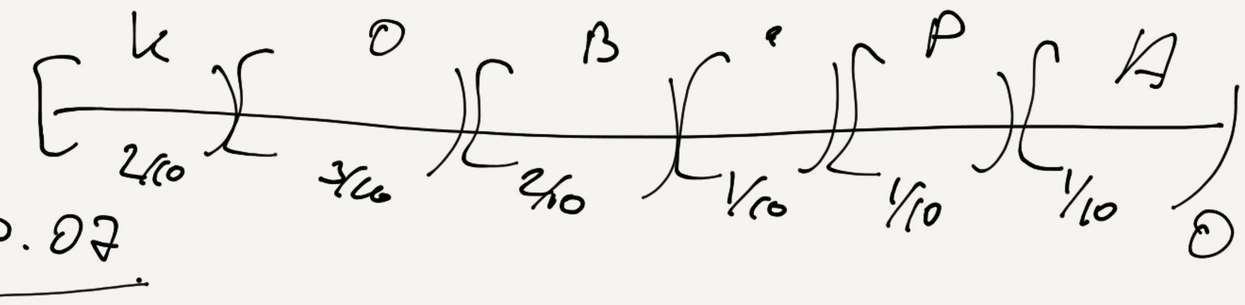
К − 2
О − 3
В − 2
• − 1
Р − 1
А − 1

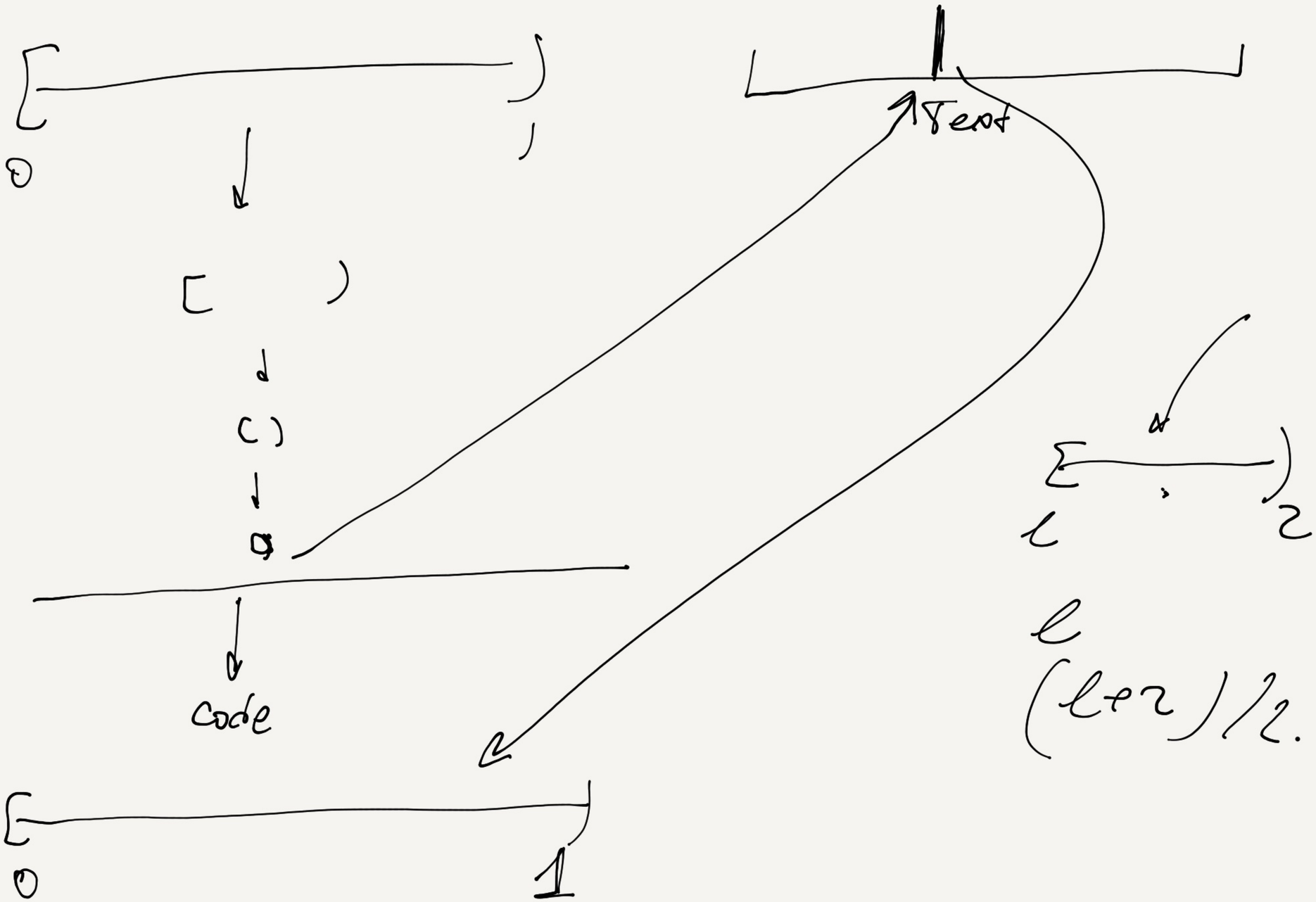






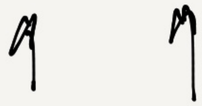
$0.06 = \frac{2}{10}$

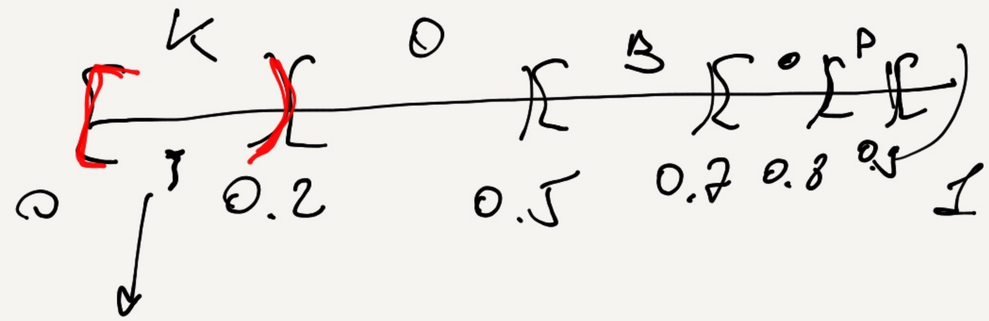



0
1
code
γ_{end}
Σ
1
2
ℓ
ℓ
$(\ell+2)/2$.

