Информатика

КОДИРОВАНИЕ ТЕКСТОВОЙ ИНФОРМАЦИИ

ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ В КОМПЬЮТЕРЕ

10 класс

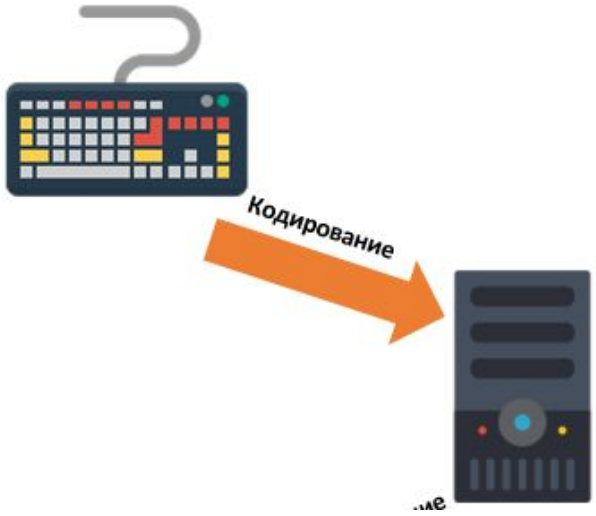
ИЗДАТЕЛЬСТВО БИНОМ

Ключевые слова

- текстовая информация
- кодирование
- кодовые таблицы

Как происходит кодирование и декодирование текстовой информации?

Текстовая информация — это информация, представленная в форме письменного текста. Для того, чтобы компьютер смог работать с такой информацией, ее надо закодировать. Как происходит кодирование и декодирование текстовой информации?



Компьютерное представление текстовой информации

Для компьютерного представления текстовой информации достаточно:

...	...
64	01000000
65	01000001
66	01000010
67	01000011
68	01000100

Определить алфавит (множество всех символов)

Присвоить каждому символу алфавита порядковый номер

Перевести номер символа в двоичную систему счисления

Как вы знаете, вся информация в компьютере хранится в двоичном коде.

Поэтому надо научиться преобразовывать символы в **двоичный код**.

Нам хорошо известна формула Хартли, определяющая количество информации в зависимости от количества возможных вариантов.

Ральф Винтон Лайон Хартли
(30 ноября 1888 — 1 мая 1970)

Если же мы преобразуем эту формулу и примем за **N** — количество символов в используемом алфавите

то мы поймем сколько памяти потребуется для кодирования одного символа.

$$N=2^i$$

N — количество возможных вариантов

i — количеств бит,потребуемых для кодирования

N = 32 → i = 5 бит

А	00000	Б	01000	В	10000	Г	11000
Д	00001	Е	01001	Ж	10001	З	11001
И	00010	К	01010	Л	10010	М	11010
Н	00011	О	01011	П	10011	Р	11011
С	00100	Т	01100	У	10100	Ф	11100
Х	00101	Ц	01101	Ч	10101	Ш	11101
Щ	00110	Ы	01110	Э	10110	Ю	11110
Я	00111	пробел	01111		10111		1111

Первая широко используемая кодировочная таблица была создана в США и называлась **ASCII.**

