Информатика. Упражнение 1 Представление текстовой информации в кодировках *KOI-8, windows-1251* и *UTF-8*

Цель работы: изучить наиболее часто встречающиеся кодировки текстов.

Последовательность выполнения упражнения 1

1. Ознакомьтесь с таблицами кодировок ASCII (коды 0 - 127), KOI-8R (коды 160 - 255), windows-1251 (коды 128 - 255) и UTF-8 (кириллица).
2. Получите у преподавателя номер варианта текста для кодирования. Варианты приведены в табл. 1.
3. Закодируйте полученный текст последовательно в кодировках *KOI-8, windows-1251* и *UTF-8* так, как показано в примере 1.

Табл. 1. Варианты заданий

№ варианта	Текст для кодирования
1	Pascal - язык для обучения начинающих
2	Автор Pascal - Никлаус Вирт(Niklaus Wirth)
3	Какой язык сложнее - C++ или Java?
4	Чем отличается язык C++ от C#?
5	FORTRAN - язык для математических задач
6	PERL - язык для генерации отчётов
7	Автор языка PERL - Ларри Уолл (1987 г.)
8	PHP - язык для динамических страниц (1996)
9	JavaScript - язык для активных страниц
10	Python (питон, 1991) - возврат к прошлому?
11	ЭВМ, ЦВМ, АВМ, ПЭВМ, computer, ipad
12	SQL -структурированный язык запросов
13	entity-relationship - сущность-связь
14	HTML - основа всемирной паутины (WWW)
15	Автор HTML - Тим Бернерс-Ли (1989 г.)
16	TCP/IP, HTTP, FTP - протоколы

	internet
17	Почтовые протоколы: POP 3, IMAP, SMTP
18	MySQL, Oracle, DBI - реляционные СУБД
19	Label - метка, этикетка, меченый атом
20	Top - верх, первое место, topless - ?
21	SELECT - главный оператор языка SQL
22	INTRANET - локальная сеть интернет
23	XML — расширяемый язык разметки;
24	Paris -Париж, London - Лондон, Roma - ?
25	Что больше - 5! * 5! или 6! * 4! +100?
26	СУБД ACCESS -сетевая или локальная?
27	Photoshop - пакет для растровой графики
28	CORELDRW - пакет для векторной графики
29	3D Studio MAX - пакет трёхмерной графики
30	MatLab - пакет программ для математиков

Пример 1

Нужно закодировать строку *write - писать (англ.)*.

Результат

koi8

```
w r i t e   -   п и с а т ь   ( а н г л . )
7 7 6 7 6 2 2 2 d c d c d d 2 2 c c c c 2 2
7 2 9 4 5 0 d 0 0 9 3 1 4 8 0 8 1 e 7 c e 9
```

cp1251

```
w r i t e   -   п и с а т ь   ( а н г л . )
7 7 6 7 6 2 2 2 e e f e f f 2 2 e e e e 2 2
7 2 9 4 5 0 d 0 f 8 1 0 2 c 0 8 0 d 3 b e 9
```

utf-8

```
w r i t e   -   п   и   с   а   т   ь   (   а   н   г   л   .   )
7 7 6 7 6 2 2 2 d b d b d 8 d b d 8 d 8 2 2 d b d b d b d b 2 2
```

7	2	9	4	5	0	d	0	0	f	0	8	1	1	0	0	1	2	1	c	0	8	0	0	0	d	0	3	0	b	e	9

Обратите внимание на то, что слово *write*, пробел, скобки и точка кодируются одинаково во всех трёх кодировках.

Кодировки *KOI-8R, windows-1251* и *UTF-8* состоят из двух частей. В качестве первой части всех перечисленных кодировкок используется кодировка ASCII (коды 0 - 127), служащая для представления латиницы,цифр и специальных знаков,. Вторые части однобайтных кодировок KOI-8R и windows-1251 содержат коды (128 - 255) кириллицы и ряда специальных символов.

Кодировка UTF-8 - многобайтная. Предусмотрена длина кода одного символа от одного до шести байт. На практике используются коды длиной до четырёх байт. Русские буквы (кириллица) представляются 16-битными (двухбайтными) кодами:

110XXXXX 10XXXXXX,

где X обозначены двоичные разряды для размещения кода символа в соответствии с таблицей *UNICODE*.

