



Unicode, code points, characters & glyphs

Драль Алексей, study@bigdatateam.org

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

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



Неправильное указание кодировки

А. С. Пушкин. Евгений Онегин

Александр Пушкин

ЕВГЕНИЙ ОНЕ

РОМАН В СТИХАХ

Pétri
espèce d'
indifférenc
actions, s
être imagi

Не мысля гордый свет заб
Внимание дружбы возлюб
Хотел бы я тебе представ
Залог достойнее тебя,
Достойнее души прекрас
Святой исполненной мечт

стр. 1 из 12

ю, я, оСЬЙХМ, еБЦЕМХИ нМЕЦХМ

ЮКЕЙЯЮМДП ОСЬЙХМ

ЕБЦЕМХИ НМЕЦХМ

ПНЛЮМ Б ЯРХУЮУ

Pétri de vanitéil avait encore plus de cette espèce d'orgueil quifait avouer avec la même indifférence lesbonnes comme les mauvaises actions, suite d'un sentiment desupériorité peut-être imaginaire.

Tiré d'une lettre particulière. ¹

Pétri de vanitéil avait encore
espèce d'orgueil quifait avouer :
indifférence lesbonnes comme
actions, suite d'un sentiment desu
être imaginaire.

Tiré d'une lettre

МЕ ЛШЯКЪ ЦНПДШИ ЯБЕР ГЮАЮБХРЭ,
БМХЛЮМЭ ДПСФАШ БНГКЧАЪ,
УНРЕК АШ Ъ РЕАЕ ОПЕДЯРЮБХРЭ
ГЮКНЦ ДНЯРНМЕЕ РЕАЪ,
ДНЯРНМЕЕ ДСЪХ ОПЕЙПОЯМНИ,
ЯБЪРНИ ХЯОНКМЕММНИ ЛЕВРШ.

ЯРП. 1 ХГ 12

[illegible]

◆◆◆, 1 ◆◆ 12

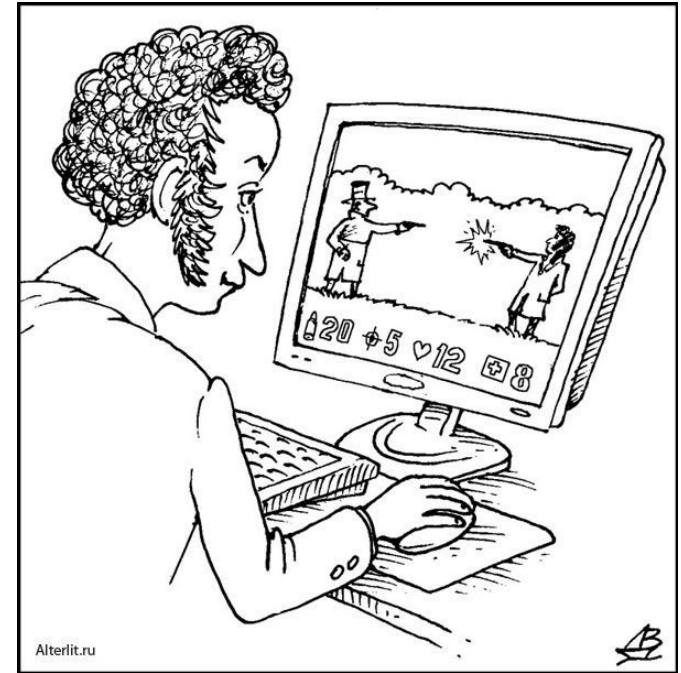


...

Unicode... как много в этом звуке

Для сердца русского слилось!

...



Цитата из романа в стихах "е-Гений Онегин" (1993 – 2005 гг.)
русского поэта Александра Сергеевича Пушкина



Фрост vs Пушкин, раунд #1



Some say the world will end in fire,
Some say in ice.
From what I've tasted of desire
I hold with those who favor fire.

...

