**Сожержание**

I. История кодирования информации………………………………..3

II. Кодирование информации…………………………………………4

III. Кодирование текстовой информации…………………………….4

IV. Виды таблиц кодировок…………………………………………...6

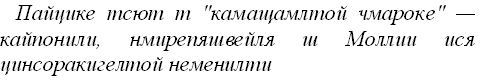
V. Расчет количества текстовой информации………………………14

Список используемой литературы…………………………………..16

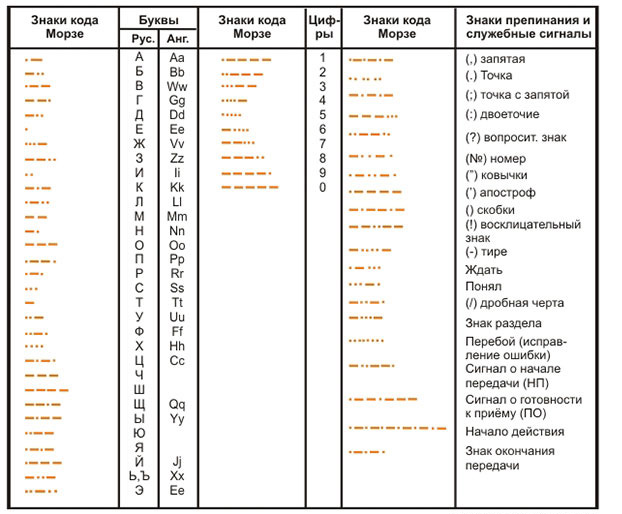
**I. История кодирования информации**

Человечество использует шифрование (кодировку) текста с того самого момента, когда появилась первая секретная информация. Перед вами несколько приёмов кодирования текста, которые были изобретены на различных этапах развития человеческой мысли:

* криптография – это тайнопись, система изменения письма с целью сделать текст непонятным для непосвященных лиц;



* азбука Морзе или неравномерный телеграфный код, в котором каждая буква или знак представлены своей комбинацией коротких элементарных посылок электрического тока (точек) и элементарных посылок утроенной продолжительности (тире);



* сурдожесты – язык жестов, используемый людьми с нарушениями слуха.



Один из самых первых известных методов шифрования носит имя римского императора Юлия Цезаря (I век до н.э.) . Этот метод основан на замене каждой буквы шифруемого текста, на другую, путем смещения в алфавите от исходной буквы на фиксированное количество символов, причем алфавит читается по кругу, то есть после буквы я рассматривается а. Так слово «байт» при смещении на два символа вправо кодируется словом «гвлф». Обратный процесс расшифровки данного слова – необходимо заменять каждую зашифрованную букву, на вторую слева от неё.

**II. Кодирование информации**

Код – это набор условных обозначений (или сигналов) для записи (или передачи) некоторых заранее определенных понятий.

Кодирование информации – это процесс формирования определенного представления информации. В более узком смысле под термином «кодирование» часто понимают переход от одной формы представления информации к другой, более удобной для хранения, передачи или обработки.

Обычно каждый образ при кодировании (иногда говорят – шифровке) представлении отдельным знаком.

Знак - это элемент конечного множества отличных друг от друга элементов.

В более узком смысле под термином "кодирование" часто понимают переход от одной формы представления информации к другой, более удобной для хранения, передачи или обработки.

На компьютере можно обрабатывать текстовую информацию. При вводе в компьютер каждая буква кодируется определенным числом, а при выводе на внешние устройства (экран или печать) для восприятия человеком по этим числам строятся изображения букв. Соответствие между набором букв и числами называется кодировкой символов.

Как правило, все числа в компьютере представляются с помощью нулей и единиц (а не десяти цифр, как это привычно для людей). Иными словами, компьютеры обычно работают в двоичной системе счисления, поскольку при этом устройства для их обработки получаются значительно более простыми. Ввод чисел в компьютер и вывод их для чтения человеком может осуществляться в привычной десятичной форме, а все необходимые преобразования выполняют программы, работающие на компьютере.

**III. Кодирование текстовой информации**

Одна и та же информация может быть представлена (закодирована) в нескольких формах. C появлением компьютеров возникла необходимость кодирования всех видов информации, с которыми имеет дело и отдельный человек, и человечество в целом. Но решать задачу кодирования информации человечество начало задолго до появления компьютеров. Грандиозные достижения человечества - письменность и арифметика - есть не что иное, как система кодирования речи и числовой информации. Информация никогда не появляется в чистом виде, она всегда как-то представлена, как-то закодирована.

