

Двоичное кодирование текстовой информации

Информация и информационные процессы

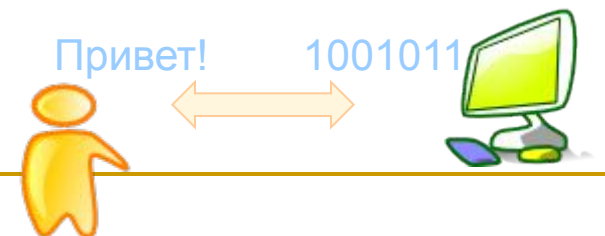
Двоичное кодирование в компьютере

Вся информация, которую обрабатывает компьютер должна быть представлена двоичным кодом с помощью двух цифр: **0** и **1**. *Эти два символа принято называть двоичными цифрами или битами.*

С помощью двух цифр 0 и 1 можно закодировать любое сообщение. Это явилось причиной того, что в компьютере обязательно должно быть организовано два важных процесса: кодирование и декодирование.

Кодирование – преобразование входной информации в форму, воспринимаемую компьютером, т.е. двоичный код.

Декодирование – преобразование данных из двоичного кода в форму, понятную человеку.



Почему двоичное кодирование

С точки зрения технической реализации использование двоичной системы счисления для кодирования информации оказалось намного более простым, чем применение других способов. Действительно, удобно кодировать информацию в виде последовательности нулей и единиц, если представить эти значения как два возможных устойчивых состояния электронного элемента:

0 – отсутствие электрического сигнала;

1 – наличие электрического сигнала.

Эти состояния легко различать. *Недостаток* двоичного кодирования – *длинные коды*. Но в технике легче иметь дело с большим количеством простых элементов, чем с небольшим числом сложных.

Способы кодирования и декодирования информации в компьютере, в первую очередь, зависят от вида информации, а именно, что должно кодироваться: числа, текст, графические изображения или звук.

Двоичное кодирование текстовой информации

Начиная с 60-х годов, компьютеры все больше стали использовать для обработки текстовой информации и в настоящее время большая часть ПК в мире занято обработкой именно текстовой информации.

Традиционно для кодирования одного символа используется количество информации = 1 байту (1 байт = 8 битов).

1 СИМВОЛ – 1 байт (8 бит)

Для кодирования **одного символа** требуется **один байт** информации.

Учитывая, что каждый бит принимает значение 1 или 0, получаем, что с помощью 1 байта можно закодировать 256 различных символов.

$$2^8=256$$

Двоичное кодирование текстовой информации

Кодирование заключается в том, что каждому символу ставится в соответствие уникальный двоичный код от 00000000 до 11111111 (или десятичный код от 0 до 255).

Важно, что присвоение символу конкретного кода – это вопрос соглашения, которое фиксируется кодовой таблицей.

Таблица кодировки

Таблица, в которой всем символам компьютерного алфавита поставлены в соответствие порядковые номера (коды), называется таблицей кодировки.

Для разных типов ЭВМ используются различные кодировки. С распространением IBM PC международным стандартом стала таблица кодировки **ASCII** (**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange) – Американский стандартный код для информационного обмена.

Таблица кодировки ASCII

Стандартной в этой таблице является только первая половина, т.е. символы с номерами от 0 (00000000) до 127 (01111111). Сюда входят буква латинского алфавита, цифры, знаки препинания, скобки и некоторые другие символы.

Остальные 128 кодов используются в разных вариантах. В русских кодировках размещаются символы русского алфавита.

В настоящее время существует 5 разных кодовых таблиц для русских букв (КОИ8, **CP1251**, CP866, Mac, ISO).

В настоящее время получил широкое распространение новый международный стандарт **Unicode**, который отводит на каждый символ два байта. С его помощью можно закодировать 65536 ($2^{16} = 65536$) различных символов.

| символ | 10-й код | 2-й код | символ | 10-й код | 2-й код | символ | 10-й код | 2-й код | символ | 10-й код | 2-й 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 |