МЕ ЛШЯКЪ ЦНПДШИ ЯБЕР ГЮАЮБХРЭ,
БМХЛЮМЭЕ ДПСФАШ БНГКЧАЪ,
УНРЕК АШ Ъ РЕАЕ ОПЕДЯРЮБХРЭ
ГЮКНЦ ДНЯРНИМЕЕ РЕАЪ

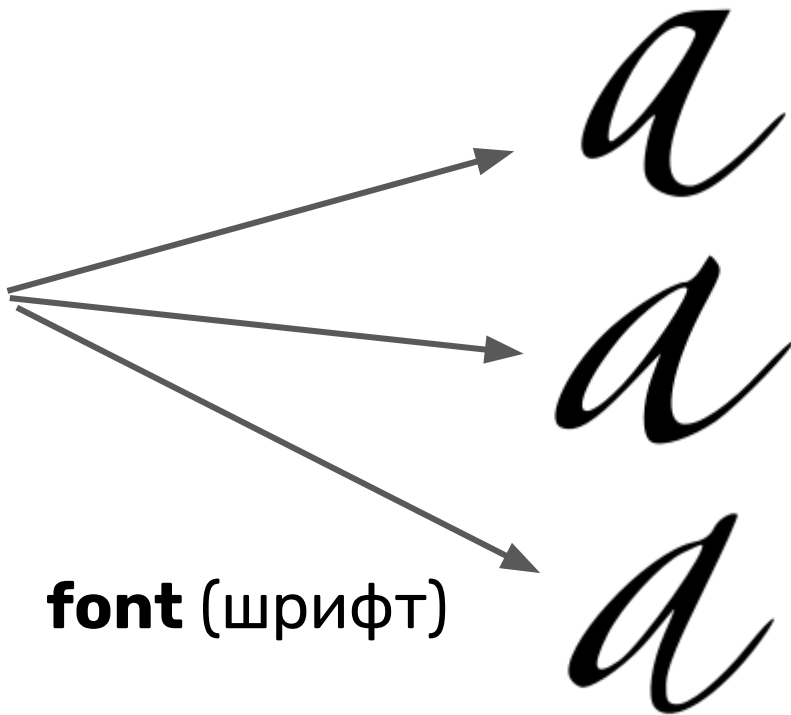
...



'A', 'B', 'C', 'È', 'Í', '췘', ...
character (графема)

font (шрифт)

glyph (глиф)









Unicode

0061 LATIN SMALL LETTER A
0062 LATIN SMALL LETTER B
0063 LATIN SMALL LETTER C
...
007B LEFT CURLY BRACKET
...
2167 ROMAN NUMERAL EIGHT
2168 ROMAN NUMERAL NINE
...
265E BLACK CHESS KNIGHT
265F BLACK CHESS PAWN
...
1F600 GRINNING FACE
1F609 WINKING FACE
...



0061 a
0062 b
0063 c
...
007B {
...
2167 VIII
2168 IX
...
265E 
265F 
...
1F600 
1F609 
...

glyphs



Adopt a Character



Emoji



Basic Info



News

Events

Connect



Membership



Press



U+1F481

Ε

U+0A98

И

U+30A7



U+1F5A4

ᲀ

U+0669

<

U+27E8



U+2042



U+1F3EF

ᱚ

U+0D05

Φ

U+03A6

‘

U+2018

®

U+00AE

尸

U+5C38

3

U+0437

첼

U+CDF5



U+1F604

ᄇ

U+0B9C



U+3006

人

U+4EBA

•

U+2022

—

U+FE3B

Რ

U+0254

ξ

U+0EC2

᳚

U+0534

Everyone in the world should be able to use
their own language on phones and computers.

[▶ LEARN MORE ABOUT UNICODE](#)



[ADOPT A CHARACTER ↗](#)

