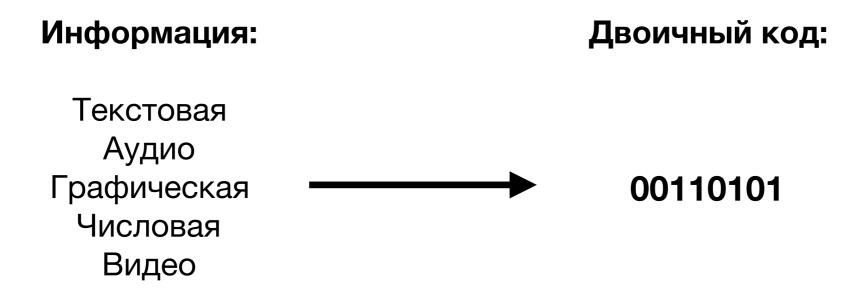
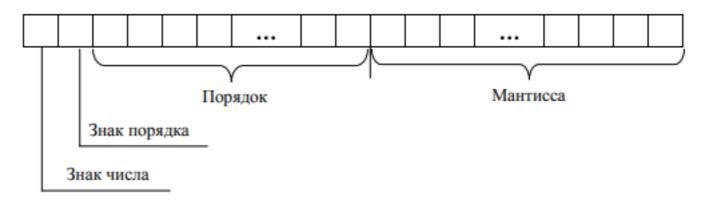
Представление текстовой информации в ЭВМ



Единица информации в компьютере - один бит, т.е. двоичный разряд, который может принимать значение 0 или 1 В одном байте можно закодировать значение одного символа из 256 возможных (2⁸).

1Кбайт = 2^10 1Мбайт = 1024 Кбайт И тд



Представление числа с плавающей точкой в виде набора битов

Число половинной точности (Binary16, Half precision)

	Знак															
		Порядок					Мантисса									
0	0	0	0	0	0	1,	0	0	0	0	0	0	0	0	0	0
	14 10			10		9									0	

Число одинарной точности (Binary32, Single precision, float)



Число двойной точности (Binary64, Double precision, double)

	Знак	
	Порядок	Мантисса
	(11 бит)	(52+1 бит)
0	0 0 0 0 0 0 0 0 0 0 0	1, 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	62 52	51

Число четверной точности (Binary128, Quadruple precision)

Знак	
Порядок	Мантисса
(15 бит)	(112+1 бит)
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	1, 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
126 112	111

ASCII

ASCII (American Standard Code for Informational Interchange — Американский стандартный код информационного обмена).

٦	0	1	2	3	4	5	6	7	8	9	ı A	В	С	D	E	<u> </u>
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	S0	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		-:	-	#	\$	%	&	-	()	*	+	,	-	•	/
3	0	1	2	3	4	5	6	7	8	9		;	٧	=	۸	?
4	0	Α	В	С	D	Ε	F	G	Н	Ι	J	K	Г	М	N	0
5	Р	Q	R	S	T	U	V	W	Х	Υ	Z]	\]	^	_
6	,	а	b	C	d	е	f	g	h	i	j	k	l	m	n	0
7	р	q	r	S	t	u	V	W	Х	у	Z	{		}	~	DEL

1963 год: 1 код = 7 бит. $2^7 = 128$ символов.

Сейчас: 1 код = 8 бит = 1 байт. 2^8 = 256 символов.

ASCII Code Chart

لـ	0	1	2	3	4	5	6	7	8	9	ı A	В	C	D	E	L F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	S0	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2			=	#	\$	%	&	-	()	*	+	,	-	•	/
3	0	1	2	3	4	5	6	7	8	9		;	٧	=	۸	?
4	0	Α	В	C	D	Е	F	G	Н	Ι	J	K	L	М	N	0
5	Р	Q	R	S	T	U	V	W	Х	Υ	Z]	\]	^	_
6	,	а	b	С	d	е	f	g	h	i	j	k	l	m	n	0
7	р	q	r	S	t	u	V	W	Х	у	Z	{		}	~	DEL