| символ | 10-Б код | 2-Б код | символ | 10-Б код | 2-Б код | символ | 10-Б код | 2-Б код | символ | 10-Б код | 2-Б код |
|--------|----------|----------|--------|----------|----------|--------|----------|----------|--------|----------|----------|
| Ъ | 128 | 10000000 | | 160 | 10100000 | А | 192 | 11000000 | а | 224 | 11100000 |
| Ѓ | 129 | 10000001 | Ѹ | 161 | 10100001 | Б | 193 | 11000001 | б | 225 | 11100001 |
| , | 130 | 10000010 | ѹ | 162 | 10100010 | В | 194 | 11000010 | в | 226 | 11100010 |
| ѓ | 131 | 10000011 | Ј | 163 | 10100011 | Г | 195 | 11000011 | г | 227 | 11100011 |
| „ | 132 | 10000100 | Ѡ | 164 | 10100100 | Д | 196 | 11000100 | д | 228 | 11100100 |
| … | 133 | 10000101 | Ѓ | 165 | 10100101 | Е | 197 | 11000101 | е | 229 | 11100101 |
| † | 134 | 10000110 | ‡ | 166 | 10100110 | Ж | 198 | 11000110 | ж | 230 | 11100110 |
| ‡ | 135 | 10000111 | § | 167 | 10100111 | З | 199 | 11000111 | з | 231 | 11100111 |
| € | 136 | 10001000 | Е | 168 | 10101000 | И | 200 | 11001000 | и | 232 | 11101000 |
| ‰ | 137 | 10001001 | © | 169 | 10101001 | Ѓ | 201 | 11001001 | й | 233 | 11101001 |
| Љ | 138 | 10001010 | € | 170 | 10101010 | К | 202 | 11001010 | к | 234 | 11101010 |
| ‹ | 139 | 10001011 | « | 171 | 10101011 | Л | 203 | 11001011 | л | 235 | 11101011 |
| Њ | 140 | 10001100 | – | 172 | 10101100 | М | 204 | 11001100 | м | 236 | 11101100 |
| Ќ | 141 | 10001101 | - | 173 | 10101101 | Н | 205 | 11001101 | н | 237 | 11101101 |
| Ћ | 142 | 10001110 | ® | 174 | 10101110 | О | 206 | 11001110 | о | 238 | 11101110 |
| Ќ | 143 | 10001111 | Ѓ | 175 | 10101111 | П | 207 | 11001111 | п | 239 | 11101111 |
| ђ | 144 | 10010000 | ° | 176 | 10110000 | Р | 208 | 11010000 | р | 240 | 11110000 |
| ‘ | 145 | 10010001 | ± | 177 | 10110001 | С | 209 | 11010001 | с | 241 | 11110001 |
| ’ | 146 | 10010010 | І | 178 | 10110010 | Т | 210 | 11010010 | т | 242 | 11110010 |
| “ | 147 | 10010011 | і | 179 | 10110011 | У | 211 | 11010011 | у | 243 | 11110011 |
| ” | 148 | 10010100 | г | 180 | 10110100 | Ф | 212 | 11010100 | ф | 244 | 11110100 |
| • | 149 | 10010101 | и | 181 | 10110101 | Х | 213 | 11010101 | х | 245 | 11110101 |
| – | 150 | 10010110 | ¶ | 182 | 10110110 | Ц | 214 | 11010110 | ц | 246 | 11110110 |
| — | 151 | 10010111 | · | 183 | 10110111 | Ч | 215 | 11010111 | ч | 247 | 11110111 |
| □ | 152 | 10011000 | ё | 184 | 10111000 | Ш | 216 | 11011000 | ш | 248 | 11111000 |
| ™ | 153 | 10011001 | № | 185 | 10111001 | Щ | 217 | 11011001 | щ | 249 | 11111001 |
| љ | 154 | 10011010 | € | 186 | 10111010 | Ъ | 218 | 11011010 | ъ | 250 | 11111010 |
| › | 155 | 10011011 | » | 187 | 10111011 | Ы | 219 | 11011011 | ы | 251 | 11111011 |
| њ | 156 | 10011100 | ј | 188 | 10111100 | Ь | 220 | 11011100 | ь | 252 | 11111100 |
| ќ | 157 | 10011101 | š | 189 | 10111101 | Э | 221 | 11011101 | э | 253 | 11111101 |
| ћ | 158 | 10011110 | s | 190 | 10111110 | Ю | 222 | 11011110 | ю | 254 | 11111110 |
| џ | 159 | 10011111 | ï | 191 | 10111111 | Я | 223 | 11011111 | я | 255 | 11111111 |

- Таблица расширенного кода ASCII Кодировка Windows-1251 (CP1251)

Информационный объем текста

Сегодня очень многие люди для подготовки писем, документов, статей, книг и пр. используют **компьютерные текстовые редакторы**. Компьютерные редакторы, в основном, работают с **алфавитом размером 256 символов**.

В этом случае легко подсчитать объем информации в тексте. Если **1 символ алфавита несет 1 байт информации**, то надо просто сосчитать количество символов; полученное число даст информационный объем текста в байтах.

Пусть небольшая книжка, сделанная с помощью компьютера, содержит 150 страниц; на каждой странице — 40 строк, в каждой строке — 60 символов. Значит страница содержит $40 \times 60 = 2400$ байт информации. Объем всей информации в книге: $2400 \times 150 = 360\,000$ байт.

Обратите внимание!



Цифры кодируются по стандарту ASCII в двух случаях – при вводе-выводе и когда они встречаются в тексте. Если цифры участвуют в вычислениях, то осуществляется их преобразование в другой двоичный код (см. урок «представление чисел в компьютере»).

Возьмем число **57**.

При использовании **в тексте** каждая цифра будет представлена своим кодом в соответствии с таблицей ASCII. В двоичной системе это – **0011010100110111**.

При использовании **в вычислениях**, код этого числа будет получен по правилам перевода в двоичную систему и получим – **00111001**.

Вопросы и задания:

- В чем заключается кодирование текстовой информации в компьютере?
- Закодируйте с помощью ASCII-кода свою фамилию, имя, номер класса.
- Какое сообщение закодировано в кодировке Windows-1251:
0011010100100000111000011110000011101011111010111110
111011100010
- Считая, что каждый символ кодируется одним байтом, оцените информационный объем следующего предложения из пушкинского четверостишия:
Певец-Давид был ростом мал, Но повалил же Голиафа!