Юникод (англ. Unicode) — стандарт кодирования символов, позволяющий представить знаки почти всех письменных языков. Представляемые в юникоде символы кодируются целыми числами без знака. Эти числа будем называть кодами символов в юникоде или просто *UNICODE*. Юникод имеет несколько форм представления символов в компьютере: *UTF-8, UTF-16 (UTF-16BE, UTF-16LE) и UTF-32 (UTF-32BE, UTF-32LE)*. (Англ. Unicode transformation format - UTF).

Рассмотрим, как кодируется в *UTF-8* буква *Ж*. Её *UNICODE* - 1046_{10} или 0416_{16} или 10000 010110_2. *UNICODE* в двоичном виде разбивается на две части: пять левых бит и шесть правых. Левая часть дополняется до байта признаком **110** двухбайтного кода *UTF-8*: **110**10000. К правой части приписываются два бита **10** признака продолжения многобайтного кода: **10**010110. Окончательно код буквы *Ж* в *UTF-8* выглядит так:

11010000 **10**010110_2
или D0 96_{16}

Таким образом, русская буква кодируется дважды: сначала в 11-битный *UNICODE*, а затем - в 16-битный UTF-8.

Рассмотрим, как отличить в закодированном в UTF-8 тексте однобайтные коды от двухбайтных. Представим часть текста

ь (а,

содержащую двухбайтные коды русских букв *ь* и *а* и заключённые между ними однобайтные коды пробела и открывающей скобки в шестнадцатиричном и двоичном коде (табл. 2). Первый байт букв *ь* и *а* начинается признаком первого байта двухбайтного кода **110**. В начале второго байта двухбайтного кода стоит признак продолжения кода **10**. Все однобайтные коды начинаюся битом **0**.

Табл. 2. Отличия однобайтных кодов от двухбайтных				
текст	ь	пробел	)	а
Шестн. код	d1 8c	20	28	d0 b0
Двоич. код	**110**10001 **10**001100	**0**0100000	**0**0101000	**110**10000 **10**110000

Таблица ASCII
(коды 0 - 127 дес. или 0 - 7F шестн.)

Код		Сим-вол	Код		Сим-вол	Код		Сим-вол	Код		Сим-вол
DEC	**HEX**		**DEC**	**HEX**		**DEC**	**HEX**		**DEC**	**HEX**	
0	0	NUL	32	20	пробел	64	40	@	96	60	`
1	1	SOH	33	21	!	65	41	A	97	61	a
2	2	STX	34	22	«	66	42	B	98	62	b
3	3	ETX	35	23	#	67	43	C	99	63	c
4	4	EOT	36	24	$	68	44	D	100	64	d
5	5	ENQ	37	25	%	69	45	E	101	65	e
6	6	ACK	38	26	&	70	46	F	102	66	f
7	7	BEL	39	27	‘	71	47	G	103	67	g
8	8	BS	40	28	(	72	48	H	104	68	h
9	9	TAB	41	29	)	73	49	I	105	69	i
10	A	LF	42	2A	*	74	4A	J	106	6A	j
11	B	VT	43	2B	+	75	4B	K	107	6B	k
12	C	FF	44	2C	,	76	4C	L	108	6C	l
13	D	CR	45	2D	-	77	4D	M	109	6D	m
14	E	SO	46	2E	.	78	4E	N	110	6E	n
15	F	SI	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	48	30	0	80	50	P	112	70	p
17	11	DC1	49	31	1	81	51	Q	113	71	q
18	12	DC2	50	32	2	82	52	R	114	72	r
19	13	DC3	51	33	3	83	53	S	115	73	s

20	14	DC4	52	34	4	84	54	T	116	74	t
21	15	NAK	53	35	5	85	55	U	117	75	u
22	16	SYN	54	36	6	86	56	V	118	76	v
23	17	ETB	55	37	7	87	57	W	119	77	w
24	18	CAN	56	38	8	88	58	X	120	78	x
25	19	EM	57	39	9	89	59	Y	121	79	y
26	1A	SUB	58	3A	:	90	5A	Z	122	7A	z
27	1B	ESC	59	3B	"	91	5B	[	123	7B	{
28	1C	FS	60	3C	<	92	5C	\	124	7C	|
29	1D	GS	61	3D	=	93	5D	]	125	7D	}
30	1E	RS	62	3E	>	94	5E	^	126	7E	~
31	1F	US	63	3F	?	95	5F	_	127	7F	DEL

Таблица KOI-8
(коды 160 - 255 дес. или A0 - FF шестн.)