Что в переводе означало **American standard code for information interchange.**

Кодировочная таблица ASCII

двоичный код	символ	двоичный код	символ	двоичный код	символ	двоичный код	символ
000 0000	[NUL]	010 0000	space	100 0000	@	110 0000	`
000 0001	[SOH]	010 0001	!	100 0001	A	110 0001	a
000 0010	[STX]	010 0010	"	100 0010	B	110 0010	b
000 0011	[ETX]	010 0011	#	100 0011	C	110 0011	c
000 0100	[EOT]	010 0100	$	100 0100	D	110 0100	d
000 0101	[ENQ]	010 0101	%	100 0101	E	110 0101	e
000 0110	[ACK]	010 0110	&	100 0110	F	110 0110	f
000 0111	[BEL]	010 0111	'	100 0111	G	110 0111	g
000 1000	[BS]	010 1000	(	100 1000	H	110 1000	h
000 1001	[TAB]	010 1001	)	100 1001	I	110 1001	i
000 1010	[LF]	010 1010	*	100 1010	J	110 1010	j
000 1011	[VT]	010 1011	+	100 1011	K	110 1011	k
000 1100	[FF]	010 1100	,	100 1100	L	110 1100	l
000 1101	[CR]	010 1101	-	100 1101	M	110 1101	m
000 1110	[SO]	010 1110	.	100 1110	N	110 1110	n
000 1111	[SI]	010 1111	/	100 1111	O	110 1111	o
001 0000	[DLE]	011 0000	0	101 0000	P	111 0000	p
001 0001	[DC1]	011 0001	1	101 0001	Q	111 0001	q
001 0010	[DC2]	011 0010	2	101 0010	R	111 0010	r
001 0011	[DC3]	011 0011	3	101 0011	S	111 0011	s
001 0100	[DC4]	011 0100	4	101 0100	T	111 0100	t
001 0101	[NAK]	011 0101	5	101 0101	U	111 0101	u
001 0110	[SYN]	011 0110	6	101 0110	V	111 0110	v
001 0111	[ETB]	011 0111	7	101 0111	W	111 0111	w
001 1000	[CAN]	011 1000	8	101 1000	X	111 1000	x
001 1001	[EM]	011 1001	9	101 1001	Y	111 1001	y
001 1010	[SUB]	011 1010	:	101 1010	Z	111 1010	z
001 1011	[ESC]	011 1011	;	101 1011	[	111 1011	{
001 1100	[FS]	011 1100	<	101 1100	\	111 1100	|
001 1101	[GS]	011 1101	=	101 1101	]	111 1101	}
001 1110	[RS]	011 1110	>	101 1110	^	111 1110	~
001 1111	[US]	011 1111	?	101 1111	_	111 1111	[DEL]

Кодировка ASCII

American Standard Code for Information Interchange – американский стандартный код для обмена информацией, разработанный в 1960-х годах в США.

Расширение кодировки ASCII

Расширение кодировки ASCII

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	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	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

Windows-1251 **КОИ-8**

	Ъ	Ѓ	,	ѓ	„	…	†	‡	€	‰	Љ	‹	Њ	Ќ	Ћ	Џ
9	ђ	‘	’	“	”	•	–	—		™	љ	›	њ	ќ	ћ	џ
A		Ў	ў	Ј	¤	Ґ	¦	§	Ё	©	Є	«	¬		®	Ї
B	°	±	І	і	ґ	µ	¶	·	ё	№	є	»	ј	Ѕ	ѕ	ї
C	А ю	Б а	В б	Г ц	Д д	Е е	Ж ф	З г	И х	Й и	К й	Л к	М л	Н м	О н	П о
D	Р п	С я	Т р	У с	Ф т	Х у	Ц ж	Ч в	Ш ь	Щ ы	Ъ з	Ы ш	Ь э	Э щ	Ю ч	Я ъ
E	а Ю	б А	в Б	г Ц	д Д	е Е	ж Ф	з Г	и Х	й И	к Й	л К	м Л	н М	о Н	п О
F	р П	с Я	т Р	у С	ф Т	х У	ц Ж	ч В	ш Ь	щ Ы	ъ З	ы Ш	ь Э	э Щ	ю Ч	я Ъ

Но так как этого количества было недостаточно, стали создаваться другие таблицы, в которых можно было закодировать и другие символы.