—

U+203E



U+25C8

᳚

U+179C

Რ

U+0A68

✂

U+30E1

—

U+203F

•

U+0F0B

᳚

U+30FE

/

U+FF0F

땡

U+D1E1

И

U+FF74



U+1F31E

᳚

U+12CE

᳚

U+1795

ඳ

U+0DAA

«

U+00AB

ガ

U+30AC

○

U+26AC

。

U+FF61

᳚

U+0E9F



Примеры использования Unicode

```
$ echo "\U1F5A4"
```



U+1F5A4



```
$ echo "\U1F5A4"
```



```
$ python -c "print('\U0001F5A4')"
```



U+1F5A4



```
101010000000000000000000000000000011010000000000000000000000000000
00011001010000000000000000000000000000100000000000000000000000000000
00000101101000000000000000000000000000001100101000000000000000000000
00000001101110000000000000000000000000000001000000000000000000000000
000000000110111100000000000000000000000000000000000110011000000000000000
00000000000010000000000000000000000000000000000000000001010000000000000000
000000000000000000000000000000000000000000000000000000001101000000000000
0000000000000000000000000000000000000000000000000000000000110111000000
0000000000000000000000000000000000000000000000000000000000001011000000
00000000000000000000000000000000000000000000000000000000000001011000000
00000000000000000000000000000000000000000000000000000000000000000000...
```

*отгадайте, на основе какого текста получены указанные code points
(текст можно получить с помощью одного вызова Python)



Unicode, начало...

+	0	1	2	3	4	5	6	7	8	9
0	?									
10										
20										
30				!	"	#	\$	%	&	'
40	()	*	+	,	-	.	/	0	1
50	2	3	4	5	6	7	8	9	:	;
60	<	=	>	?	@	A	B	C	D	E
70	F	G	H	I	J	K	L	M	N	O
80	P	Q	R	S	T	U	V	W	X	Y
90	Z	[\]	^	_	`	a	b	c
100	d	e	f	g	h	i	j	k	l	m
110	n	o	p	q	r	s	t	u	v	w
120	x	y	z	{		}	~		€	
130	,	f	„	...	†	‡	^	‰	Š	<
140	Œ		Ž		‘	’	“	”		•



