



**Технологии
обработки
текстовой
информации.
Кодовые таблицы**

Технологии обработки текстовой информации.

Кодовые таблицы



- *Текстовая информация* – это информация, представленная в виде букв, знаков препинания и специальных символов некоторой знаковой системы. Буквы и другие знаки принято называть символами. Набор их конечен. Текстовую информацию иногда называют символьной. Знаковая система содержит еще и правила выполнения операций над знаками (грамматика, синтаксис).
- *Кодирование* – это процесс представления каждого символа в виде кода.
- *Код* – набор условных обозначений для представления информации.
- Количество знаков в коде называется *длиной кода*.

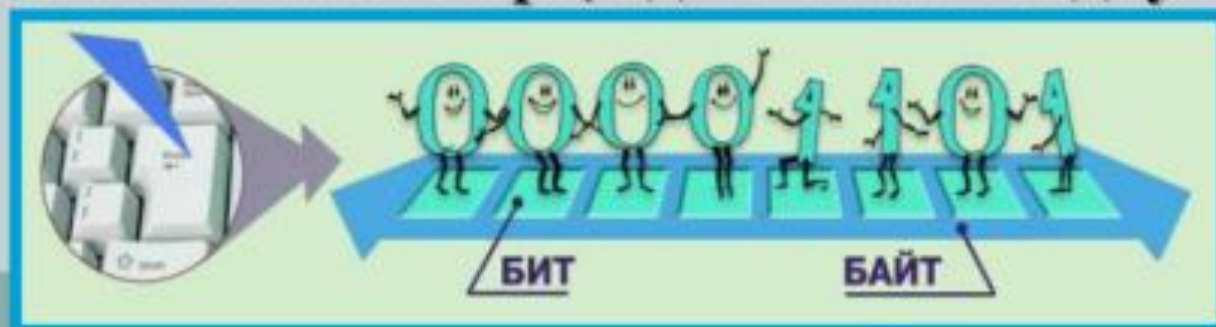


- *Естественные языки* – это знаковые системы с исключениями из правил. Поэтому их нельзя использовать для кодирования информации с последующей обработкой на компьютере. Знаковые системы со строгими правилами называются *формальными*.
- Для компьютерного кодирования информации используется формальная двоичная знаковая система. Физическая природа знаков двоичного компьютерного кода – это электрические импульсы (наличие импульса или его отсутствие). При кодировании используется *кодировочная таблица*. Таблица устанавливает взаимно однозначное соответствие между знаками и их кодами.

Кодирование текстовой информации

Человек различает знаки по их начертанию, а компьютер - по их двоичным кодам. При вводе в компьютер текстовой информации происходит ее двоичное кодирование.

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.



Кодирование текстовой информации



Для представления текстовой информации достаточно 256 различных знаков.

По формуле $N=2^I$ можно вычислить, какое количество информации необходимо, чтобы закодировать каждый знак:

$$N = 2^I \Rightarrow 256 = 2^I \Rightarrow 2^8 = 2^I \Rightarrow$$

$$I = 8 \text{ бит} = 1 \text{ байт}$$

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

Кодирование текстовой информации



- **Декодирование** – процесс обратный кодированию, т.е. код символа преобразуется в его изображение. Процесс декодирования информации осуществляется при выводе информации из оперативной памяти компьютера на экран монитора, например, или на листинг с помощью принтера.

Таблица кодировки



При кодировании каждому символу алфавита ставится в соответствие уникальный двоичный код.

Таблица кодировки – это таблица, в которой всем символам компьютерного алфавита поставлены в соответствие порядковые номера (коды).

Кодировки знаков

| 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 | - | ь | А | А | к |
| ... | | | | | | |
| 11000010 | 194 | б | В | - | - | Т |
| ... | | | | | | |
| 11001100 | 204 | л | М | | | ь |
| ... | | | | | | |
| 11011101 | 221 | щ | Э | _ | Е | н |
| ... | | | | | | |
| 11111111 | 255 | ь | я | нераз. пробел | нераз. пробел | п |

В существующих кодовых таблицах десятичные коды :



- **от 0 до 32** соответствуют операциям (перевод строки, ввод пробела и т.д.);
- **от 33 по 127** соответствуют знакам латинского алфавита, цифрам, знакам арифметических операций и знакам препинания;
- **от 128 по 255** в различных национальных кодировках одному и тому же коду соответствуют разные знаки.

Десятичные коды некоторых символов в различных кодировках



В настоящее время существуют пять различных кодовых таблиц для русских букв (Windows, MS-DOS, КОИ-8, Mac, ISO) поэтому тексты, созданные в одной кодировке, не будут правильно отображаться в другой.

| 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 |

Таблицы кодировки русскоязычных символов



КОИ-8

| | | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|
| — | | Г | Г | └ | ┘ | ┌ | ┐ | └ | ┘ | ┌ | ┐ | ■ | ■ | ■ | ■ | ■ |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 | 143 |
| ▬ | ▬ | ▬ | Г | ● | ● | √ | ≈ | ≤ | ≥ | nbsp | ┘ | ● | 2 | ● | ÷ | |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 | 159 |
| = | | F | ё | П | Г | Г | П | П | Е | Ц | Ц | ┘ | ┘ | ┘ | ┘ | ┘ |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 | 175 |
| ┌ | ┐ | ┐ | Ё | | | └ | ┐ | ┐ | └ | ┐ | ┐ | ┐ | ┐ | ┐ | ┐ | ⊙ |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 | 191 |
| ю | а | б | ц | д | е | ф | г | х | и | й | к | л | м | н | о | о |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 | 207 |
| п | я | р | с | т | у | ж | в | ь | ы | з | ш | э | щ | ч | ъ | ъ |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 | 223 |
| Ю | А | Б | Ц | Д | Е | Ф | Г | Х | И | Й | К | Л | М | Н | О | О |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 | 239 |
| П | Я | Р | С | Т | У | Ж | В | Ь | Ы | З | Ш | Э | Щ | Ч | Ъ | Ъ |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 | 255 |

