Кодирование и обработка текстовой информации.

Кодирование текстовой информации

Информация, выраженная с помощью естественных и формальных языков (системы счисления, языки программирования) в письменной форме, обычно называется текстовой информацией. Начиная с 60 – х годов прошлого века, компьютеры всё больше стали использоваться для обработки текстовой информации

Кодирование и декодирование текстовой информации

Для кодирования прописных и строчных букв русского и латинского алфавитов, цифр и ряда специальных знаков (знаки арифметических операций, знаки препинания) достаточно использовать 256 различных символов. По формуле, связывающей количество сообщений N и количество информации I, можно вычислить, какое количество информации необходимо, чтобы закодировать каждый знак:

$$N = 2^I \rightarrow 256 = 2^I \rightarrow 2^8 \rightarrow 2^I \rightarrow I = 8 \text{ битов} = 1 \text{ байт}$$

Кодирование заключается в том, что каждому символу ставится в соответствие уникальный десятичный код от 0 до 255 или соответствующий ему двоичный код от 00000000 до 11111111. Таким образом, человек различает символы по их начертанию, а компьютер – по их коду.

При вводе в компьютер текстовой информации происходит её двоичное кодирование, изображение символа преобразуется в его двоичный код. Пользователь нажимает на клавиатуре клавишу с символом, и в компьютер поступает определённая последовательность из восьми электрических импульсов (двоичный код символа).код символа хранится в оперативной памяти компьютера, где занимает одну ячейку.

В процессе вывода символа на экран компьютера производится обратный процесс – декодирование, т.е. преобразование кода символа в его изображение.

Кодировки русского алфавита

Важно, что присваивание символу конкретного кода – это вопрос соглашения, которое фиксируется в кодовой таблице. Первые 33 кода (0 – 32) этой таблицы соответствует не символам, а операциям (перевод строки, ввод пробела).

Коды с 33 по 127 являются интернациональными и соответствуют символам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания.

Коды с 128 по 255 являются национальными, т.е. в национальных кодировках одному и тому же коду соответствуют различные символы. Существуют пять однобайтовых кодовых таблиц для русских букв (Windows, MS – DOS, КОИ – 8, Mac, ISO), поэтому тексты, созданные в одной кодировке, не будут правильно отображаться в другой

Двоичный код	Десятичный код	КОИ-8	Windows	MS-DOS	Mac	ISO
00000000	0					
. . .						
00001000	8	удаление последнего символа (клавиша {Backspace})				
. . .						
00001101	13	перевод строки (клавиша {Enter})				
. . .						
00100000	32	клавиша {Пробел}				
00100001	33	!				
. . .						
01011010	90	Z				
. . .						
01111111	127	□				
10000000	128	—	Ъ	А	А	к
. . .						
11000010	194	б	В	—	—	Т
. . .						
11001100	204	л	М	|	|	Ь
. . .						
11011101	221	щ	Э	_	Ё	н
. . .						
11111111	255	ь	я	нераз. пробел	нераз. пробел	п

Символ	Windows	MS-DOS	КОИ-8	Mac	ISO	Unicode
А	192	128	225	128	176	1040
В	194	130	247	130	178	1042
М	204	140	237	140	188	1052
Э	221	157	252	157	205	1069
я	255	239	241	223	239	1103

В настоящее время широкое распространение получил новый международный стандарт Unikode, который отводит на каждый символ не один байт, а два, и поэтому с его помощью можно закодировать не 256 символов, а 65 534 различных символов. Такого количества символов достаточно, чтобы закодировать не только русский и латинский алфавиты, цифры, знаки и математические символы, но и греческий, арабский, иврит и другие алфавиты.