Юникод, кодировки, pytest:capsys

Спецификации utf-8 и koi8-r, читаем байты правильно

Драль Алексей, study@bigdatateam.org CEO at BigData Team, https://bigdatateam.org https://www.facebook.com/bigdatateam



KOI8-R SP 0021 0022 0023 0024 0025 0026 0027 0028 0029 002A 002B 002C 002D 002E 002F 0 5 0030 0031 0032 0033 0034 0035 0036 0037 0038 0039 003A 003B 003C 003D 003E 003F M 0040 0041 0042 0043 0044 0045 0046 0047 0048 0049 004A 004B 004C 004D 004E 004F U 80 0050 0052 005E 005F 0051 0053 0054 0055 0056 0057 0058 0059 005A 005B 005C 005D d e m 006C 006E 0060 0061 0062 0063 0064 0065 0066 0067 0068 0069 006A 006B 006D 006F u Z 112 0070 0071 0072 0073 0074 0075 0076 0077 0078 0079 007A 007B 007C 007D 007E 128 2502 250C 2510 2518 251C 2524 252C 2534 253C 2580 2588 258C 2590 NBSP 144 2591 2592 2593 25A0 2219 221A 2264 2265 00A0 00B0 00B2 00B7 00F7 160 2552 2553 2550 2551 0451 2554 2555 2556 2557 2558 2559 255A 255B 255C 255D 255E 176 255F 2560 2561 0401 2562 2563 2564 2565 2566 2567 2568 2569 256A 256B 256C 00A9 M Ю 192 044E 0430 0431 0446 0434 0435 0444 0433 0445 0438 0439 043A 043B 043C 043D 043E ы П ш Щ 043F 044F 0440 0441 0442 0443 0436 0432 044C 044B 0437 0448 044D 0449 0447 044A Й Ц И Л M 0 A 224 042E 0410 0411 0426 0414 0415 0424 0413 0425 0418 0419 041A 041B 041C 041D 041E П C Ы 3 Ш Э Щ В 041F 0420 042F 0421 0422 0423 0416 0412 042C 042B 0417 0428 042D 0429 0427

koi8-r encoding

	TT	n	TIT	n	TTT	TT
5	0418	0419	041A	041B	041C	041D
	И	Й	K	Л	M	Н
С	044B	0437	0448	044D	0449	0447
	ы	3	ш	Э	щ	Ч
5	0438	0439	043A	043B	043C	043D



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	0xxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxx	10xxxxx	10xxxxx

© https://en.wikipedia.org/wiki/UTF-8

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