HSE: ASCII: 48 53 45 или в двоичной: 010010000101001101000101

Десятичное	Восьмеричное	Двоичная	Шестнадцатеричное	Двоичная
число	число	запись	число	запись
0	0	000	0	0000
1	1	001	1	0001
2	2	010	2	0010
3	3	011	3	0011
4	4	100	4	0100
5	5	101	5	0101
6	6	110	6	0110
7	7	111	7	0111
8	10	000	8	1000
9	11	001	9	1001
10	12	010	A	1010
11	13	011	В	1011
12	14	100	С	1100
13	15	101	D	1101
14	16	110	E	1110
15	17	111	F	1111

В настоящее время существует 5 кодовых таблиц для русских букв: Windows CP1251, MS – DOS CP866, КОИ – 8 (Код обмена информацией, 8-битный) (используется в OS UNIX), Mac (Macintosh), ISO (OS UNIX). Одним из первых стандартов кодирования кириллицы на компьютерах был стандарт **КОИ-8.**

Кодировка КОИ-8R

		0	1	2	3	4	5	6	7	8	9	Α	В	C	D	Ε	F
		2500	2502	250C	2510	2514	2518	251C	2524	252C	2534	253C	2580	2584	2588	2 <u>5</u> 8C	2590
8	30	_		Г	 7	L	–	 -	-	Т	ㅗ	+					
		2591	2592	2593	2320	25A0	2219	221 _A	2248	2264	2265	A0	2321	B0	B2	B7	F7
	90		******				•	√	≈	≤	≥			0	2	•	÷
		2550	2551	2552	451	2553	2554	2555	2556	2557	2558	2559	255A	255B	255C	255D	255E
1	10	=		F	ë	г	F	7	Ιп	la	E	╙	L	╛	Ш	一一	=
		255F	2560	2561	401	2562	2563	2564	2565	2566	2567	2568	2569	256A	256B	256C	A9
E	30	l ⊪	▎▕⊦		Ë			₹	Ι π	l ╦	📥	Ш	业	‡	#	#	©
		44E	430	431	446	434	435	444	433	445	438	439	43A	43B	43C	43D	43E
(00	ю	а	б	Ц	Д	е	ф	Г	х	и	Й	К	Л	М	Н	0
		43F	44F	440	441	442	443	436	432	44C	44B	437	448	44D	449	447	44A
Ĺ	00	П	Я	р	С	Т	У	ж	В	ь	ы	3	Ш	Э	Щ	4	ъ
		42E	410	411	426	414	415	424	413	425	418	419	41A	41B	41C	41D	41E
E	E0	Ю	Α	Б	Ц	Д	E	Φ	Г	X	и	Й	К	Л	M	Н	0
		41F	42F	420	421	422	423	416	412	42C	42B	417	428	42D	429	427	42A
F	FO	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	3	3	ტ	Щ	J	Ъ

Кодировка СР1251

	0	1	2	3	4	5	6	7	8	9	Α	В	С	D	Ε	F
	402	403	201A	453	201E	2026	2020	2021	20AC	2030	409	2039	40A	40C	40B	40F
80	ъ	Ĺ	,	ŕ	,,		†	#	€	%	љ	₹	њ	ΙŔ	Ъ	Ų
	452	2018	2019	201C	201D	2022	2013	2014		2122	459	203A	45A	45C	45B	45F
90	ħ	•	,	66	"	•	_	—		TM	љ	}	њ	Ŕ	ħ	Ų
	Α0	40E	45E	408	Α4	490	A6	A7	401	A9	404	AB	AC	AD	AE	407
A0		Ӱ́	ÿ	J	#			§	Ë	©	ε	«	¬	-	®	Ϊ
	B0	B1	406	456	491	B5	B6	B7	451	2116	454	BB	458	405	455	457
B0	٥	±	- 1	i	۲	μ	¶	•	ë	Nº	ε	»	j	S	S	Ϊ
	410	411	412	413	414	415	416	417	418	419	41A	41B	41C	41D	41E	41F
CO	Α	Б	В	Г	Д	E	Ж	3	и	Й	К	Л	М	H	0	П
	420	421	422	423	424	425	426	427	428	429	42A	42B	42C	42D	42E	42F
D0	Р	С	Т	У	Φ	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	ю	Я
	430	431	432	433	434	435	436	437	438	439	43A	43B	43C	43D	43E	43F
E0	а	б	В	Γ	Д	е	ж	3	и	Й	К	Л	м	н	0	П
	440	441	442	443	444	445	446	447	448	449	44A	44B	44C	44D	44E	44F
F0	р	С	Т	У	ф	х	ц	4	ш	щ	ъ	ы	ь	э	ю	Я