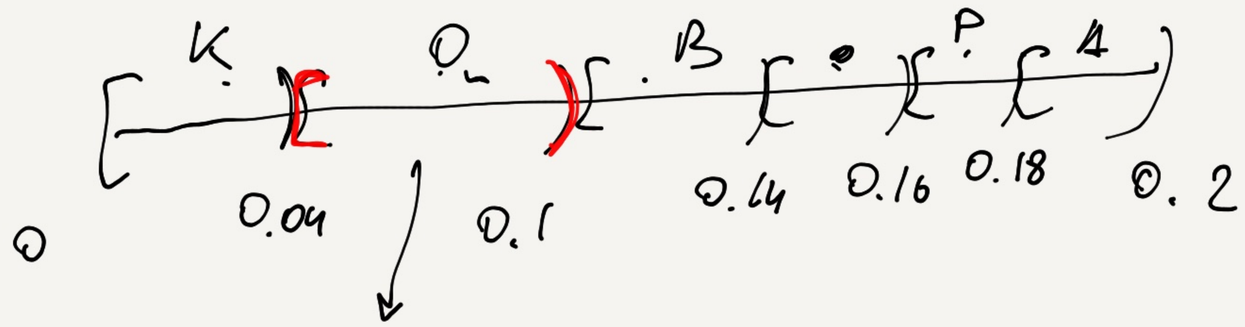
(3, 0.076)

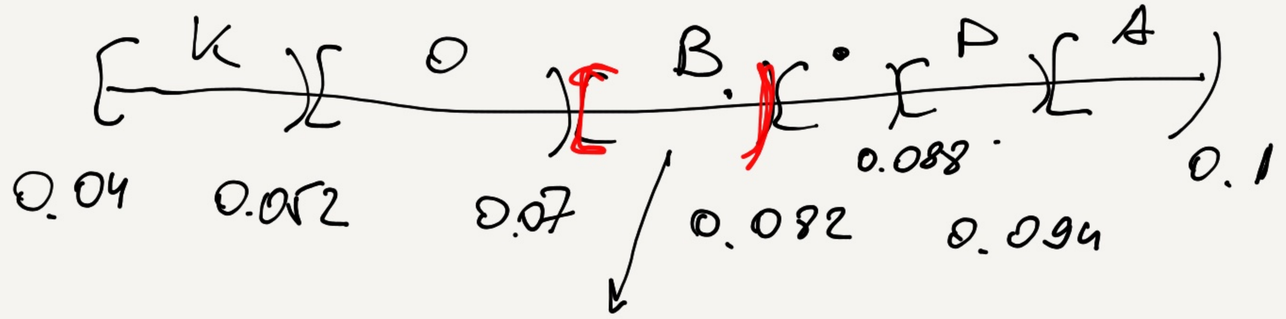
К – 2
О – 3
В – 2
• – 1
Р – 1
А – 1

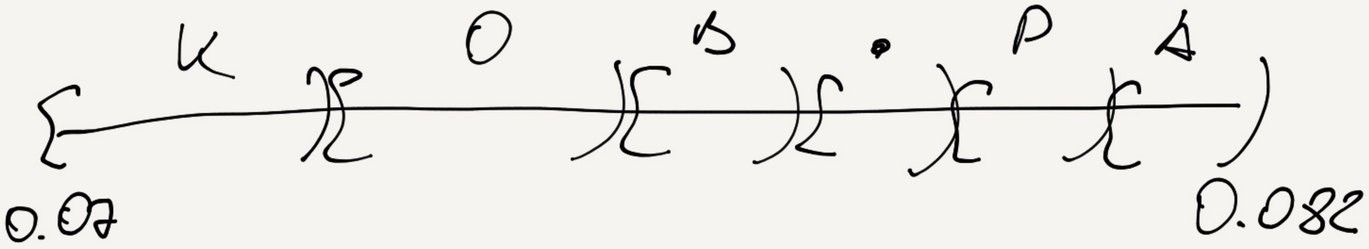


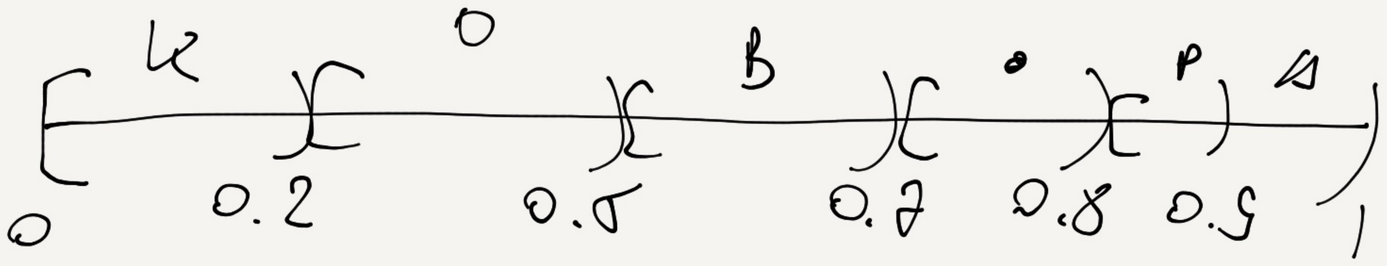


КОВ









A : 0 – 0.03

Б : 0.03 – 0.041

В : 0.041 – 0.056

. – – –

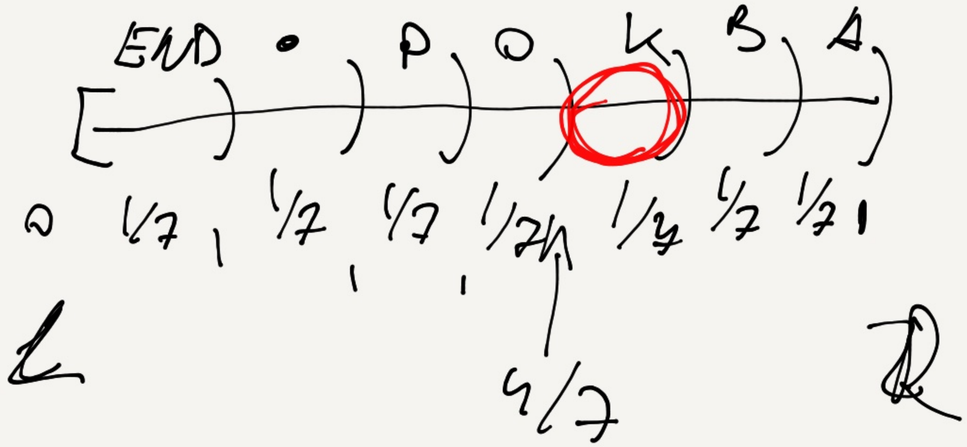
Г. 0.995 – 0.9956

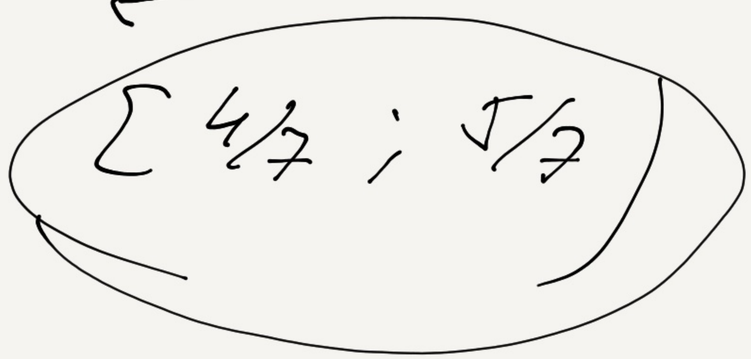
КОВ. ЛЕОПОВА

А В К О Р .

