§ 4.6

Оценка количественных параметров текстовых документов

Ключевые слова:

- кодовая таблица
- восьмиразрядный двоичный код
- информационный объём текста

4.6.1. Представление текстовой информации в памяти компьютера

Текст состоит из символов — букв, цифр, знаков препинания и т. д., которые человек различает по начертанию. *Компьютер различает вводимые символы по их двоичному коду.* Вы нажимаете на клавиатуре символьную клавишу, и в компьютер поступает определённая последовательность электрических импульсов разной силы, которую можно представить в виде цепочки из восьми нулей и единиц (двоичного кода).

Мы уже говорили о том, что разрядность двоичного кода i и количество возможных кодовых комбинаций N связаны соотношением: $2^i = N$. Восьмиразрядный двоичный код позволяет получить 256 различных кодовых комбинаций: $2^8 = 256$.

С помощью такого количества кодовых комбинаций можно закодировать все символы, расположенные на клавиатуре компьютера, — строчные и прописные русские и латинские буквы, цифры, знаки препинания, знаки арифметических операций, скобки и т. д., а также ряд управляющих символов, без которых невозможно создание текстового документа (удаление предыдущего символа, перевод строки, пробел и др.).

Соответствие между изображениями символов и кодами символов устанавливается с помощью кодовых таблиц.

Все кодовые таблицы, используемые в любых компьютерах и любых операционных системах, подчиняются международным стандартам кодирования символов.

Кодовая таблица содержит коды для 256 различных символов, пронумерованных от 0 до 255. Первые 128 кодов во всех кодовых таблицах соответствуют одним и тем же символам:

- коды с номерами от 0 до 32 соответствуют управляющим символам;
- коды с номерами от 33 до 127 соответствуют изображаемым символам — латинским буквам, знакам препинания, цифрам, знакам арифметических операций и т. д.

Эти коды были разработаны в США и получили название ASCII (American Standart Code for Information Interchange — Американский стандартный код для обмена информацией).

В таблице 4.1 представлен фрагмент кодировки ASCII.

Коды с номерами от 128 до 255 используются для кодирования букв национального алфавита, символов национальной валюты и т. п. Поэтому в кодовых таблицах для разных языков одному и тому же коду соответствуют разные символы. Более того, для многих языков

Таблица 4.1

Фрагмент кодировки ASCII

Символ	Десятичный код (номер)	Двоичный код	Символ	Десятичный код (номер)	Двоичный код
Пробел	32	00100000	0	48	00110000
!	33	00100001	1	49	00110001
#	35	00100011	2	50	00110010
$	36	00100100	3	51	00110011
*	42	00101010	4	52	00110100
+	43	00101011	5	53	00110101
,	44	00101100	6	54	00110110
-	45	00101101	7	55	00110111
.	46	00101110	8	56	00111000
/	47	00101111	9	57	00111001

Продолжение табл. 4.1

Символ	Десятичный код (номер)	Двоичный код	Символ	Десятичный код (номер)	Двоичный код
A	65	01000001	N	78	01001110
B	66	01000010	O	79	01001111
C	67	01000011	P	80	01010000
D	68	01000100	Q	81	01010001
E	69	01000101	R	82	01010010
F	70	01000110	S	83	01010011
G	71	01000111	T	84	01010100
H	72	01001000	U	85	01010101
I	73	01001001	V	86	01010110
J	74	01001010	W	87	01010111
K	75	01001011	X	88	01001000
L	76	01001100	Y	89	01001001
M	77	01001100	Z	90	01011010

существует несколько вариантов кодовых таблиц (например, для русского языка их около десятка!).

В таблице 4.2 представлены десятичные и двоичные коды нескольких букв русского алфавита в двух различных кодировках.

Таблица 4.2

Коды русских букв в разных кодировках

Символ	Кодировка			
	Windows		КОИ-8	
	десятичный код	двоичный код	десятичный код	двоичный код
А	192	11000000	225	11100001
Б	193	11000001	226	11100010
В	194	11000010	247	11110111

Например, последовательности двоичных кодов

11010010 11000101 11001010 11010001 11010010

в кодировке Windows будет соответствовать слово «ТЕКСТ», а в кодировке КОИ-8 — бессмысленный набор символов «рейяр».

Как правило, пользователь не должен заботиться о перекодировании текстовых документов, так как это делают специальные программы-конверторы, встроенные в операционную систему и приложения.

Восьмиразрядные кодировки обладают одним серьёзным ограничением: количество различных кодов символов в этих кодировках недостаточно велико, чтобы можно было одновременно пользоваться более чем двумя языками. Для устранения этого ограничения был разработан новый стандарт кодирования символов, получивший название **Unicode**. В Unicode каждый символ кодируется шестнадцатиразрядным двоичным кодом. Такое количество разрядов позволяет закодировать 65 536 различных символов:

$$2^{16} = 65\ 536.$$

Первые 128 символов в Unicode совпадают с таблицей ASCII; далее размещены алфавиты всех современных языков, а также все математические и иные научные символьные обозначения. С каждым годом Unicode получает всё более широкое распространение.

В Единой коллекции цифровых образовательных ресурсов (http://sc.edu.ru) размещены анимации «Клавиатура ПЭВМ: принципы работы; устройство клавиши» (134923), «Клавиатура ПЭВМ: принципы работы; сканирование клавиш» (135019), «Клавиатура ПЭВМ: формирование кода введенного символа» (134868), которые помогут вам наглядно увидеть, как формируется код символа, введённого с клавиатуры.

4.6.2. Информационный объём фрагмента текста

Вам известно, что информационный объём сообщения I равен произведению количества символов K в сообщении на информационный вес символа алфавита i: $I = K \cdot i$.

В зависимости от разрядности используемой кодировки информационный вес символа текста, создаваемого на компьютере, может быть равен:

- 8 битов (1 байт) — восьмиразрядная кодировка;
- 16 битов (2 байта) — шестнадцатиразрядная кодировка.