Практическая работа

Тема: Кодирование текстовой информации

Цели работы: познакомиться со способами кодирования и декодирования текстовой информации с помощью кодовых таблиц и компьютера; научиться определять числовые коды символов, вводить символы с помощью числовых кодов и осуществлять перекодировку русскоязычного текста в текстовом редакторе.

I. Теоретическая часть

Начиная с конца 60-х годов компьютеры все больше стали использоваться для обработки текстовой информации, и в настоящее время большая часть персональных компьютеров в мире значительную часть времени занято обработкой именно ТЕКСТОВОЙ информации.

Для представления текстовой информации обычно используется 256 различных символов (прописные и заглавные буквы русского и латинского алфавита, цифры, знаки, графические символы и т. д.). Поставим вопрос: “Какое количество бит информации или двоичных разрядов необходимо, чтобы закодировать 256 различных символов?”

256 различных символов можно рассматривать как 256 различных состояний (событий). В соответствии с вероятностным подходом к измерению количества информации необходимое количество информации для двоичного кодирования 2 56 символов равно:

$I = \log_2 256 = 8$ бит = 1 байт

Следовательно, для двоичного кодирования 1 символа необходим 1 байт информации или 8 двоичных разрядов. Таким образом, каждому символу соответствует своя уникальная последовательность из восьми нулей и единиц.

Присвоение символу конкретного двоичного кода — это вопрос соглашения, которое фиксируется в кодовой таблице. К сожалению, существуют пять различных кодировок русских букв, поэтому тексты — созданные в одной кодировке, не будут правильно отображаться в другой.

Хронологически одним из первых стандартов кодирования русских букв на компьютерах был КОИ8 (“Код обмена информацией, 8-битный”). Эта кодировка применяется на компьютерах с операционной системой UNIX.

Наиболее распространенная кодировка — это стандартная кириллическая кодировка Microsoft Windows, обозначаемая сокращением CP1251 (“CP” означает “Code Page”, “кодовая страница”). Все Windows-приложения, работающие с русским языком, поддерживают эту кодировку.

Для работы в среде операционной системы MS DOS используется “альтернативная” кодировка, в терминологии фирмы Microsoft — кодировка CP866.

Фирма Apple разработала для компьютеров Macintosh свою собственную кодировку русских букв (Mac).

Международная организация по стандартизации (*International Standards Organization, ISO*) утвердила в качестве стандарта для русского языка еще одну кодировку под названием ISO 8859-5.

Наконец, появился новый международный стандарт Unicode, который отводит на каждый символ не один байт, а два, и потому с его помощью можно закодировать не 256 символов, а целых 65 536. Эту кодировку поддерживает пакет Microsoft Office 97.

Двоичное кодирование текста происходит следующим образом: при нажатии на определенную клавишу в компьютер передается определенная последовательность электрических импульсов, причем каждому символу соответствует своя последовательность электрических импульсов (нулей и единиц на машинном языке). Программа драйвер клавиатуры и экрана по кодовой таблице определяет символ и создает его изображение на экране.

Таким образом, тексты хранятся в памяти компьютера в двоичном коде и программным способом преобразуются в изображения на экране.

II. Практическая часть

Задание 1: в текстовом редакторе Блокнот ввести с помощью числовых кодов последовательность символов в кодировках *Windows* и *MS-DOS.*

Ход работы:

Запустить стандартное приложение Блокнот командой *[Все программы-Стандартные-Блокнот]*.

а) С помощью дополнительной цифровой клавиатуры при нажатой клавише {Alt} ввести число 0224, отпустить клавишу {Alt}, в документе появится символ «а». Повторить процедуру для числовых кодов от 0225 до 0233. Какая последовательность появится в кодировке *Windows*?

б) С помощью дополнительной цифровой клавиатуры при нажатой клавише {Alt} ввести число 224, в документе появится символ «р». Повторить процедуру для числовых кодов от 225 до 233. Какая последовательность появится в кодировке *MS-DOS*?

