Введение

«Кодирование текста»

13.12.2015

Компьютер является универсальным устройством обработки информации и способен работать с различными ее типами. В том числе и с текстовой. Давайте рассмотрим эту возможность подробнее.

Проблема состоит в том, что компьютер в памяти может хранить только числа, причем числа двоичные (о системах счисления). Каким же образом в память можно поместить текст? Очень просто. Ведь символы можно пронумеровать, т. е. дать каждому символу цифровой код и уже его хранить в памяти. Собственно, все именно так устроено. Но тут возникает проблема — Вася пронумеровал буквы так, что прописная буква А имеет код 1, а у Пети прописная буква А имеет код 34. В итоге текст, закодированный на компьютере Васи будет некорректно отображаться на компьютере Пети и наоборот. Как поступить? Очень просто — закодировать символы и принять такое кодирование как стандартное. Таким образом появилась таблица кодировок ASCII (произносится аски).

Разработчики первых компьютеров использовали английский язык, поэтому им необходимо было закодировать 26 прописных букв, 26 строчных букв (строчная буква А и прописная буква а для компьютера совершенно разные символы и имеют разные коды), 10 цифр, знаки препинания, знаки арифметических операций, пробел (да, пробел — тоже символ и имеет свой код), различные спецзнаки. В итоге получается немногим более 100 символов. Сколько памяти необходимо, чтобы сохранить код одного символа? Давайте посчитаем. Воспользуемся формулой

$$2^i=N$$

где **N** — количество символов, а **i** — количество памяти в битах, необходимое для хранения одного символа. Значение N примем равным 100. Чему же равно i? Если i = 6, то $2^6=64$. Этого мало, ведь у нас 100 символов. Если i=7, то $2^7=128$ — то, что нам нужно. Если необходимо закодировать 128 символов, на каждый необходимо 7 бит памяти.

А что же делать тем, кто использует кириллические буквы? Ведь места в получившейся таблице не хватает. А почему бы не расширить ее? Если на каждый символ отвести 8 бит памяти, то можно будет закодировать уже $2^8=256$ символов. Таким образом, появилась расширенная таблица кодировок ASCII, в которой первая часть (символы с десятичными кодами от 0 до 127) содержит латинский алфавит, цифры, знаки препинания, знаки арифметических операций, спецсимволы, а вторая часть (символы с кодами от 128 до 255) — национальные символы разных стран. В России это русские буквы. К сожалению, было несколько вариантов второй части таблицы кодировок ASCII для кириллического алфавита, что часто приводило к некорректному отображению текста. К примеру, прописная буква А в различных таблицах кодировок имеет такие коды:

Получается, что русский текст, закодированный в кодировке Windows, будет нечитаем в кодировке КОИ-8. Аналогично и с другими кодировками. Как же решить эту проблему? Может сделать действительно единую международную кодовую таблицу, в которой можно будет поместить гораздо больше, чем 256 символов? Так и поступили в 1991 году, когда консорциум UNICODE представил стандарт кодирования Unicode (читается как юникод), который позволил закодировать символы практически всех языков Мира. Если в ASCII для хранения одного символа требуется 8 бит или 1 байт памяти, то в Unicode — 2 байта или 16 бит. Соответственно, используя 16бит мы можем закодировать $2^{16} = 65536$ различных символов! Кроме того, стандарт Unicode развивается и на данный момент позволяет закодировать гораздо больше, чем 65536 символов.

Задание 1.

Статья, набранная на компьютере, содержит 16 страниц, на каждой странице 30 строк, в каждой строке 32 символа. Определите информационный объём статьи в одной из кодировок Unicode, в которой каждый символ кодируется 16 битами.
1) 24 Кбайт 2) 30 Кбайт 3) 480 байт 4) 240 байт

Решение:

Найдем общее количество символов на одной странице, для этого умножим количество строк на странице на количество символов в строке — 30 * 32 = 960 символов.

Найдем общее количество символов во всем тексте, для этого умножим количество страниц на количество символов на одной странице — 16 * 960 = 15360 символов.

Так как каждый символ кодируется 16 битами, а 16 бит = 2 байта, то весь текст займет 15360 * 2 байта = 30720 байта. Как видим, из предложенных вариантов ответа в байтах полученного нами нет, поэтому переведем полученный результат в килобайты. Для этого разделим 30720 на 1024: 30720 / 1024 = 30Кбайт.

Правильный ответ 2) 30Кбайт.

Второй вариант решения задачи предполагает знание степеней двойки и единиц измерения информации.

Итак, количество символов во всем тексте, учитывая, что $32 = 2^5$, а $16 = 2^4$ будет равно $30 * 32 * 16 = 30 * 2^5 * 2^4 = 30 * 2^9$ символов.

Так как каждый символ занимает 2 байта, то для всего текста потребуется $30 * 2^9 * 2 = 30 * 2^{10}$ байт.

А так как 2^{10} байт это 1Кбайт, то в итоге получим **30Кбайт**.

Задание 2.

В одной из кодировок Unicode каждый символ кодируется 16 битами. Определите размер следующего предложения в данной кодировке.

Я к вам пишу – чего же боле? Что я могу ещё сказать?

1) 52 байт 2) 832 бит 3) 416 байт 4)104 бит

Решение:

Для начала посчитаем количество символов в предложении. Именно символов, не букв! То есть знак пробела, знак вопроса мы тоже считаем. В итоге у нас получается 52 символа. Из условия известно, что каждый символ кодируется 16 битами. Значит, чтобы найти информационный объем всего предложения, мы должны умножить 52 на 16.

