



**BIGDATA  
TEAM**

Юникод, кодировки, pytest:capsys

# Спецификации utf-8 и koï8-r, читаем байты правильно

**Драль Алексей**, [study@bigdatateam.org](mailto:study@bigdatateam.org)

CEO at BigData Team, <https://bigdatateam.org>

<https://www.facebook.com/bigdatateam>



## koi8-r encoding

KOI8-R																
	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_0																
1_16																
2_32	SP 0020	! 0021	" 0022	# 0023	\$ 0024	% 0025	& 0026	' 0027	( 0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[ 005B	\ 005C	] 005D	^ 005E	_ 005F
6_96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	 007C	}	~ 007D	
8_128	— 2500	 2502	Г 250C	г 2510	Л 2514	л 2518	Т 251C	т 2524	т 252C	т 2534	■ 253C	■ 2584	■ 2588	■ 258C	■ 2590	
9_144	☒ 2591	☒ 2592	☒ 2593	☒ 2320	☒ 25A0	☒ 2219	☒ 221A	☒ 2248	☒ 2264	☒ 2265	NBSP 00A0	J 2321	o 00B0	2 00B2	• 00B7	+ 00F7
A_160	= 2550	 2551	F 2552	ѐ 0451	ѓ 2553	ѓ 2554	ѓ 2555	ѓ 2556	ѓ 2557	ѓ 2558	ѓ 2559	ѓ 255A	ѓ 255B	ѓ 255C	ѓ 255D	ѓ 255E
B_176	Ѡ 255F	ѡ 2560	Ѣ 2561	Ѥ 0401	Ѧ 2562	Ѩ 2563	Ѭ 2564	Ѯ 2565	Ѱ 2566	Ѳ 2567	Ѵ 2568	Ѷ 2569	Ѹ 256A	Ѻ 256B	Ѽ 256C	Ѿ 00A9
C_192	ю 044E	а 0430	б 0431	ц 0446	д 0434	е 0435	ф 0444	г 0433	х 0445	и 0438	й 0439	к 043A	л 043B	м 043C	н 043D	о 043E
D_208	п 043F	я 0447	р 0440	с 0441	т 0442	у 0443	ж 0436	в 044C	ь 044B	ы 0448	э 0447	ш 0448	э 044D	щ 0449	ч 0447	ъ 044A
E_224	Ю 042E	А 0410	Б 0411	Ц 0426	Д 0414	Е 0415	Ф 0424	Г 0413	Х 0425	И 0418	Й 0419	К 041A	Л 041B	М 041C	Н 041D	О 041E
F_240	П 041F	Я 042F	Р 0420	С 0421	Т 0422	У 0423	Ж 0416	В 0412	Ь 042C	Ы 042B	Э 0417	Ш 0428	Щ 042D	Ч 0429	Ъ 0427	

5	0438	0439	043A	043B	043C	043D
	Ы	Э	Ш	Э	Щ	Ч
С	044B	0437	0448	044D	0449	0447
	И	Й	К	Л	М	Н
5	0418	0419	041A	041B	041C	041D
	И	Й	Ш	Э	Щ	Ч

©

<https://en.wikipedia.org/wiki/KOI8-R>



# utf-8 encoding

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx



- ▶ koir-8 и utf-8 - знаем что внутри
- ▶ encoding = подмножество Unicode + правила
- ▶ utf-8-be / utf-8-le / utf-8-BOM