Задание №2: Какое текстовое сообщение закодировано:

143 174 162 239 167 160 171 160 32 174 225 165 173 236 32 175 165
225 226 224 235 169 32 228 160 224 226 227 170
136 32 162 165 164 165 224 170 168 32 225 32 170 224 160 225 170 160
172 168 32 162 167 239 171 160 46

144 160 173 168 172 32 227 226 224 174 172 44 32 175 224 174 229 174 164 239 32 175 174 32 175 160 224 170 227 44

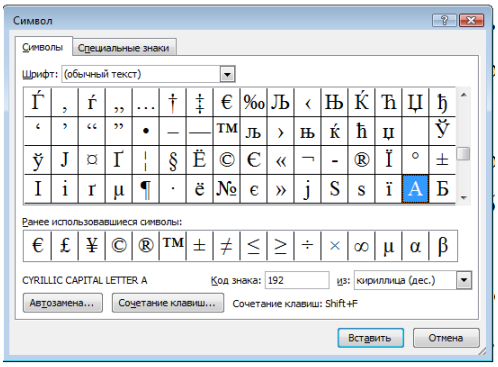
138 168 225 226 236 239 32 175 174 167 174 171 174 226 174 169 32 174 161 162 165 171 160 46

Задание №3: в текстовом редакторе Word определить числовые коды нескольких символов: (в кодировке *Windows)*

Ход работы:

1. Запустить текстовый процессор Word.

2. Ввести команду *[Вставка-Символ-Другие символы...]*. На экране появится диалоговое окно *Символ.* Центральную часть диалоговой панели занимает таблица символов.



3. Для определения десятичного числового кода символа в кодировке *Windows(CP1251)* с помощью раскрывающегося списка *из:* выбрать тип кодировки *кириллица (дес).*

4. В таблице символов выбрать символ (например, прописную букву «А»). В текстовом поле *Код знака:* появится десятичный числовой код символа (в данном случае 192).

5. Закодировать слово: *Информатика*

Задание №4: Решите задачу и оформите её по действиям.

Главный редактор журнала отредактировал статью, и её объём уменьшился на 2 страницы. Каждая страница содержит 32 строки, в каждой строке 64 символа. Информационный объём статьи до редактирования был равен 2 Мбайт. Статья представлена в кодировке Unicode, в которой каждый символ кодируется 2 байтами. Определите информационный объём статьи в Кбайтах в этом варианте представления Unicode после редактирования.

Задание №5: Определите размер следующего предложения закодированное при помощи таблицы кодировки Unicode. **Любишь кататься — люби и саночки возить!**

III. После выполнения данной практической работы оформите отчет, ответив на следующие вопросы:

Задание 1. Почему при кодировании текстовой информации в компьютере в большинстве кодировок используется 256 различных символов, хотя русский алфавит включает только 33 буквы?

Задание 2. С какой целью ввели кодировку Unicode, которая позволяет закодировать 65 530 различных символов?

Задание 3. Представить слово «Кодировка» в пяти различных кодировках Windows, MS-DOS, КОИ 8, Mac, ISO. Для работы используете файлы из папки **Данные**: **кодировка ISO.png**, **кодировка MS-DOS.png**, **кодировка Mac.png**, **кодировка КОИ 8.png**, **кодировка Windows.png**.

Задание 4. Какое текстовое сообщение закодировано:

135 173 160 165 226 32 164 165 162 174 231 170 160 32 168 32 172 160 171 236 231 168 170 44

136 32 167 165 171 165 173 235 169 32 175 174 175 227 163 160 169 44

133 225 171 168 32 164 162 168 166 165 226 225 239 32 226 224 160 172 162 160 169 231 168 170

144 165 171 236 225 235 32 173 165 32 175 165 224 165 161 165 163 160 169 46

Задание 5. Решите задачу и оформите ее по действиям.

Главный редактор журнала отредактировал статью, и её объём уменьшился на 4 страницы. Каждая страница содержит 32 строки, в каждой строке 64 символа. Информационный объём статьи до редактирования был равен 1 Мбайт. Статья представлена в кодировке Unicode, в которой каждый символ кодируется 2 байтами. Определите информационный объём статьи в Кбайтах в этом варианте представления Unicode после редактирования.