Unicode

0400 Cyrillie 04FF

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F
0	È	A 0410	P	a 0430	p	è	©	Ψ	<u>C</u>	Т 0490	K	¥	I 0400	Ă 04D0	3	$\ddot{\mathbf{y}}$
1	Ë	Б	C	б ₀₄₃₁	C 0441	ë 0451	W 0461	Ψ 0471	Ç	Γ΄ 0491	K	Y	Ж	ă	3 04E1	ÿ
2	T	B	T	B	T	ħ	Ђ 0462	O	*	F	H 04A2	X	Ж 04C2	Ä	Л	Ý
3	Ý	Г	y	Г 0433	y	Ѓ 0453	5	O	ে ৪৪৪	F	H	X ,	Б	ä	Д	ý
4	6 3	Д	Ф 0424	Д 0434	ф	E 0454	К	V 0474	ऽं इ	5	H	T I	5	Æ	Й	Ü
5	S 0405	E 0415	X 0425	e 0435	X 0445	S	Ю	V 0475	_1 8485	5	H	T I	Л 04C5	æ	Ё	ü
6	I 0406	Ж	Ц	Ж	Ц 0446	i 0456	A	V" 0476	ा ₈₄₈₆	Ж	ПЬ	Ч	Л ,	Ĕ	Ö	
7	i	3	4	3	4	ï 0457	A	V 0477		Ж,	П	4	H	ĕ	Ö	
8	J 0408	И	111 0428	И	111 0448	j 0458	IA 0468	Oy	्रें 0488	3	Q 04A8	4 04B8	H	3	O	Ӹ очгв
9	Љ	Й	Щ 0429	Й	Щ 0449	Љ 0459	I-A 0469	O Y	ر 489 م	3	Q	Ч	H 04C9	ə	O	Ӹ
Α	Њ	K	Ъ	K	Ъ	Њ 045A	X	047A	Й	K	Ç	h 04BA	H ,	ä 04DA	Ö	
В	T	Л 041В	Ы	Л 043B	Ы 0448	ħ	X	O47B	Й ,	K ,	Ç	h MABB	4 04CB	Ö 04DB	Ö	
С	Κ́	M 0410	b	M	Ь	Κ́ 0450	X	ن	b	K	T	C	y MCC	Ж 04DC	Э	
D	À 0400	H 041D	Э	H	Э	ѝ	1-X	<u>ن</u> ۵47D	b	K	T	e	M, O4CD	Ж	Ä	
Е	y	O 041E	Ю	O 043E	Ю 044Е	ў 045E	à	©	P 048E	K	Y	Ç O4BE	M ,	З	$ar{\mathbf{y}}$	
F	Ц	11 041F	Я	П 043F	Я	Ц 045F	à	W	p	ҟ	Y	ę O4BF		3 04DF	$ar{y}_{_{\scriptscriptstyle{04EF}}}$	

- Результат сотрудничества Международной организации по стандартизации (ISO) с ведущими производителями компьютеров и программного обеспечения.
- В **Unicode** на кодирование символов отводится 16 бит, т.е. 2^16 = 65 536 кодов

Распространенные стандарты:

- UTF-8
- UTF-16
- UTF-32

One-hot. Унитарный код

Унитарный код - двоичный код фиксированной длины. Длина кода определяется количеством кодируемых объектов, то есть каждому объекту соответствует отдельный разряд кода, а значение кода положением 1 или 0 в кодовом слове.

Размер вектора = размеру словаря

Cat =
$$[1,0,0,...,0]$$

$$Dog = [0,1,0,...,0]$$

Bird =
$$[0,0,0,...,1]$$

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50