(ART) END ↓

	→	A	B	K	O	P	.	(ART) END
index:	0	1	2	3	4	5	6	7
Freq:	0	1	1	1	1	1	1	1
Cum_Freq:	7	6	5	4	3	2	1	0

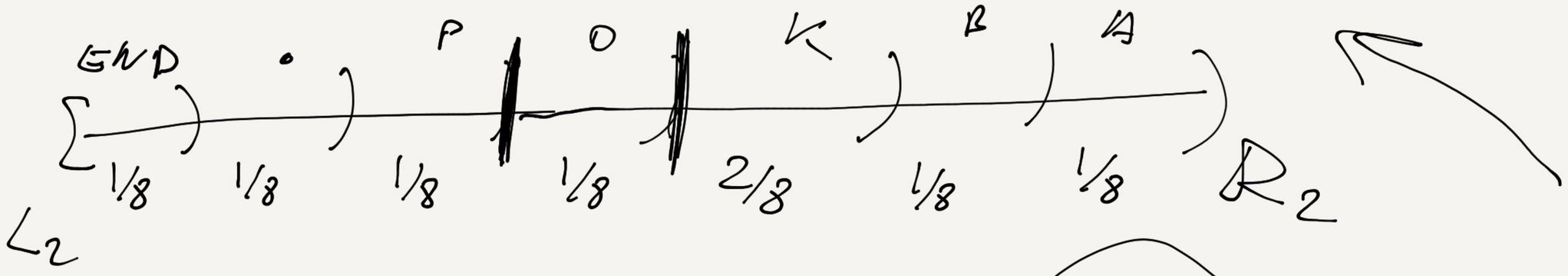




$$L + (R - L) \cdot 4/7 = L_2$$

$$L + (R - L) \cdot 5/7 = R_2$$

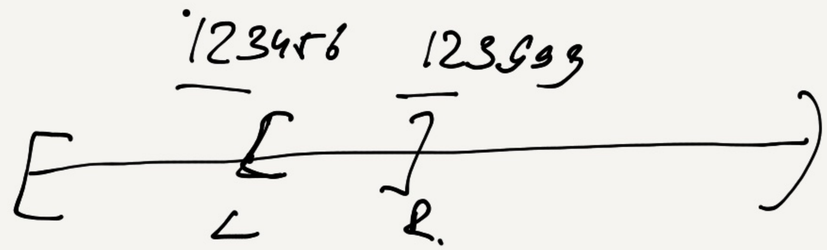
		A	B	K	O	D	.	END
index	0	1	2	3	4	5	6	7
Freq:	0	1	1	2	~~1~~ 2	1	1	1
Cum_Freq:	~~8~~ 9	~~7~~ 8	~~6~~ 7	~~4~~ 5	3	2	1	0



$$\begin{cases} L_2 + (R_2 - L_2) \cdot 3/8 = L_3 \\ L_2 + (R_2 - L_2) \cdot 4/8 = R_3 \end{cases}$$

$L = 0$ | $R = 1$ ⟶ $L = 0$ $\quad R = 2^n - 1$

$n = 3$



$L = 00000000_2$ $\quad R = 11111111_2$

400000 $\quad$ 999999

000456 $\quad$ 999999

$L = 00110011_2$ $\quad R = 0011\,1111_2$

$L = 00110000$ $\quad R = 11111111$

Code ⟶ Stream

0011 $\quad$ Output

$R - L <$ Coded_Symbols

||

Cum_Freq[0]

Prediction by Partial Matching

PPM.

u — 2%

q u — 95%

DBCDEX

encoded | To Do

ABCDE | X | y

- не встречалась
- встречалась, но без x
- встречалась с x.

ABCDE

ABCDE: A — 5, B — 10, C — 20, x — 10, ART-1

freq = 5/46, 10/46, 20/46, 10/46, 1/46 → code(X)

ABCDE: X, +1

5 → 4

ABCDE → BCDE

BCDE: x+1

CDE: x+1

DE: x+1

ABCDE: A — 5, B — 20, C — 30, E: x+1, x+1

ART — 1.

ABCDE → BCDE

code[ART] →

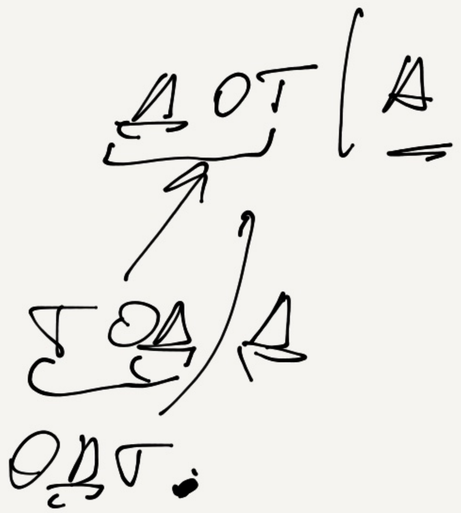
freq(ART) = 1/56

freq(A) = 5/56

freq(B) = 20/56

freq(C) = 30/56

Context : Symbol : Frequency



ABCDE

↓

BCDE

↓

CDE

↓

DE

↓

E → ∅

A = 20

B = 30

C = 40

~~X = 10 → Code~~

:

~~ARE~~ = 1. → ~~[X]~~

X = 1

ABCABD

Context = 3

Code (A) ART code (B) ART code (C) A B D

UTF-8 1 UTF-8 11 UTF-8 00 1

ctx=1

ctx=2 ctx=1 ctx=0

ART ART ART code (D)

0 0 000 UTF-8

Context = 1

Context = 0 :

ART = 1.

A = 2

B = 2

C = 1

D = 1

A : ART = 1

B = 2

B : ART = 1.

C = 1.

D = 1

C : ART = 1

A = 1

PPMA ART = 1

PPMC : Freq(ART) = R / Rsum

ART = Freq(ART) = 4/29

A - 5.

B - 10. Freq(B) = 10/29

C - 10.

Context = 2 :

CA : ART = 1
B = 1

CAB : ART = 1
D = 1

AB : ART = 1
C = 1
D = 1

BC : ART = 1
A = 1

Context = 3 :