Двоичное кодирование – один из распространенных способов представления информации. В вычислительных машинах, в роботах и станках с числовым программным управлением, как правило, вся информация, с которой имеет дело устройство, кодируется в виде слов двоичного алфавита.

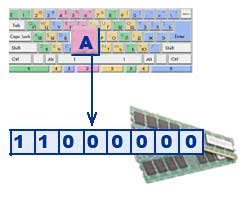
Начиная с конца 60-х годов, компьютеры все больше стали использоваться для обработки текстовой информации, и в настоящее время основная доля персональных компьютеров в мире (и большая часть времени) занята обработкой именно текстовой информации. Все эти виды информации в компьютере представлены в двоичном коде, т. е. используется алфавит мощностью два (всего два символа 0 и 1). Связано это с тем, что удобно представлять информацию в виде последовательности электрических импульсов: импульс отсутствует (0), импульс есть (1).

Такое кодирование принято называть двоичным, а сами логические последовательности нулей и единиц - машинным языком.

С точки зрения ЭВМ текст состоит из отдельных символов. К числу символов принадлежат не только буквы (заглавные или строчные, латинские или русские), но и цифры, знаки препинания, спецсимволы типа "=", "(", "&" и т.п. и даже (обратите особое внимание!) пробелы между словами.

Тексты вводятся в память компьютера с помощью клавиатуры. На клавишах написаны привычные нам буквы, цифры, знаки препинания и другие символы. В оперативную память они попадают в двоичном коде. Это значит, что каждый символ представляется 8-разрядным двоичным кодом.

Традиционно для кодирования одного символа используется количество информации, равное 1 байту, т. е. I = 1 байт = 8 бит. При помощи формулы, которая связывает между собой количество возможных событий К и количество информации I, можно вычислить сколько различных символов можно закодировать (считая, что символы - это возможные события): К = 2I = 28 = 256, т. е. для представления текстовой информации можно использовать алфавит мощностью 256 символов.



Такое количество символов вполне достаточно для пред­ставления текстовой информации, включая прописные и строчные буквы русского и латинского алфавита, цифры, знаки, графические символы и пр.

Кодирование заключается в том, что каждому символу ставится в соответствие уникальный десятичный код от 0 до 255 или соответствующий ему двоичный код от 00000000 до 11111111. Таким образом, человек различает символы по их начертанию, а компьютер - по их коду.

Удобство побайтового кодирования символов очевидно, поскольку байт - наименьшая адресуемая часть памяти и, следовательно, процессор может обратиться к каждому символу отдельно, выполняя обработку текста. С другой стороны, 256 символов – это вполне достаточное количество для представления самой разнообразной символьной информации.

В процессе вывода символа на экран компьютера произ­водится обратный процесс — декодирование, то есть преоб­разование кода символа в его изображение. Важно, что присвоение символу конкретного кода — это вопрос соглашения, которое фиксируется в кодовой табли­це.

Теперь возникает вопрос, какой именно восьмиразрядный двоичный код поставить в соответствие каждому символу. Понятно, что это дело условное, можно придумать множество способов кодировки.

Все символы компьютерного алфавита пронумерованы от 0 до 255. Каждому номеру соответствует восьмиразрядный двоичный код от 00000000 до 11111111. Этот код просто порядковый номер символа в двоичной системе счисления.

**IV. Виды таблиц кодировок**

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

Для разных типов ЭВМ используются различные таблицы кодировки.

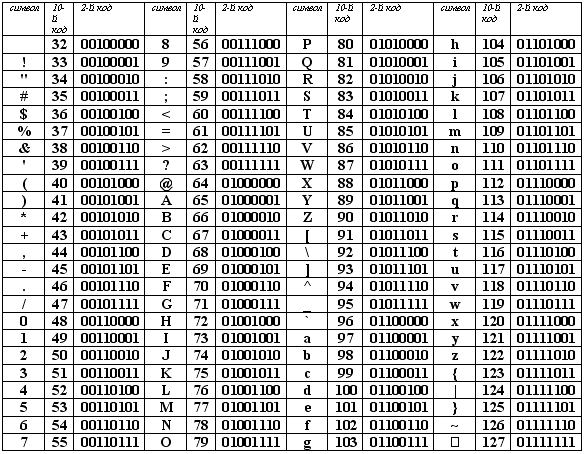
В качестве международного стандарта принята кодовая таблица ASCII (American Standard Code for Information Interchange - Американский стандартный код для информационного обмена), кодирующая первую половину символов с числовыми кодами от 0 до 127 ( коды от 0 до 32 отведены не символам, а функциональным клавишам).

Таблица кодов ASCII делится на две части.