Задание 6. Определите размер следующего предложения, закодированное при помощи таблицы кодировки Unicode. **Семь раз отмерь, один раз отрежь!**

Приложение

Таблица ASCII - коды

Базовая таблица ASCII											
32		00100000	56	8	00111000	80	P	01010000	104	h	01101000
33	!	00100001	57	9	00111001	81	Q	01010001	105	i	01101001
34	"	00100010	58	:	00111010	82	R	01010010	106	j	01101010
35	#	00100011	59	;	00111011	83	S	01010011	107	k	01101011
36	$	00100100	60	<	00111100	84	T	01010100	108	l	01101100
37	%	00100101	61	=	00111101	85	U	01010101	109	m	01101101
38	&	00100110	62	>	00111110	86	V	01010110	110	n	01101110
39		00100111	63	?	00111111	87	W	01010111	111	o	01101111
40	(	00101000	64	@	01000000	88	X	01011000	112	p	01110000
41	)	00101001	65	A	01000001	89	Y	01011001	113	q	01110001
42	*	00101010	66	B	01000010	90	Z	01011010	114	r	01110010
43	+	00101011	67	C	01000011	91	[	01011011	115	s	01110011
44	,	00101100	68	D	01000100	92	\	01011100	116	t	01110100
45	-	00101101	69	E	01000101	93	]	01011101	117	u	01110101
46	.	00101110	70	F	01000110	94	^	01011110	118	v	01110110
47	/	00101111	71	G	01000111	95	_	01011111	119	w	01110111
48	0	00110000	72	H	01001000	96	`	01100000	120	x	01111000
49	1	00110001	73	I	01001001	97	a	01100001	121	y	01111001
50	2	00110010	74	J	01001010	98	b	01100010	122	z	01111010
51	3	00110011	75	K	01001011	99	c	01100011	123	{	01111011
52	4	00110100	76	L	01001100	100	d	01100100	124	|	01111100
53	5	00110101	77	M	01001101	101	e	01100101	125	}	01111101
54	6	00110110	78	N	01001110	102	f	01100110	126	~	01111110
55	7	00110111	79	O	01001111	103	g	01100111	127		01111111
Расширенная таблица ASCII											
128	Ђ	10000000	160		10100000	192	А	11000000	224	а	11100000
129	Ѓ	10000001	161	Ў	10100001	193	Б	11000001	225	б	11100001
130	‚	10000010	162	ў	10100010	194	В	11000010	226	в	11100010
131	ѓ	10000011	163	Ј	10100011	195	Г	11000011	227	г	11100011
132	„	10000100	164	¤	10100100	196	Д	11000100	228	д	11100100
133	…	10000101	165	Ґ	10100101	197	Е	11000101	229	е	11100101
134	†	10000110	166	¦	10100110	198	Ж	11000110	230	ж	11100110
135	‡	10000111	167	§	10100111	199	З	11000111	231	з	11100111
136	€	10001000	168	Ё	10101000	200	И	11001000	232	и	11101000
137	‰	10001001	169	©	10101001	201	Й	11001001	233	й	11101001
138	Љ	10001010	170	Є	10101010	202	К	11001010	234	к	11101010
139	‹	10001011	171	«	10101011	203	Л	11001011	235	л	11101011
140	Њ	10001100	172	¬	10101100	204	М	11001100	236	м	11101100
141	Ќ	10001101	173	-	10101101	205	Н	11001101	237	н	11101101
142	Ћ	10001110	174	®	10101110	206	О	11001110	238	о	11101110
143	Џ	10001111	175	Ї	10101111	207	П	11001111	239	п	11101111
144	ђ	10010000	176	°	10110000	208	Р	11010000	240	р	11110000
145	‘	10010001	177	±	10110001	209	С	11010001	241	с	11110001
146	’	10010010	178	І	10110010	210	Т	11010010	242	т	11110010
147	“	10010011	179	і	10110011	211	У	11010011	243	у	11110011
148	”	10010100	180	ґ	10110100	212	Ф	11010100	244	ф	11110100
149	••	10010101	181	µ	10110101	213	Х	11010101	245	х	11110101
150	–	10010110	182	¶	10110110	214	Ц	11010110	246	ц	11110110
151	—	10010111	183	·	10110111	215	Ч	11010111	247	ч	11110111
152	•	10011000	184	ё	10111000	216	Ш	11011000	248	ш	11111000

153	™	10011001	185	№	10111001	217	Щ	11011001	249	щ	11111001
154	љ	10011010	186	є	10111010	218	Ъ	11011010	250	Ъ	11111010
155	›	10011011	187	»	10111011	219	Ы	11011011	251	ы	11111011
156	њ	10011100	188	j	10111100	220	Ь	11011100	252	ь	11111100
157	ќ	10011101	189	S	10111101	221	Э	11011101	253	э	11111101
158	ћ	10011110	190	s	10111110	222	Ю	11011110	254	ю	11111110
159	џ	10011111	191	ї	10111111	223	Я	11011111	255	я	11111111