52 * 16 = 832 бита.

Среди вариантов ответа есть найденный нами. **Правильный ответ 2.**

Задание 3.

*В одной из кодировок Unicode каждый символ кодируется 16 битами. **О**пределите информационный объем следующего предложения в данной кодировке.*

Я памятник себе воздвиг нерукотворный.

1) 76 бит 2) 608 бит 3) 38 байт 4) 544 бит

Решение:

Принцип решения подобного класса задач остается прежним — посчитать количество символов и умножить полученное число на информационный объем одного символа. В условии сказано, что каждый символ кодируется 16 битам (рекомендую ознакомиться со статьей кодирование текста для понимания принципов хранения текста в памяти компьютера). Итак, считаем количество символов в строке. Напомню очередной раз, что пробелы, знаки препинания — это тоже символы и их тоже надо считать. В предложении 38 символов. Умножив 38 символов на 16 бит получим 608 бит. В предложенных вариантах такой встречается, значит **правильный ответ 2.**

Задание 4.

*Текст рассказа набран на компьютере. Информационный объём получившегося файла 15 Кбайт. Текст занимает 10 страниц, на каждой странице одинаковое количество строк, в каждой строке 64 символа. Все символы представлены в кодировке Unicode. В используемой версии Unicode каждый символ кодируется 2 байтами. **О**пределите, сколько строк помещается на каждой странице.*
1) 48 2) 24 3) 32 4) 12

Решение:

15Кбайт = 15 * 2^{10} байт.

Обозначим количество строк X. Тогда во всем тексте будет 10 * 64 * X символов. А если каждый символ занимает 2 байта, то во всем тексте будет 10 * 64 * X * 2 байта. Осталось вспомнить степени двойки и решить простейшее уравнение:

$15*2^{10} = 10*64*X*2$

$15*2^{10} = 10*2^6 *X*2$

$15*2^{10} = 10*2^7 *X$

$X=15 * 2^{10} / 10 * 2^7 = 3 * 2^3 / 2 = 3 * 2^2 = 3 * 4 = 12$

Ответ: 12 строк — это **4-й вариант**.

Задание 5.

Статья, набранная на компьютере, содержит 64 страницы, на каждой странице 40 строк, в каждой строке 64 символа. Определите размер статьи в кодировке КОИ-8, в которой каждый символ кодируется 8 битами.

1) 160 Кбайт
2) 320 Кбайт
3) 1280 байт
4) 2560 байт

Решение:

Найдем количество символов в статье:

$$64 \cdot 40 \cdot 64 = 2^6 \cdot 5 \cdot 2^3 \cdot 2^6 = 5 \cdot 2^{15}.$$

Один символ кодируется одним байтом, 2^{10} байт составляют 1 килобайт, поэтому информационный объем статьи составляет

$$5 \cdot 2^{15} \text{ байт} = 5 \cdot 2^5 \text{ килобайт} = 160 \text{ Кб}.$$

Правильный ответ указан под номером 1.

Задание 6.

Реферат, набранный на компьютере, содержит 14 страниц, на каждой странице 36 строк, в каждой строке 64 символа. Для кодирования символов используется кодировка Unicode, при которой каждый символ кодируется 2 байтами. Определите информационный объём реферата.

1) 12 Кбайт
2) 24 Кбайта
3) 58 Кбайт
4) 63 Кбайта

Решение:

Найдем количество символов в статье:

$$14 \cdot 36 \cdot 64 = 2^3 \cdot 63 \cdot 2^6 = 63 \cdot 2^9.$$

Один символ кодируется двумя байтами, 2^{10} байт составляют 1 килобайт, поэтому информационный объем статьи составляет

$$63 \cdot 2^{10} \text{ байт} = 63 \text{ Кб}.$$

Правильный ответ указан под номером 4.

Задания для самоконтроля

1. **Задание 1 № 1.** Статья, набранная на компьютере, содержит 32 страницы, на каждой странице 40 строк, в каждой строке 48 символов. Определите размер статьи в кодировке КОИ-8, в которой каждый символ кодируется 8 битами.

1) 120 Кбайт
2) 480 байт
3) 960 байт
4) 60 Кбайт

2. **Задание 1 № 101.** Статья, набранная на компьютере, содержит 64 страницы, на каждой странице 40 строк, в каждой строке 48 символов. Определите размер статьи в кодировке КОИ-8, в которой каждый символ кодируется 8 битами.

1) 1920 байт
2) 960 байт
3) 120 Кбайт
4) 240 Кбайт

3. **Задание 1 № 161.** В одной из кодировок Unicode каждый символ кодируется 16 битами. Определите размер следующего предложения в данной кодировке: **Я вас любил: любовь ещё, быть может, в душе моей угасла не совсем.**

1) 66 байт
2) 1056 бит
3) 528 байт
4) 132 бит

4. **Задание 1 № 342.** Реферат, набранный на компьютере, содержит 16 страниц, на каждой странице 50 строк, в каждой строке 64 символа. Для кодирования символов используется кодировка Unicode, при которой каждый символ кодируется 16 битами. Определите информационный объём реферата.

1) 320 байт
2) 100 Кбайт
3) 128 Кбайт
4) 1 Мбайт

Учитель информатики и ИКТ : Шаповалов И.Л.