Первая цифра кода	**Вторая цифра кода**															
	. 0	**. 1**	**. 2**	**. 3**	**. 4**	**. 5**	**. 6**	**. 7**	**. 8**	**. 9**	**. A**	**. B**	**. C**	**. D**	**. E**	**. F**
A.	═	║	╒	ё	╓	╔	╕	╖	╗	╘	╙	╚	╛	╜	╝	╞
B.	╟	╠	╡	Ё	╢	╣	╤	╥	╦	╧	╨	╩	╪	╫	╬	©
C.	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D.	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ
E.	Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F.	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ

Таблица Windows-1251 (коды 128 - 255)

Windows-1251 (cp1251) — это стандартная 8-битная кодировка, разработанная компанией Microsoft. Она содержит практически все символы, которые Вы можете встретить на стандартной русской клавиатуре. Символы с кодами с 0 по 127 дес. такие же, как в первой половине таблицы ASCII.

DEC	HEX	СИМВ	DEC	HEX	СИМВ	DEC	HEX	СИМВ
128	80	Ђ	171	AB	«	214	D6	Ц
129	81	Ѓ	172	AC	¬	215	D7	Ч
130	82	‚	173	AD		216	D8	Ш
131	83	ѓ	174	AE	®	217	D9	Щ
132	84	„	175	AF	Ї	218	DA	Ъ
133	85	…	176	B0	°	219	DB	Ы
134	86	†	177	B1	±	220	DC	Ь
135	87	‡	178	B2	І	221	DD	Э
136	88	€	179	B3	і	222	DE	Ю
137	89	‰	180	B4	ґ	223	DF	Я
138	8A	Љ	181	B5	µ	224	E0	а
139	8B	‹	182	B6	¶	225	E1	б
140	8C	Њ	183	B7	·	226	E2	в
141	8D	Ќ	184	B8	Ё	227	E3	г
142	8E	Ћ	185	B9	№	228	E4	д
143	8F	Џ	186	BA	Є	229	E5	е
144	90	ђ	187	BB	»	230	E6	ж
145	91	‘	188	BC	ј	231	E7	з
146	92	’	189	BD	S	232	E8	и
147	93	“	190	BE	S	233	E9	й
148	94	”	191	BF	ї	234	EA	к
149	95	•	192	C0	А	235	EB	л
150	96	–	193	C1	Б	236	EC	м
151	97	—	194	C2	В	237	ED	н
152	98		195	C3	Г	238	EE	о
153	99	™	196	C4	Д	239	EF	п
154	9A	љ	197	C5	Е	240	F0	р
155	9B	›	198	C6	Ж	241	F1	с
156	9C	њ	199	C7	З	242	F2	т
157	9D	ќ	200	C8	И	243	F3	у
158	9E	ћ	201	C9	Й	244	F4	ф
159	9F	џ	202	CA	К	245	F5	х
160	A0		203	CB	Л	246	F6	ц
161	A1	Ў	204	CC	М	247	F7	ч
162	A2	ў	205	CD	Н	248	F8	ш
163	A3	Ј	206	CE	О	249	F9	щ
164	A4	¤	207	CF	П	250	FA	ъ
165	A5	Ґ	208	D0	Р	251	FB	ы
166	A6	¦	209	D1	С	252	FC	ь
167	A7	§	210	D2	Т	253	FD	э
168	A8	Ё	211	D3	У	254	FE	ю
169	A9	©	212	D4	Ф	255	FF	я
170	AA	Є	213	D5	Х			