ABC : ART = 1
A = 1

BCA : ART = 1
B = 1

ABC : ART = 1

A = 10

B = 20

C = 30

BC : ART = 1

~~A~~ = 50

~~B~~ = 30

~~C~~ = 40

D = 20

E = 60

X = 10

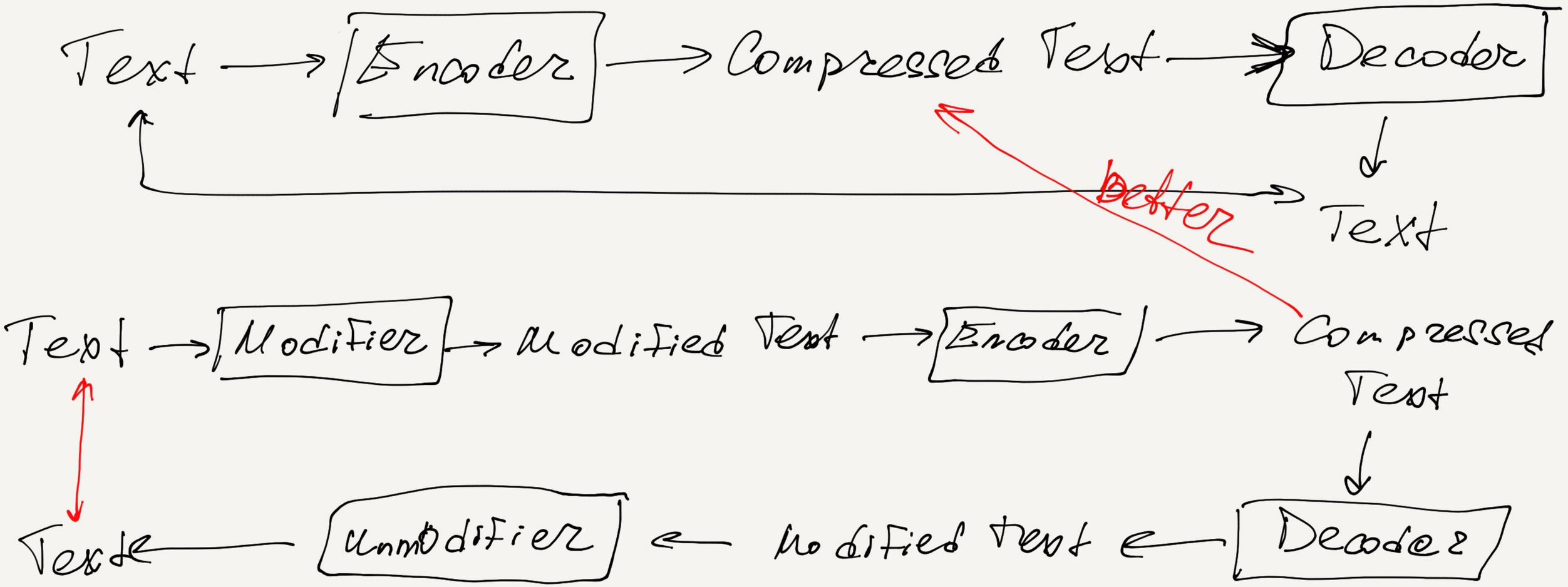
ABC, X

BCX

Freq(X) = $\frac{10}{191}$ → Freq(X) = $\frac{10}{71}$

~~C~~ X = 3

~~ART~~


Text
Encoder
Compressed Text
Decoder
Better
Text
Text
Modifier
Modified Text
Encoder
Compressed Text
Text
Unmodifier
Modified Text
Decoder

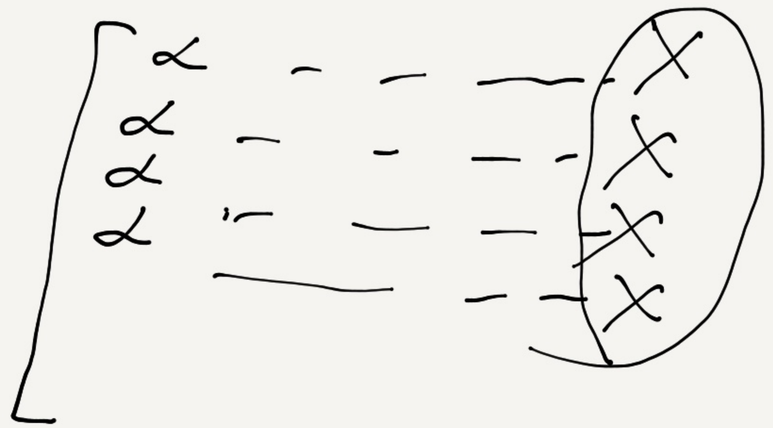
Барроуза – Уилера
Burrows – Wheeler Transform (BWT)

word = Xα

- АБРАКАДАБРА
- БРАКАДАБРАА
- РАКАДАБРААБ
- АКАДАБРААБР
- КАДАБРААБРА
- АДАБРААБРАК
- ДАБРААБРАКА
- АБРААБРАКАД
- БРААБРАКАДА
- РААБРАКАДАБ
- ААБРАКАДАБР

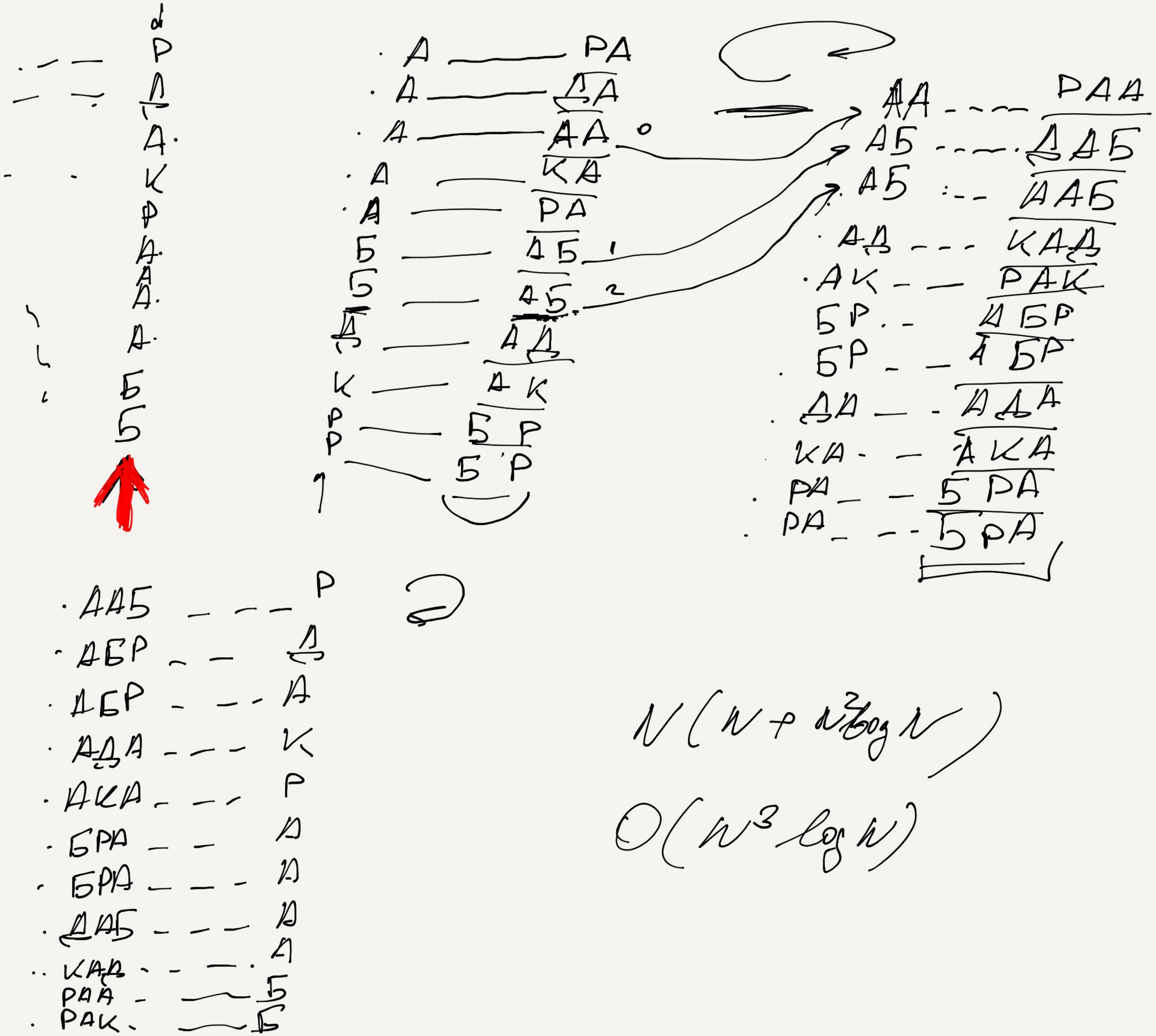
0. ААБРАКАДАБР
1. АБРААБРАКАД
2. АБРАКАДАБРА
3. АДАБРААБРАК
4. АКАДАБРААБР
5. БРААБРАКАДА
6. БРАКАДАБРААА
7. ДАБРААБРАКАА
8. КАДАБРААБРАА
9. РААБРАКАДАБ
10. РАКАДАБРААБ