Международным стандартом является лишь первая половина таблицы, т.е. символы с номерами от 0 (00000000), до 127 (01111111).

**Структура таблицы кодировки ASCII**

|  |  |  |
| --- | --- | --- |
| **Порядковый номер** | **Код** | **Символ** |
| 0 - 31 | 00000000 - 00011111 | Символы с номерами от 0 до 31 принято называть управляющими.  Их функция – управление процессом вывода текста на экран или печать, подача звукового сигнала, разметка текста и т.п. |
| 32 - 127 | 0100000 - 01111111 | Стандартная часть таблицы (английский). Сюда входят строчные и прописные буквы латинского алфавита, десятичные цифры, знаки препинания, всевозможные скобки, коммерческие и другие символы.  Символ 32 - пробел, т.е. пустая позиция в тексте.  Все остальные отражаются определенными знаками. |
| 128 - 255 | 10000000 - 11111111 | Альтернативная часть таблицы (русская).  Вторая половина кодовой таблицы ASCII, называемая кодовой страницей (128 кодов, начиная с 10000000 и кончая 11111111), может иметь различные варианты, каждый вариант имеет свой номер.  Кодовая страница в первую очередь используется для размещения национальных алфавитов, отличных от латинского. В русских национальных кодировках в этой части таблицы размещаются символы русского алфавита. |

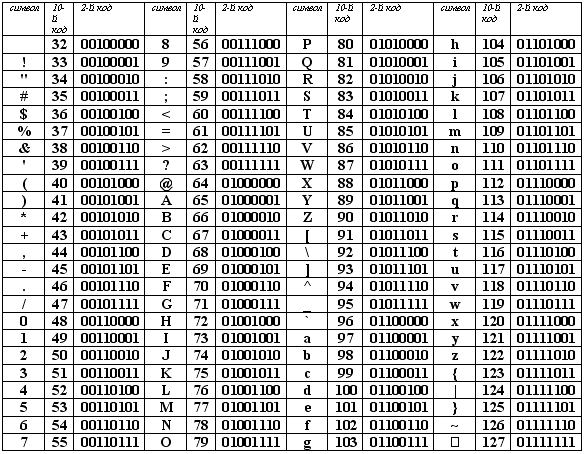


**Первая половина таблицы кодов ASCII**

Обращается внимание на то, что в таблице кодировки буквы (прописные и строчные) располагаются в алфавитном порядке, а цифры упорядочены по возрастанию значений. Такое соблюдение лексикографического порядка в расположении символов называется принципом последовательного кодирования алфавита.

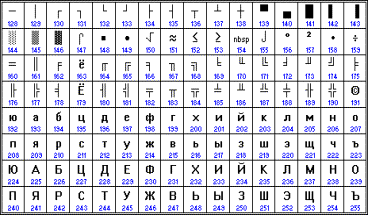
Для букв русского алфавита также соблюдается принцип последовательного кодирования.

**Вторая половина таблицы кодов ASCII**

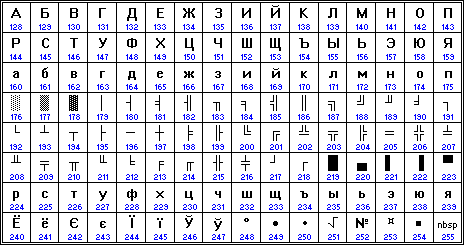


К сожалению, в настоящее время существуют пять различных кодировок кириллицы (КОИ8-Р, Windows. MS-DOS, Macintosh и ISO). Из-за этого часто возникают проблемы с переносом русского текста с одного компьютера на другой, из одной программной системы в другую.

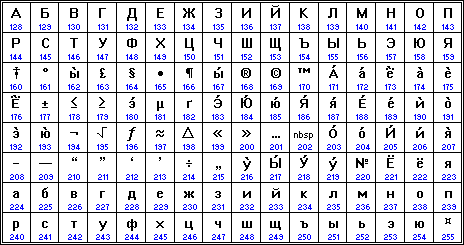
Хронологически одним из первых стандартов кодирования русских букв на компьютерах был КОИ8 ("Код обмена информацией, 8-битный"). Эта кодировка применялась еще в 70-ые годы на компьютерах серии ЕС ЭВМ, а с середины 80-х стала использоваться в первых русифицированных версиях операционной системы UNIX.



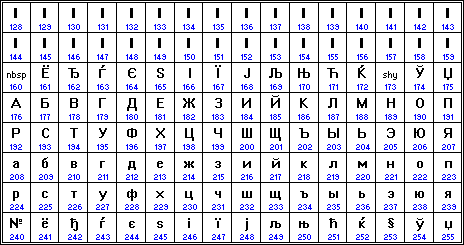
От начала 90-х годов, времени господства операционной системы MS DOS, остается кодировка CP866 ("CP" означает "Code Page", "кодовая страница").