─ 128	│ 129	┌ 130	┐ 131	└ 132	┘ 133	├ 134	┤ 135	┬ 136	┴ 137	┼ 138	▀ 139	▄ 140	▌ 141	▐ 142	▐ 143
▒ 144	▒ 145	▓ 146	147	■ 148	• 149	√ 150	≈ 151	≤ 152	≥ 153	nbsp 154	155	° 156	² 157	· 158	÷ 159
= 160	‖ 161	┌ 162	е 163	┌ 164	┌ 165	┐ 166	┐ 167	┐ 168	└ 169	└ 170	└ 171	┘ 172	┘ 173	┘ 174	├ 175
├ 176	├ 177	┤ 178	Ё 179	┤ 180	┤ 181	┬ 182	┬ 183	┬ 184	┴ 185	┴ 186	┴ 187	┼ 188	┼ 189	┼ 190	© 191
ю 192	а 193	б 194	ц 195	д 196	е 197	ф 198	г 199	х 200	и 201	й 202	к 203	л 204	м 205	н 206	о 207
п 208	я 209	р 210	с 211	т 212	у 213	ж 214	в 215	ь 216	ы 217	з 218	ш 219	э 220	щ 221	ч 222	ъ 223
Ю 224	А 225	Б 226	Ц 227	Д 228	Е 229	Ф 230	Г 231	Х 232	И 233	Й 234	К 235	Л 236	М 237	Н 238	О 239
П 240	Я 241	Р 242	С 243	Т 244	У 245	Ж 246	В 247	Ь 248	Ы 249	З 250	Ш 251	Э 252	Щ 253	Ч 254	Ъ 255

Проблемой использования таких различных таблиц приводила к тому, что текст, написанный на одном компьютере, мог некорректно читаться на другом.

Была разработана международная таблица кодировки **Unicode.**

Включающая в себя как символы английского, русского, немецкого, арабского и других языков.

На каждый символ в такой таблице отводится **16 бит**.

А значит она позволяет кодировать **65536 символов**.

Однако, использование такой таблицы сильно «утяжеляет» текст. Поэтому существуют различные алгоритмы неравномерной кодировки текста, например, **алгоритм Хаффмана**.

Стандарт Unicode

Unicode — это «уникальный код для любого символа, независимо от платформы, независимо от программы, независимо от языка» (www.unicode.org).

Стандарт Unicode был разработан в 1991 году и описывает алфавиты всех известных, в том числе и «мертвых», языков. Для языков, имеющих несколько алфавитов или вариантов написания (японского и индийского), закодированы все варианты.
В кодировку Unicode внесены все математические и иные научные символьные обозначения и даже некоторые придуманные языки (язык эльфов из трилогии Дж. Р. Р. Толкина «Властелин колец»).