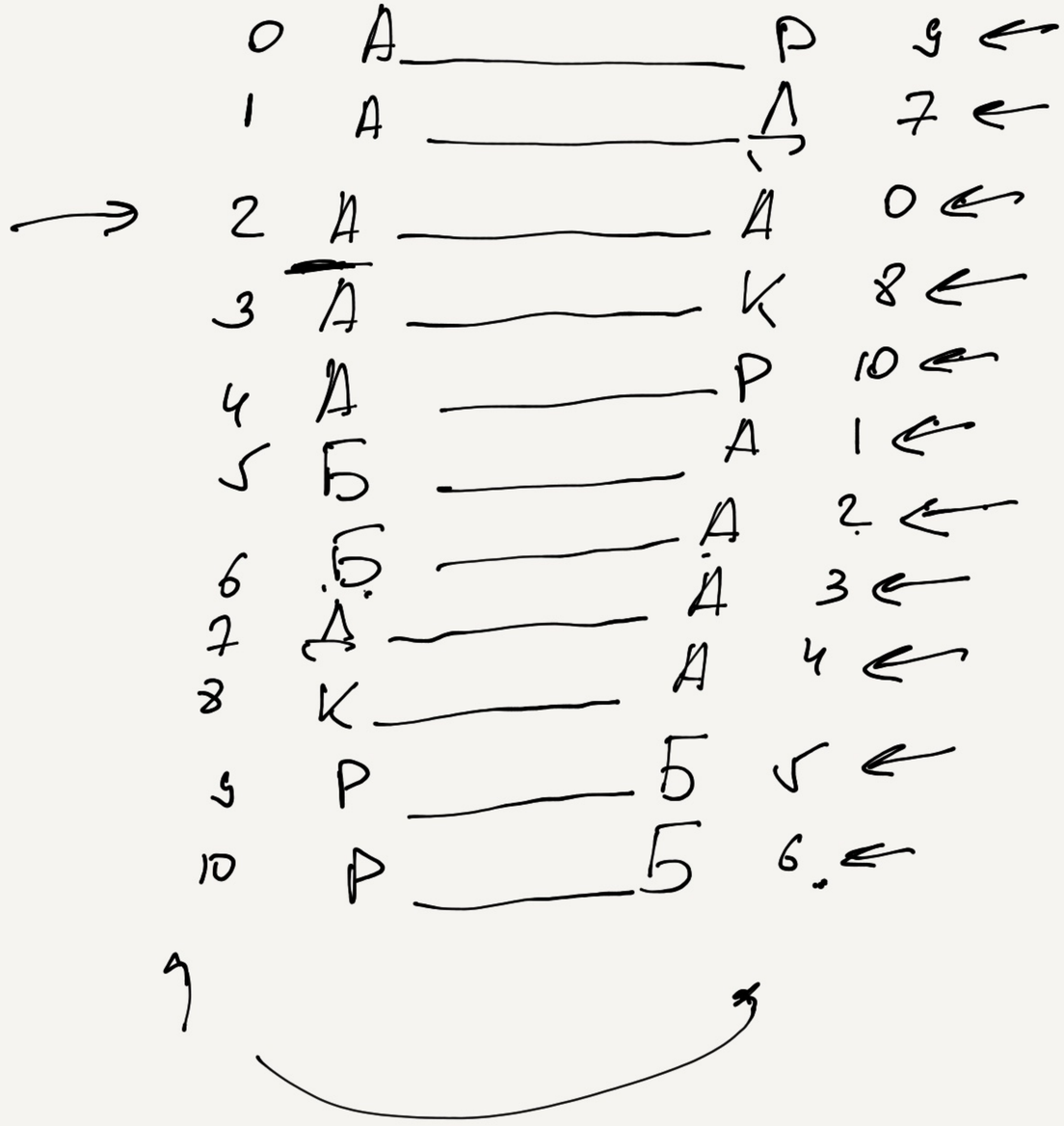
Xα
Xα
Xα
Xα



АБДАКАДАБРА

< РДАКРАААААББ 2 >


АА — РАА
АБ — ДАБ
АБ — ААБ
АД — КАД
АК — РАК
БР — АБР
БР — АБР
ДА — ДДА
КА — АКА
РА — БРА
РА — БРА
ААБ — Р
АБР — Д
АБР — А
АДА — К
АКА — Р
БРА — Д
БРА — Д
ДАБ — Д
КАД — Д
РАА — Б
РАК — Б
N(N + n³ log n)
O(N³ log N)



2 Р Д А К Р А А А Б Б, 2>

А Б Р А К А Д А Б Р А

O(N)

А Б Р А К А Д А Б Р А → Р Д А К Р А А А Б Б

Move To Front (MTF)

РДАКРАДАДАББ

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

4 3 2 4 3 2 0 0 0 4 0

АБРАКАДАБРА → 23

↓ Modified = Huffman

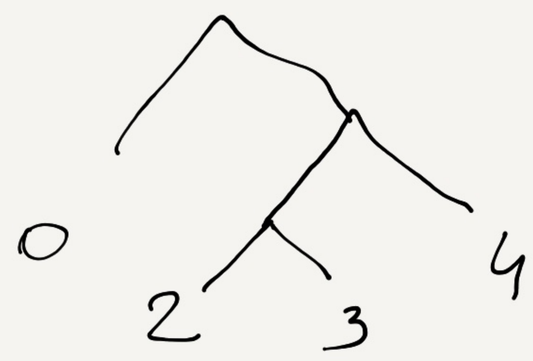
4 3 2 4 3 2 0 0 0 4 0

0 – 4

2 – 2 } 4

3 – 2 }

4 – 3



word bits

0 – 1

4 – 2

2 – 3

3 – 3

АБДКР

0 1 2 3 4

РАБДК

ДРАБК

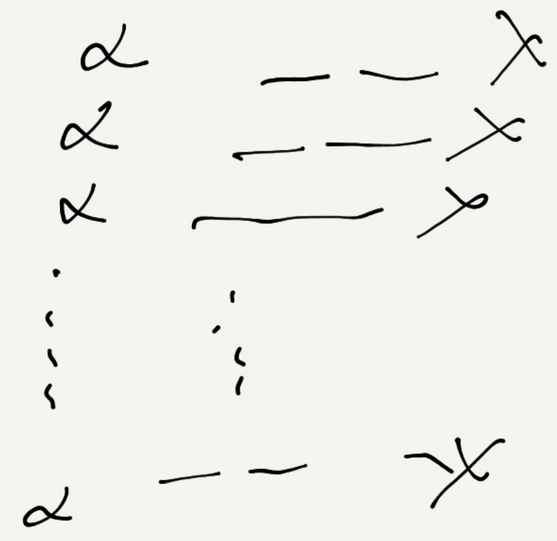
АДРБК

КАДРБ

РКАДБ

ДРКАБ

БАРКД



XXX... ... XY...