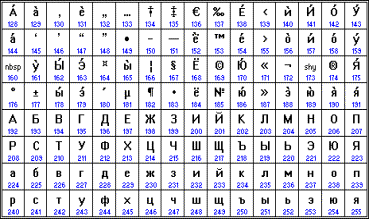
Компьютеры фирмы Apple, работающие под управлением операционной системы Mac OS, используют свою собственную кодировку Mac.



Кроме того, Международная организация по стандартизации (International Standards Organization, ISO) утвердила в качестве стандарта для русского языка еще одну кодировку под названием ISO 8859-5.



Наиболее распространенной в настоящее время является кодировка Microsoft Windows, обозначаемая сокращением CP1251. Введена компанией Microsoft; с учетом широкого распространения операционных систем (ОС) и других программных продуктов этой компании в Российской Федерации она нашла широкое распространение.



С конца 90-х годов проблема стандартизации символьного кодирования решается введением нового международного стандарта, который называется Unicode.



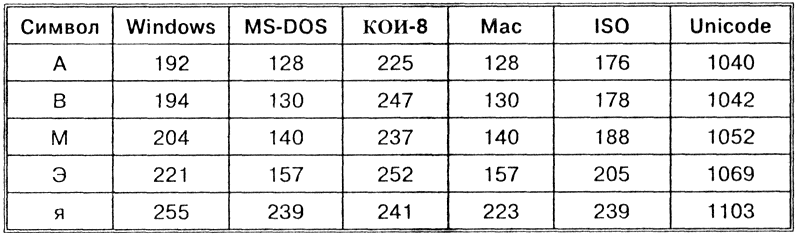
Это 16-разрядная кодировка, т.е. в ней на каждый символ отводится 2 байта памяти. Конечно, при этом объем занимаемой памяти увеличивается в 2 раза. Но зато такая кодовая таблица допускает включение до 65536 символов. Полная спецификация стандарта Unicode включает в себя все существующие, вымершие и искусственно созданные алфавиты мира, а также множество математических, музыкальных, химических и прочих символов.

**Внутреннее представление слов в памяти компьютера**

**с помощью таблицы ASCII**

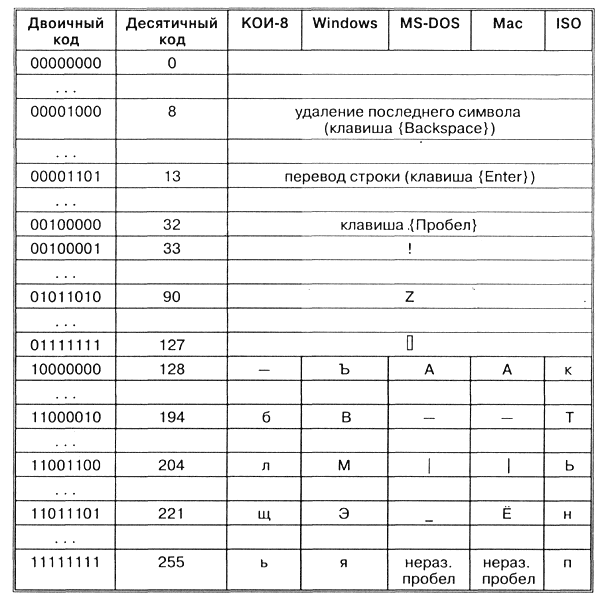
|  |  |
| --- | --- |
| **Слова** | **Память** |
| file | 01100110  01101001  01101100  01100101 |
| disk | 01100100  01101001  01110011  01101011 |

Иногда бывает так, что текст, состоящий из букв русского алфавита, полученный с другого компьютера, невозможно прочитать - на экране монитора видна какая-то "абракадабра". Это происходит оттого, что на компьютерах применяется разная кодировка символов русского языка.



Таким образом, каждая кодировка задается своей собственной кодовой таблицей. Как видно из таблицы, одному и тому же двоич­ному коду в различных кодировках поставлены в соответ­ствие различные символы.

Например, последовательность числовых кодов 221, 194, 204 в кодировке СР1251 образует слово «ЭВМ», тогда как в других кодировках это будет бессмысленный набор символов.