Представление кириллицы в UTF-8

В кодировке UTF-8 унаследованы однобайтные (точнее, 7-битные) коды символов ASCII-7 (коды от 0 до 127), т.е. одним байтом кодируются латинские буквы, цифры и специальные символы. Русские буквы (кириллица) представляются 16-битными (двухбайтными) кодами:
110XXXXX 10XXXXXX,
где X обозначены двоичные разряды для размещения кода символа в соответствии с таблицей UNICODE.
Юникод (англ. Unicode) — стандарт кодирования символов, позволяющий представить знаки почти всех письменных языков. Представляемые в юникоде символы кодируются целыми числами без знака. Эти числа будем называть кодами символов в юникоде или просто UNICODE. Юникод имеет несколько форм представления символов в компьютере: UTF-8, UTF-16 (UTF-16BE, UTF-16LE) и UTF-32 (UTF-32BE, UTF-32LE). (Англ. Unicode transformation format - UTF).
Рассмотрим, как кодируется в UTF-8 буква Ж. Её UNICODE - 1046_{10} или 0416_{16} или 10000 010110_2. UNICODE в двоичном виде разбивается на две части: пять левых бит и шесть правых. Левая часть дополняется до байта признаком 110 двухбайтного кода UTF-8: 11010000. К правой части приписываются два бита 10 признака продолжения многобайтного кода: 10010110. Окончательно код буквы Ж в UTF-8 выглядит так:
11010000 10010110_2
или $D0\ 96_{16}$
Таким образом, русская буква кодируется дважды: сначала в 11-битный UNICODE, а затем - в 16-битный UTF-8.
В приведённой ниже таблице, кроме кодов UNICODE и UTF-8 в шестнадцатиричной системе счисления, даны коды UTF-8 в десятичной системе счисления и для сравнения коды кириллицы в кодировке CP-1251, иначе называемой windovs-1251.

Таблица кодов кириллицы в UTF-8

Символ	**UNICODE**		**UTF-8**		**CP-1251**
	Шестн.	**Десят**	**Шестн.**	**Десят**	
А	0410	1040	D090	208 144	192
Б	0411	1041	D091	208 145	193
В	0412	1042	D092	208 146	194
Г	0413	1043	D093	208 147	195
Д	0414	1044	D094	208 148	196
Е	0415	1045	D095	208 149	197

Ж	0416	1046	D096	208 150	198
З	0417	1047	D097	208 151	199
И	0418	1048	D098	208 152	200
Й	0419	1049	D099	208 153	201
К	041A	1050	D09A	208 154	202
Л	041B	1051	D09B	208 155	203
М	041C	1052	D09C	208 156	204
Н	041D	1053	D09D	208 157	205
О	041E	1054	D09E	208 158	206
П	041F	1055	D09F	208 159	207
Р	0420	1056	D0A0	208 160	208
С	0421	1057	D0A1	208 161	209
Т	0422	1058	D0A2	208 162	210
У	0423	1059	D0A3	208 163	211
Ф	0424	1060	D0A4	208 164	212
Х	0425	1061	D0A5	208 165	213
Ц	0426	1062	D0A6	208 166	214
Ч	0427	1063	D0A7	208 167	215
Ш	0428	1064	D0A8	208 168	216
Щ	0429	1065	D0A9	208	217

				169	
Ъ	042A	1066	D0AA	208 170	218
Ы	042B	1067	D0AB	208 171	219
Ь	042C	1068	D0AC	208 172	220
Э	042D	1069	D0AD	208 173	221
Ю	042E	1070	D0AE	208 174	222
Я	042F	1071	D0AF	208 175	223
а	0430	1072	D0B0	208 176	224
б	0431	1073	D0B1	208 177	225
в	0432	1074	D0B2	208 178	226
г	0433	1075	D0B3	208 179	227
д	0434	1076	D0B4	208 180	228
е	0435	1077	D0B5	208 181	229
ж	0436	1078	D0B6	208 182	230
з	0437	1079	D0B7	208 183	231
и	0438	1080	D0B8	208 184	232
й	0439	1081	D0B9	208 185	233
к	043A	1082	D0BA	208 186	234
л	043B	1083	D0BB	208 187	235
м	043C	1084	D0BC	208 188	236

н	043D	1085	D0BD	208 189	237
о	043E	1086	D0BE	208 190	238
п	043F	1087	D0BF	208 191	239
р	0440	1088	D180	209 128	240
с	0441	1089	D181	209 129	241
т	0442	1090	D182	209 130	242
у	0443	1091	D183	209 131	243
ф	0444	1092	D184	209 132	244
х	0445	1093	D185	209 133	245
ц	0446	1094	D186	209 134	246
ч	0447	1095	D187	209 135	247
ш	0448	1096	D188	209 136	248
щ	0449	1097	D189	209 137	249
ъ	044A	1098	D18A	209 138	250
ы	044B	1099	D18B	209 139	251
ь	044C	1100	D18C	209 140	252
э	044D	1101	D18D	209 141	253
ю	044E	1102	D18E	209 142	254
я	044F	1103	D18F	209 143	255
Символы вне общего правила					

Ë	0401	1025	D001	208 101	168
ë	0451	1025	D191	209 145	184