k 00... .0 l2..0

4 + 6 + 6 + 6 = 22

4 3 2 4 3 2 0 0 0 4 0

Р Д А К Р А А А А Б Б

Х . . - - - . . . Х

↓

К О . - - . ◎

А Б Р А К А Д А Б Р А

1 1 1 1 1 1 1 1 1 1 1

0 1 4 2 4 1 4 1 4 4 2

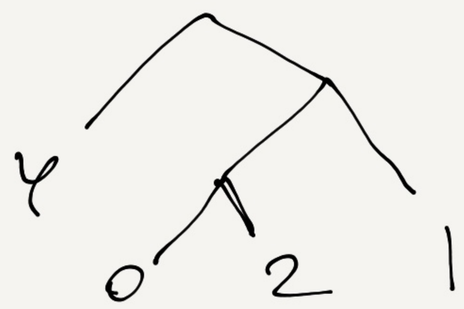
word bits

0 – 1

1 – 3

2 – 2

4 – 4



А Б Д К Р

0 1 2 3 4

Б А Д К Р

Р Б А Д К

А Р Б Д К

К А Р Б Д

Д А К Р Б

А К Р Б Д

Д А К Р Б

А Д К Р Б

А Б Д К Р

0 1 2 3 4

Р А Б Д К

Д Р А Б К

А Д Р Б К

К А Д Р Б

Р К А Д Б

А Р К Д Б

Б А Р К Д

Б А Д К Р

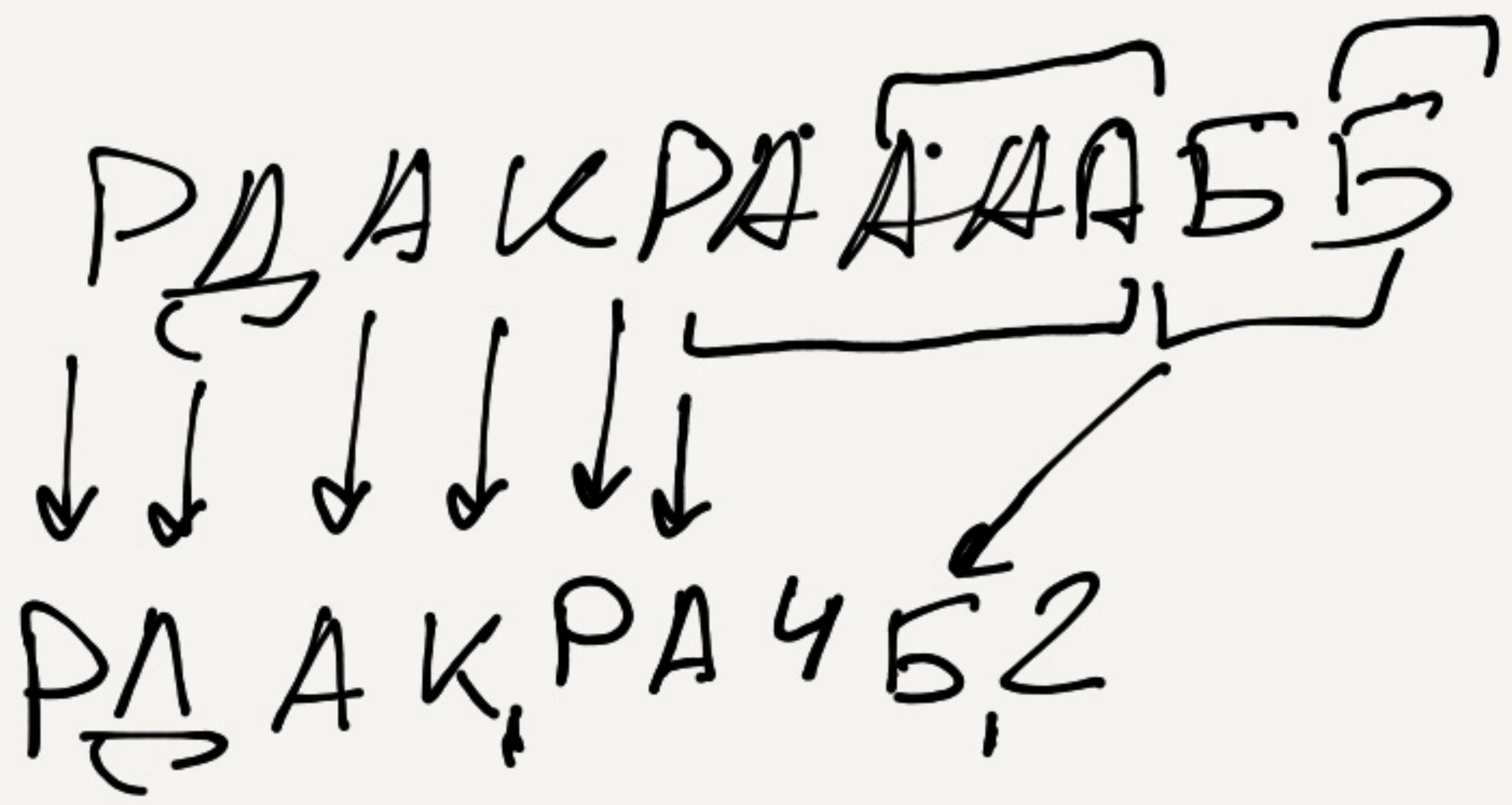
Р Б А Д К

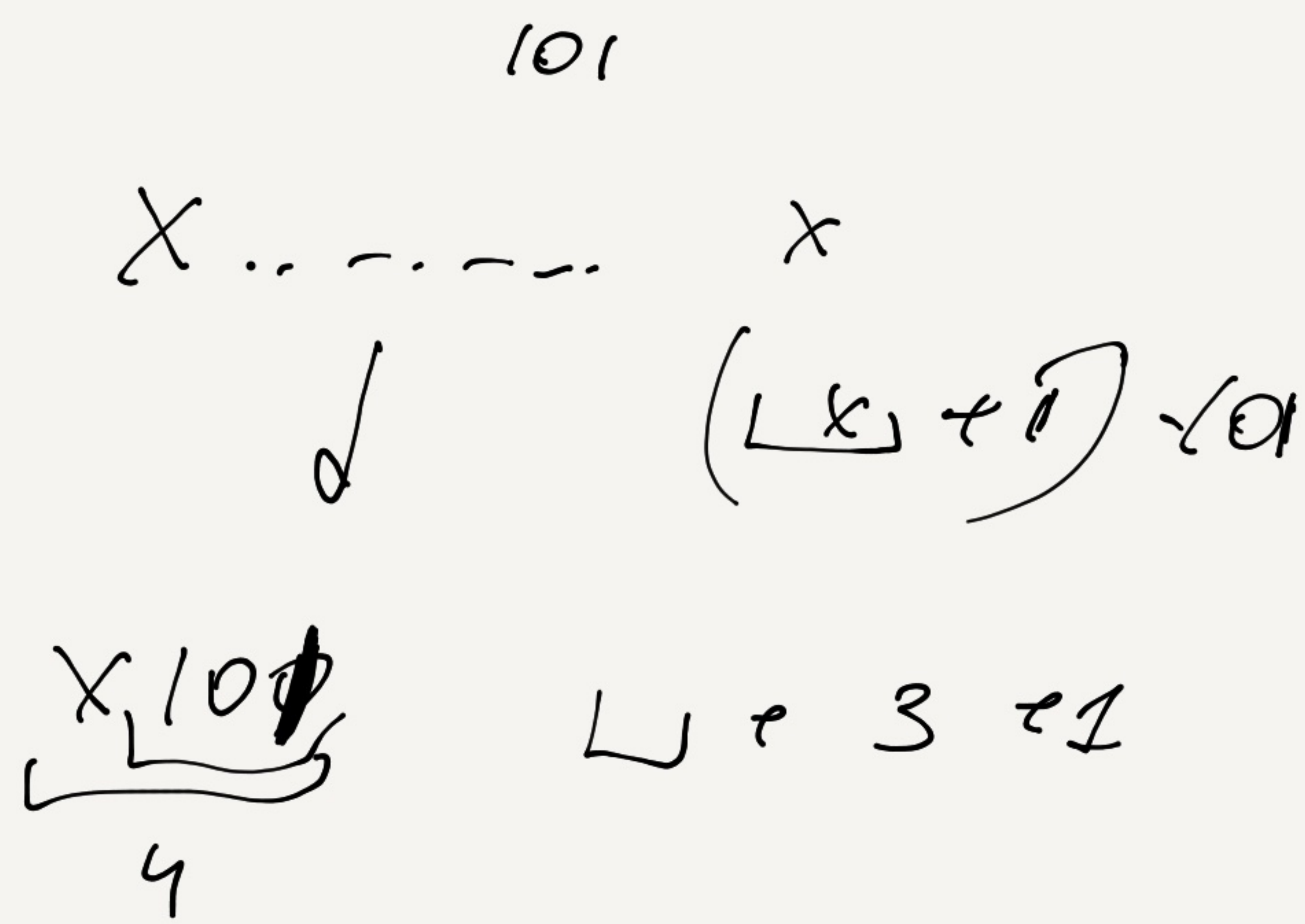
А Р Б Д К.

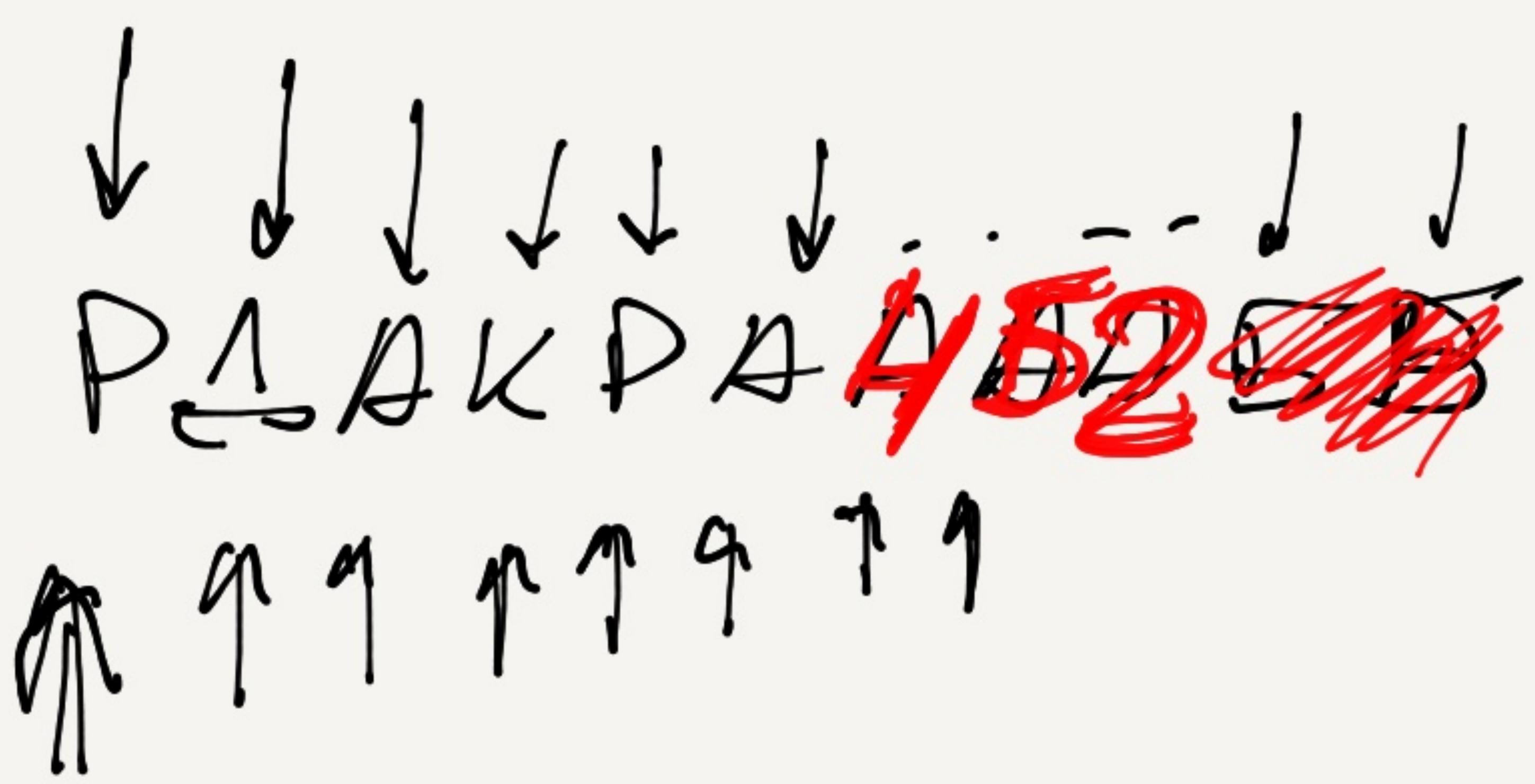
4 + 6 + 6 + 3 = 19.