Таблицы кодировки русскоязычных символов

CP1251(Windows)

| | | | | | | | | | | | | | | | |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Á | à | , | è | „ | … | † | ‡ | € | ‰ | É | < | Й | Ў | Ó | Ú |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| á | ‘ | ’ | “ | ” | • | – | — | ê | ™ | é | > | ò | í | ó | ú |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| nbsp | ÿ | Ы | Э | И | Ы | ¡ | § | Ë | © | Ю | « | ¬ | shy | ® | Я |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| • | ± | Ы | Э | ’ | μ | ¶ | • | ë | № | ю | » | э | ю | я | я |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 |
| А | Б | В | Г | Д | Е | Ж | З | И | Й | К | Л | М | Н | О | П |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 |
| Р | С | Т | У | Ф | Х | Ц | Ч | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 |
| а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 |
| р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |

Таблицы кодировки русскоязычных символов



CP866 (MS-DOS)

| | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| А | Б | В | Г | Д | Е | Ж | З | И | Й | К | Л | М | Н | О | П |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| Р | С | Т | У | Ф | Х | Ц | Ч | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| ▩ | ▨ | ▧ | | ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 |
| ┐ | ┌ | └ | ┘ | ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 |
| ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ | ┌ | ┐ | └ | ┘ | ▩ | ▨ | ▧ | ▦ |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 |
| р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 |
| Ё | ё | Є | є | Ї | ї | Ў | ў | • | • | • | √ | № | ¤ | ■ | nbsp |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |

Таблицы кодировки русскоязычных символов



ISO

| | | | | | | | | | | | | | | | |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| І | І | І | І | І | І | І | І | І | І | І | І | І | І | І | І |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| І | І | І | І | І | І | І | І | І | І | І | І | І | І | І | І |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| nbsp | Ё | Ъ | Ѓ | Є | Ѕ | І | Ї | Ј | Љ | Њ | Ћ | Ќ | shy | Ў | Џ |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| А | Б | В | Г | Д | Е | Ж | З | И | Й | К | Л | М | Н | О | П |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 |
| Р | С | Т | У | Ф | Х | Ц | Ч | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 |
| а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 |
| р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 |
| № | ё | ђ | ѓ | є | ѕ | і | ї | ј | љ | њ | ћ | ќ | ѕ | ў | џ |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |

Mac

| | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|
| А | Б | В | Г | Д | Е | Ж | З | И | Й | К | Л | М | Н | О | П |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| Р | С | Т | У | Ф | Х | Ц | Ч | Ш | Щ | Ъ | Ы | Ь | Э | Ю | Я |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| † | ° | Ы | £ | § | • | ¶ | Ы | © | © | ™ | Á | á | è | à | è |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| Ё | ± | ≤ | ≥ | э | μ | г | ó | ю | ю | я | я | É | é | й | ò |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 |
| э | ю | ¬ | √ | f | ≈ | Δ | « | » | ... | nbsp | Ó | ó | Й | й | я |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 |
| - | — | “ | ” | ‘ | ’ | ÷ | „ | ù | Ы | У | у | № | Ё | ё | я |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 |
| а | б | в | г | д | е | ж | з | и | й | к | л | м | н | о | п |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 233 | 234 | 235 | 236 | 237 | 238 | 239 |
| р | с | т | у | ф | х | ц | ч | ш | щ | ъ | ы | ь | э | ю | я |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |

Кодовая таблица ASCII

- Для разных типов ЭВМ используются различные таблицы кодировки.
- С распространением персональных компьютеров типа IBM PC международным стандартом стала таблица кодировки под названием ASCII (American Standard Code for Information Interchange) – американский стандартный код для информационного обмена

| | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| sp | ! | " | # | \$ | % | & | ' | (|) | * | + | , | - | . | / |
| 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 |
| @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 |
| P | Q | R | S | T | U | V | W | X | Y | Z | [| \ |] | ^ | _ |
| 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 |
| ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 |
| p | q | r | s | t | u | v | w | x | y | z | { | | } | ~ | |
| 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | |



- Этот формат оперирует с 256 численными кодами, имеющими значения от 0 до 255. В соответствие каждому коду ставится определенный символ (буква, цифра, знак препинания, математический символ или символ псевдографики). Это соответствие задается с помощью стандартных кодовых таблиц с различными номерами (например, таблица 866 предназначена для русскоязычных пользователей). Не содержит форматирования текста, поэтому является переносимым между различными операционными системами и программами.



- ANSI (American National Standard Interface) кодировка в среде Windows. У этих кодировок совпадают те части, которые относятся к латинскому алфавиту, специальным символам, цифрам, знакам препинания и математическим операциям, а различаются относящиеся к другим алфавитам и псевдографике

Понятие кодировки Unicode(UCS - 2)



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: $N = 2^1 = 2^{16} = 65\,536$.

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.

Текст - последовательность символов компьютерного алфавита.

Текстовая информация - это информация, выраженная с помощью естественных и формальных языков в письменной форме (прописные и строчные буквы русского и латинского алфавитов, цифры, знаки и математические символы).