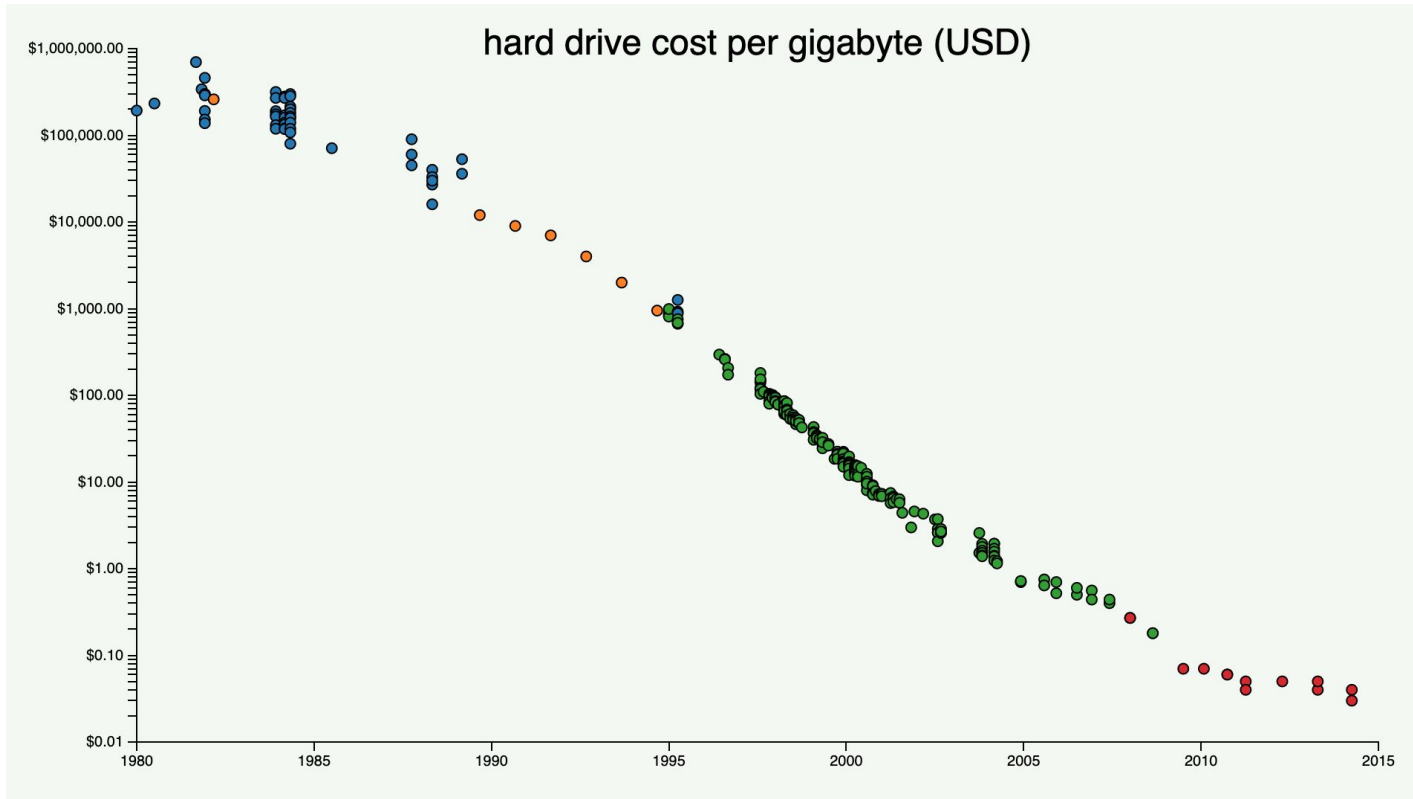


	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0	NUL 0000	SOH 0001	STX 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
1_ 16	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
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	~ 007E	DEL 007F

ASCII = American Standard Code for Information Interchange



Unicode, ascii и плюшки

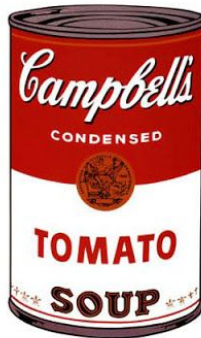
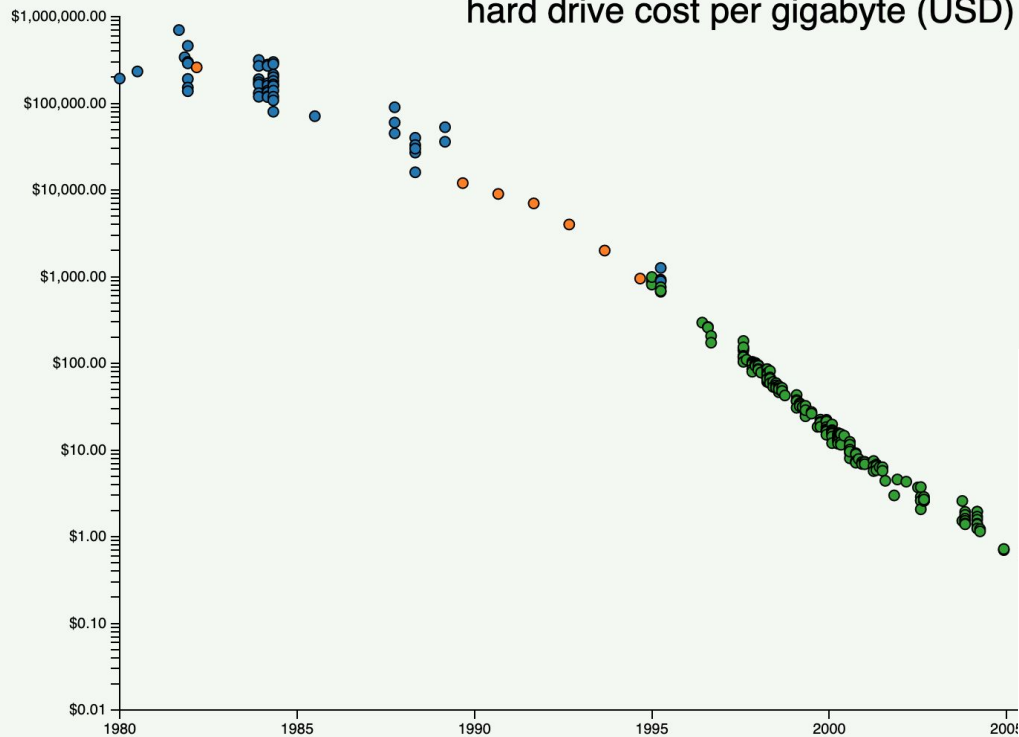