Run Length Encoding

RLE





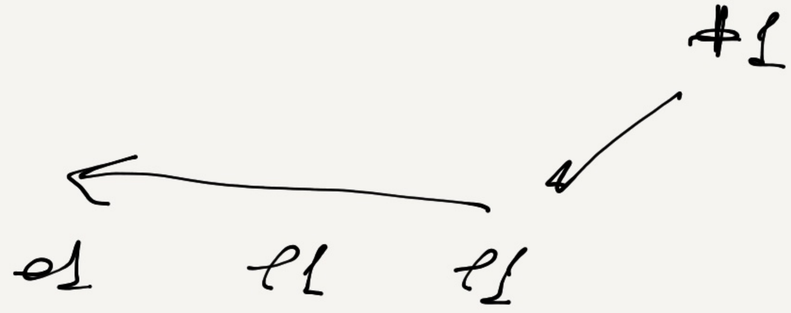


А В К О Р . ЄN

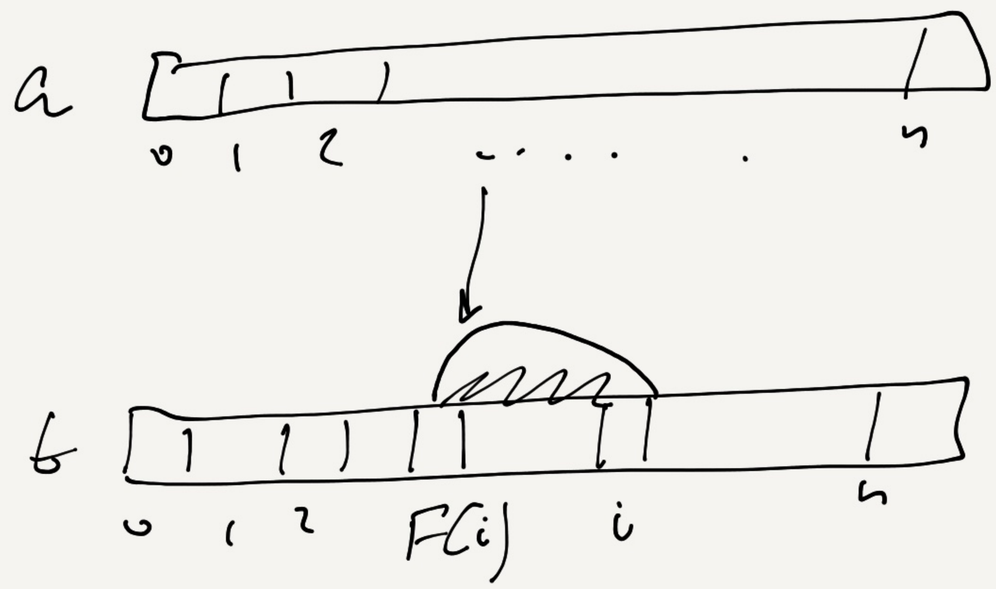
index

Freq

cum_freq



Дерево Фенвика



$$t_i = \sum_{j=F(i)+1}^{i} a_j$$

$$t_7 = \sum_{j=0}^{7} a_j$$

$$t_9 = \sum_{j=8}^{9} a_j$$

$F(i) = i \,\&\, (i+1)$

0001101001101111

0000

111 & 1000 = 0.

1001 & 1010 = 1000

$0 \ldots j$

$a[i] + a[i+1] + \ldots + a[j]$

$O(n)$

$t[j] = a[F(j)] + \ldots + a[j]$. $O(\log n)$

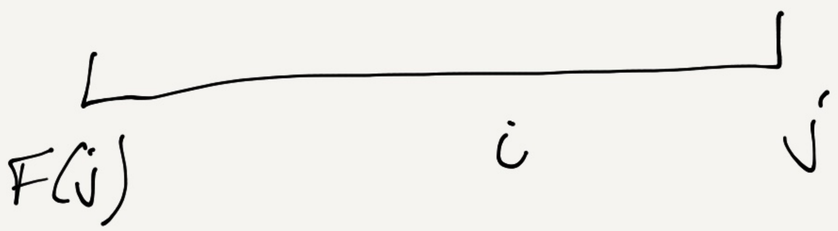
$t[F(j)-1]$

$0 \ldots j$

$0 \sim i-1$ $t[i-1] \ldots$

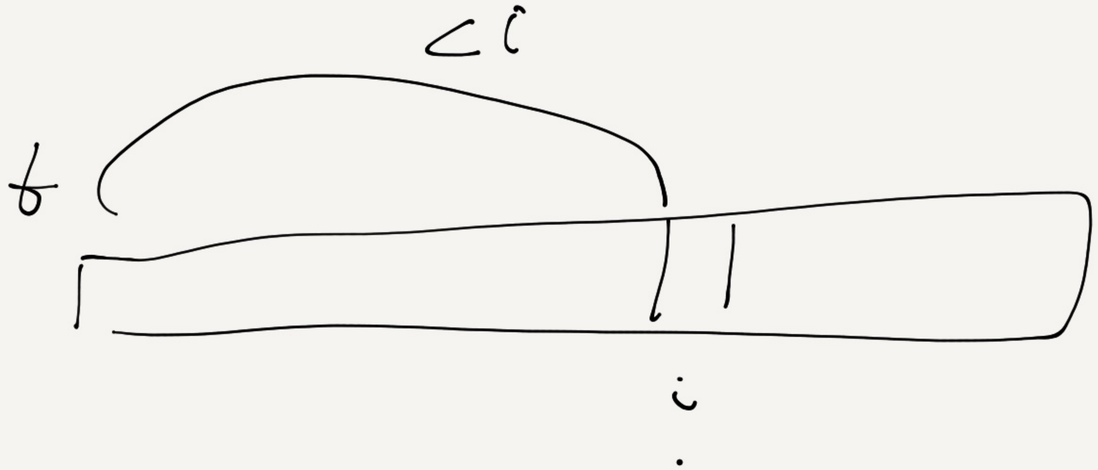
$t[F(F(j)-1)-1]$.

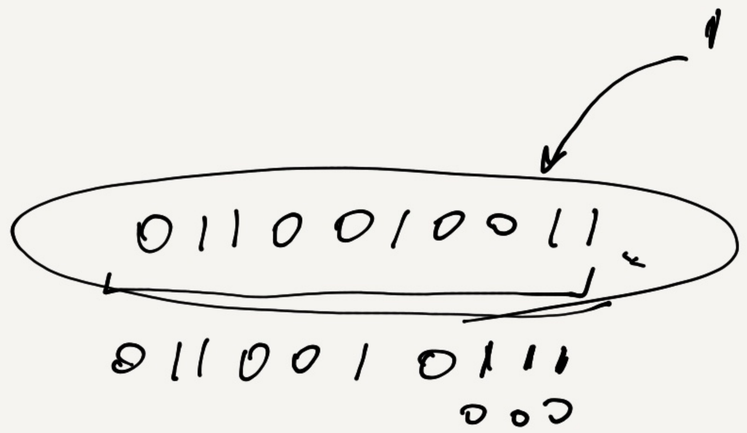
$a[i] += 3$



$\in [j]$

$F(j) \leq i \leq j$





$F(i) = i \,\&\, (i+1)$

$b[i] += 3$

$i_{new} = i \mid (i+1)$

$b[i_{new}] += 3$

$i_{new2} = i_{new} \mid (i_{new}+1)$

$b[i_{new2}] += 3$