Вопрос 1

Кодирование текстовой информации

Двоичное кодирование текстовой информации в компьютере. Информация, выраженная с помощью естественных и формальных языков в письменной форме, обычно называется текстовой информацией.

Для представления текстовой информации (прописные и строчные буквы русского и латинского алфавитов, цифры, знаки и математические символы) достаточно 256 различных знаков. По формуле можно вычислить, какое количество информации необходимо, чтобы закодировать каждый знак:

N = 2i => 256 = 2i => 28 = 2i => I = 8 битов.

Для обработки текстовой информации на компьютере необходимо представить ее в двоичной знаковой системе. Для кодирования каждого знака требуется количество информации, равное 8 битам, т. е. длина двоичного кода знака составляет восемь двоичных знаков. Каждому знаку необходимо поставить в соответствие уникальный двоичный код из интервала от 00000000 до 11111111 (в десятичном коде от 0 до 255) (табл. 3.1).

Человек различает знаки по их начертанию, а компьютер - по их двоичным кодам. При вводе в компьютер текстовой информации происходит ее двоичное кодирование, изображение знака преобразуется в его двоичный код. Пользователь нажимает на клавиатуре клавишу со знаком, и в компьютер поступает определенная последовательность из восьми электрических импульсов (двоичный код знака). Код знака хранится в оперативной памяти компьютера, где занимает одну ячейку.Таблица 3.1. Кодировки знаков



В процессе вывода знака на экран компьютера производится обратное перекодирование, т. е. преобразование двоичного кода знака в его изображение.

Различные кодировки знаков. Присваивание знаку конкретного двоичного кода - это вопрос соглашения, которое фиксируется в кодовой таблице. В существующих кодовых таблицах первые 33 кода (десятичные коды с 0 по 32) соответствуют не знакам, а операциям (перевод строки, ввод пробела и т. д.).

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

Десятичные коды с 128 по 255 являются национальными, т. е. в различных национальных кодировках одному и тому же коду соответствуют разные знаки. К сожалению, в настоящее время существуют пять различных кодовых таблиц для русских букв (Windows, MS-DOS, КОИ-8, Mac, ISO (табл. 3.1 и 3.2)), поэтому тексты, созданные в одной кодировке, не будут правильно отображаться в другой.Таблица 3.2. Десятичные коды некоторых символов в различных кодировках



Например, в кодировке Windows последовательность числовых кодов 221, 194, 204 образует слово "ЭВМ", тогда как в других кодировках это будет бессмысленный набор символов.

К счастью, в большинстве случаев пользователь не должен заботиться о перекодировках текстовых документов, так как это делают специальные программы-конверторы, встроенные в операционную систему и приложения.

В последние годы широкое распространение получил новый международный стандарт кодирования текстовых символов Unicode, который отводит на каждый символ 2 байта (16 битов). По формуле можно определить количество символов, которые можно закодировать согласно этому стандарту:

N = 2i = 216 = 65 536.

Такого количества символов оказалось достаточно, чтобы закодировать не только русский и латинский алфавиты, цифры, знаки и математические символы, но и греческий, арабский, иврит и другие алфавиты.