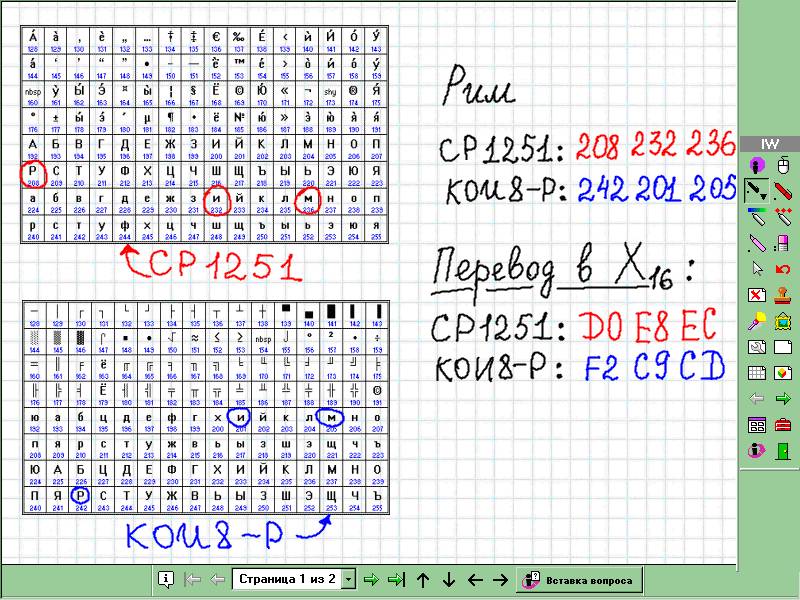


К счастью, в большинстве случаев пользователь не дол­жен заботиться о перекодировках текстовых документов, так как это делают специальные программы-конверторы, встроенные в приложения.

**V. Расчет количества текстовой информации**

Задача 1: Закодируйте слово “Рим” с помощью таблиц кодировок КОИ8-Р и CP1251.

*Решение:*



Задача 2: Считая, что каждый символ кодируется одним байтом, оцените информационный объем следующего предложения:

“Мой дядя самых честных правил,

Когда не в шутку занемог,

Он уважать себя заставил

И лучше выдумать не мог.”

*Решение:* В данной фразе 108 символов, учитывая знаки препинания, кавычки и пробелы. Умножаем это количество на 8 бит. Получаем 108\*8=864 бита.

Задача 3: Два текста содержат одинаковое количество символов. Первый текст записан на русском языке, а второй на языке племени нагури, алфавит которого состоит из 16 символов. Чей текст несет большее количество информации?

*Решение:*

1) I = К \* а (информационный объем текста равен произведению числа символов на информационный вес одного символа).

2) Т.к. оба текста имеют одинаковое число символов (К), то разница зависит от информативности одного символа алфавита (а).

3) 2а1 = 32, т.е. а1 = 5 бит, 2а2 = 16, т.е. а2 = 4 бит.

4) I1 = К \* 5 бит, I2 = К \* 4 бит.

5) Значит, текст, записанный на русском языке в 5/4 раза несет больше информации.

**Задача 4:** Объем сообщения, содержащего 2048 символов, составил 1/512 часть Мбайта. Определить мощность алфавита.

*Решение:*

1) I = 1/512 \* 1024 \* 1024 \* 8 = 16384 бит – перевели в биты информационный объем сообщения.

2) а = I / К = 16384 /1024 =16 бит – приходится на один символ алфавита.

3) 2\*16\*2048 = 65536 символов – мощность использованного алфавита.

Задача 5: Лазерный принтер Canon LBP печатает со скоростью в среднем 6,3 Кбит в секунду. Сколько времени понадобится для распечатки 8-ми страничного документа, если известно, что на одной странице в среднем по 45 строк, в строке 70 символов (1 символ – 1 байт)?

*Решение:*

1) Находим количество информации, содержащейся на 1 странице: 45 \* 70 \* 8 бит = 25200 бит

2) Находим количество информации на 8 страницах: 25200 \* 8 = 201600 бит

3) Приводим к единым единицам измерения. Для этого Мбиты переводим в биты: 6,3\*1024=6451,2 бит/сек.

4) Находим время печати: 201600: 6451,2 =31 секунда.

**Список используемой литературы**

1. Агеев В.М. Теория информации и кодирования: дискретизация и кодирование измерительной информации. — М.: МАИ, 1977.
2. Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. — Киев, Вища школа, 1986.
3. Простейшие методы шифрования текста/ Д.М. Златопольский. – М.: Чистые пруды, 2007 – 32 с.
4. Угринович Н.Д. Информатика и информационные технологии. Учебник для 10-11 классов / Н.Д.Угринович. – М.: БИНОМ. Лаборатория знаний, 2003. – 512 с.
5. <http://school497.spb.edu.ru/uchint002/les10/les.html#n>