Клавиатуры некоторых стран мира

РУССКАЯ

АМЕРИКАНСКАЯ

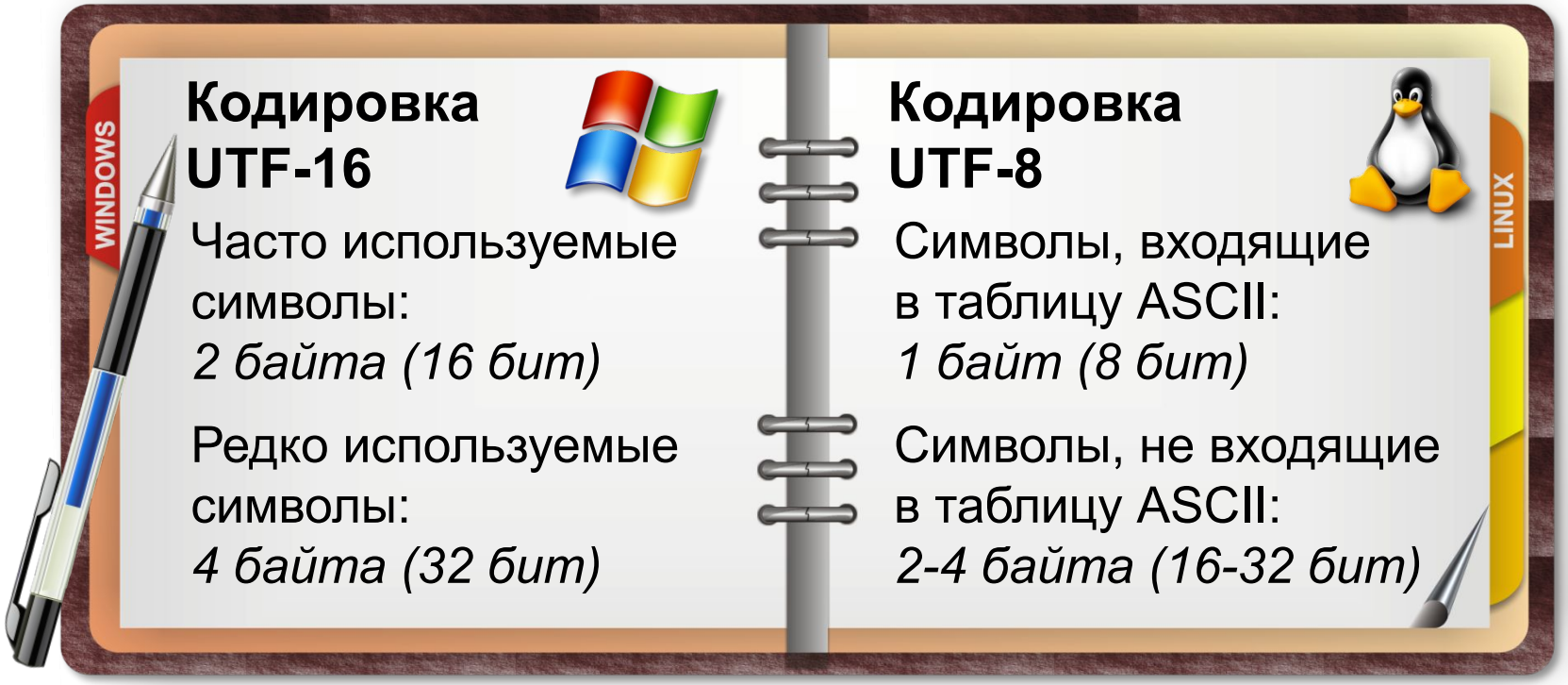
АРАБСКАЯ

АРМЯНСКАЯ

ЯПОНСКАЯ

Кодировки стандарта Unicode

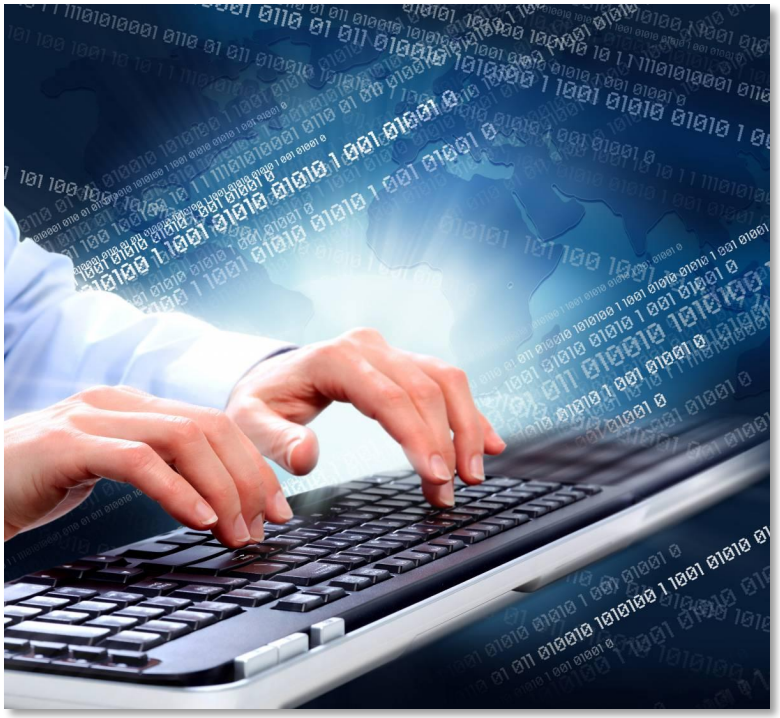
Для представления символов в памяти компьютера в стандарте Unicode имеется несколько кодировок.



Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

Информационный объем сообщения

Информационным объёмом текстового сообщения называется количество бит (байт, килобайт, мегабайт и т. д.), необходимых для записи этого сообщения путём заранее оговоренного способа двоичного кодирования.



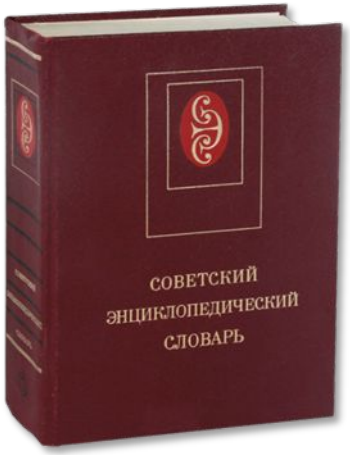
Количество символов в сообщении

$$I = K \cdot i$$

ASCII, КОИ-8, Windows-1251, …
1 символ = 1 байт

Unicode
1 символ = 2 байта

Вопросы и задания



В Советском энциклопедическом словаре (1983 года издания) 1600 страниц. На одной странице размещается в среднем 100 строк по 140 символов (включая пробелы) в каждой. Найдите объем (в Мбайтах) текстовой информации в словаре, если при записи используется кодировка «*один символ — один байт*».

Дано:

i = 1 байт

K = 1600·100·140

I - ?

I = K·i

$$I = \frac{1600 \cdot 100 \cdot 140}{1024 \cdot 1024} \text{ Мб} \approx 21{,}36 \text{ Мб}$$

Ответ: 21,36 Мбайта

Самое главное

Текстовая информация по своей природе дискретна, так как представляется последовательностью отдельных символов.

В памяти компьютера хранятся специальные кодовые таблицы, в которых для каждого символа указан его двоичный код. Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

Основой для компьютерных стандартов кодирования символов послужил код ASCII, рассчитанный на передачу только английского текста. Расширения ASCII-кодировки, в которых первые 128 символов кодовой таблицы совпадают с кодировкой ASCII, а остальные (с 128-го по 255-й) используются для кодирования букв национального алфавита, символов национальной валюты и т. п.

Самое главное

В 1991 году был разработан новый стандарт кодирования символов, получивший название Unicode (Юникод), позволяющий использовать в текстах любые символы любых языков мира. Кодировки Unicode позволяют включать в один документ символы самых разных языков, но их использование ведёт к увеличению размеров текстовых файлов.

Символ	10й код	2й код	Символ	10й код	2й код	Символ	10й код	2й код	Символ	10й код	2й код
	32	00100000	8	56	00111000	P	80	01010000	h	104	01101000
!	33	00100001	9	57	00111001	Q	81	01010001	i	105	01101001
"	34	00100010	:	58	00111010	R	82	01010010	j	106	01101010
#	35	00100011	;	59	00111011	S	83	01010011	k	107	01101011
$	36	00100100	<	60	00111100	T	84	01010100	l	108	01101100
%	37	00100101	=	61	00111101	U	85	01010101	m	109	01101101
&	38	00100110	>	62	00111110	V	86	01010110	n	110	01101110
'	39	00100111	?	63	00111111	W	87	01010111	o	111	01101111
(	40	00101000	@	64	01000000	X	88	01011000	p	112	01110000
)	41	00101001	A	65	01000001	Y	89	01011001	q	113	01110001
*	42	00101010	B	66	01000010	Z	90	01011010	r	114	01110010
+	43	00101011	C	67	01000011	[	91	01011011	s	115	01110011
,	44	00101100	D	68	01000100	\	92	01011100	t	116	01110100
-	45	00101101	E	69	01000101	]	93	01011101	u	117	01110101
.	46	00101110	F	70	01000110	^	94	01011110	v	118	01110110
/	47	00101111	G	71	01000111	_	95	01011111	w	119	01110111
0	48	00110000	H	72	01001000	`	96	01100000	x	120	01111000
1	49	00110001	I	73	01001001	a	97	01100001	y	121	01111001
2	50	00110010	J	74	01001010	b	98	01100010	z	122	01111010
3	51	00110011	K	75	01001011	c	99	01100011	{	123	01111011
4	52	00110100	L	76	01001100	d	100	01100100	|	124	01111100
5	53	00110101	M	77	01001101	e	101	01100101	}	125	01111101
6	54	00110110	N	78	01001110	f	102	01100110	~	126	01111110
7	55	00110111	O	79	01001111	g	103	01100111	□	127	01111111