ИСТОЧНИК: <https://mkomo.com/cost-per-gigabyte-update>



Unicode, ascii и плюшки

hard drive cost per gigabyte (USD)



x 1,500



x 300



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	!	"	#	\$	%	&	'	()	*	+	,	-	.	/
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
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	~ 007E	
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	☒ 2594	☒ 2595	☒ 2596	☒ 2597	☒ 2598	☒ 2599	☒ 259A	☒ 259B	☒ 259C	☒ 259D	☒ 259E	☒ 259F	
A_160	= 2550	 2551	F 2552	ѐ 2553	ѓ 2554	ѓ 2555	ѓ 2556	ѓ 2557	ѓ 2558	ѓ 2559	ѓ 255A	ѓ 255B	ѓ 255C	ѓ 255D	ѓ 255E	ѓ 255F
B_176	Ѡ 255F	ѡ 2560	Ѣ 2561	Ѥ 2562	ѥ 2563	Ѧ 2564	ѧ 2565	Ѩ 2566	ѩ 2567	Ѫ 2568	ѫ 2569	Ѭ 256A	ѭ 256B	Ѯ 256C	ѯ 256D	Ѱ 256E
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 0436	ж 0443	в 044C	ь 044B	з 0437	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	З 0417	Ш 0428	Э 042D	Щ 0429	Ч 0427	Ъ 042A

5	0438	0439	043A	043B	043C	043D
	Ы	З	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





**BIGDATA
TEAM**

Резюме



 code point, character и glyph



- ▶ code point, character и glyph
- ▶ code point --(Unicode)--> character --(font)--> glyph



- ▶ code point, character и glyph
- ▶ code point --(Unicode)--> character --(font)--> glyph
- ▶ Unicode != encoding (ascii, koi8-r, cp1251, koi8-r, ...)



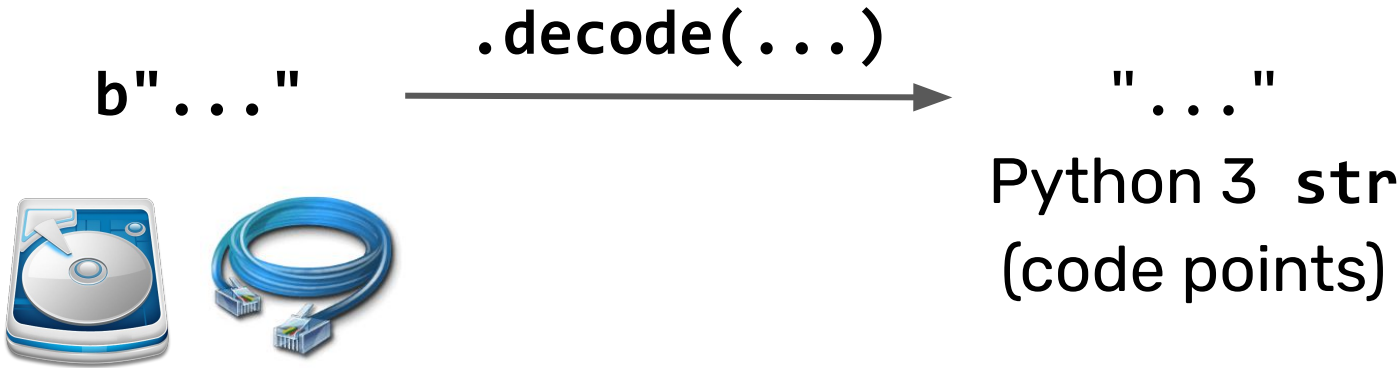
- ▶ code point, character и glyph
- ▶ code point --(Unicode)--> character --(font)--> glyph
- ▶ Unicode != encoding (ascii, koi8-r, cp1251, koi8-r, ...)

b"..."





- ▶ code point, character и glyph
- ▶ code point --(Unicode)--> character --(font)--> glyph
- ▶ Unicode != encoding (ascii, koi8-r, cp1251, koi8-r, ...)





- ▶ code point, character и glyph
- ▶ code point --(Unicode)--> character --(font)--> glyph
- ▶ Unicode != encoding (ascii, koi8-r, cp1251, koi8-r, ...)