Вопросы и задания

Задание 1. Представьте в кодировке ASCII текст

Happy New Year!

а) шестнадцатеричным кодом

48 61 70 70 79 20 4E 65 77 20 59 65 61 72 21

б) десятичным кодом

72 97 112 112 121 32 78 101 119 32 89 101 97 114 33

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	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	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

ОТВЕТ

Вопросы и задания

Задание 3. В 15-м издании энциклопедии Britannica 32 тома, в каждом из которых порядка 1000 страниц. На одной странице размещается в среднем 70 строк по 120 символов (включая пробелы) в каждой. Найдите объем текстовой информации в энциклопедии, если при записи используется кодировка Unicode («*один символ — два байта*»).

Дано:

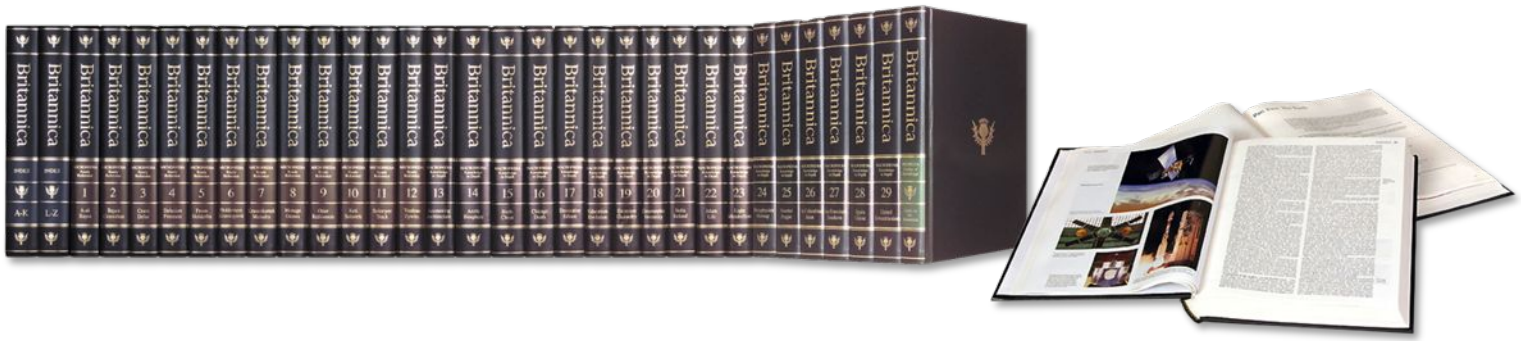
i = 2 байта

K = 32·1000·70·120

I = K·i

$$I = \frac{32 \cdot 1000 \cdot 70 \cdot 120 \cdot 2}{1024 \cdot 1024} \text{ Мб} \approx 513 \text{ Мб}$$

Ответ: 513 Мбайт



Практическая работа

Стр. 206

№ 6 устно

№ 7 практическая работа

№ 2,3 самостоятельная работа

Домашняя работа

Прочитать п. 6 стр. 43-45, выполнить № 4 стр. 206