Кодирование информации

§ 4. Язык – средство кодирования

§ 5. Дискретное кодирование

§ 6. Кодирование с обнаружением ошибок

© К.Ю. Поляков, Е.А. Ерёмин, 2018

Кодирование информации

§ 4. Язык – средство кодирования

Определения

Кодирование — это представление информации в форме, пригодной для её хранения, передачи и автоматической обработки.

Код — это правило, по которому сообщение преобразуется в цепочку знаков.

Язык — это система знаков и правил, используемая для записи и передачи информации.

Естественные языки – сформировались в результате развития общества.

Иероглифы

Египетское письмо	
	рука
	дом
	кобра
	лев
	вода

Иероглифы (Китай)	
日	солнце
月	луна
雨	дождь
山	гора
马	лошадь

Алфавитное письмо

Алфавит — это набор знаков, который используется в языке.

Мощность алфавита — это количество знаков в алфавите.

Какова мощность русского алфавита? латинского?

АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ

0123456789 .,;?!-:…«»()

мощность 56

Какие бывают языки?

•русский •английский •китайский •шведский •суахили •…	$y = 3\sin x + 1$ $2H_2 + O_2 = 2H_2O$ 1. **e2-e4 e7-e5…**

Формальный язык – это язык, в котором однозначно определяется значение каждого слова, а также правила построения предложений и придания им смысла.

Сообщения

Сообщение — это любая последовательность символов некоторого алфавита.

Сколько различных сообщений длины L можно построить, используя алфавит мощностью M?

Комбинаторика — это наука, изучающая комбинации объектов.

Пример: алфавит {0, 1}.

Сообщения длины 2:

00 01 10 11

всего 4

Сообщения

Пример: алфавит {@, #, $, %}.

Сообщения длины 1: **@ # $ %**.

всего 4

Сообщения длины 2:

@@	@#	@$	@%
#@	##	#$	#%
$@	$#	$$	$%
%@	%#	%$	%%

всего 16

Сколько сообщений длины L ?

Количество возможных сообщений

Если алфавит языка состоит из M *символов (имеет мощность* M*),* количество различных сообщений длиной L *знаков равно*

$$N = M^L$$

Сколько

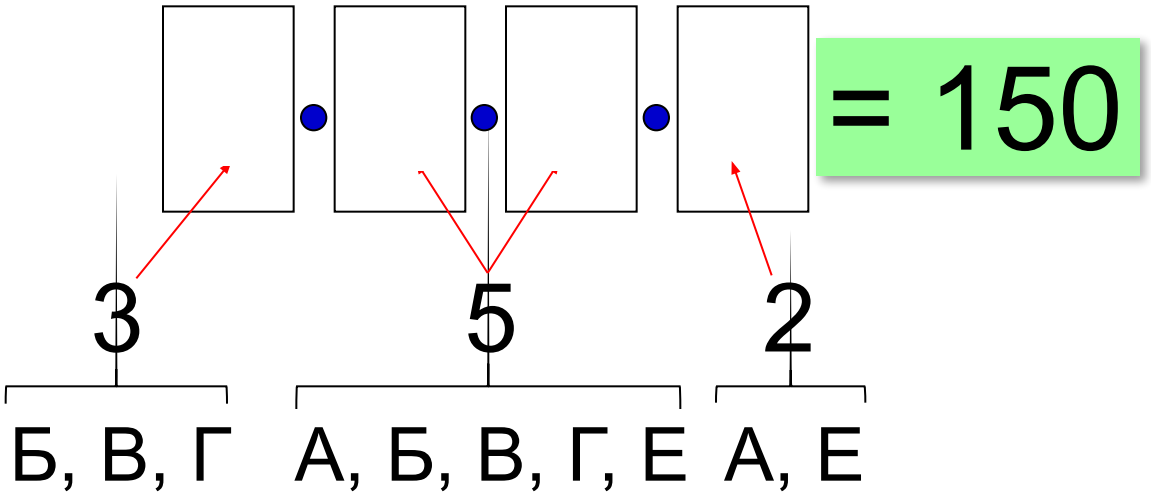
- возможных 5-буквеных слов в русском языке? 33^5
- возможных 3-буквеных слов в английском языке? 26^3
- возможных сообщений длиной L символов в алфавите {+, –}? 2^L

Правило умножения

Задача. Сколько различных сообщений длиной 4 знака можно записать с помощью алфавита

{А, Б, В, Г, Е}

если слова должны начинаться с согласной буквы и заканчиваться на гласную?

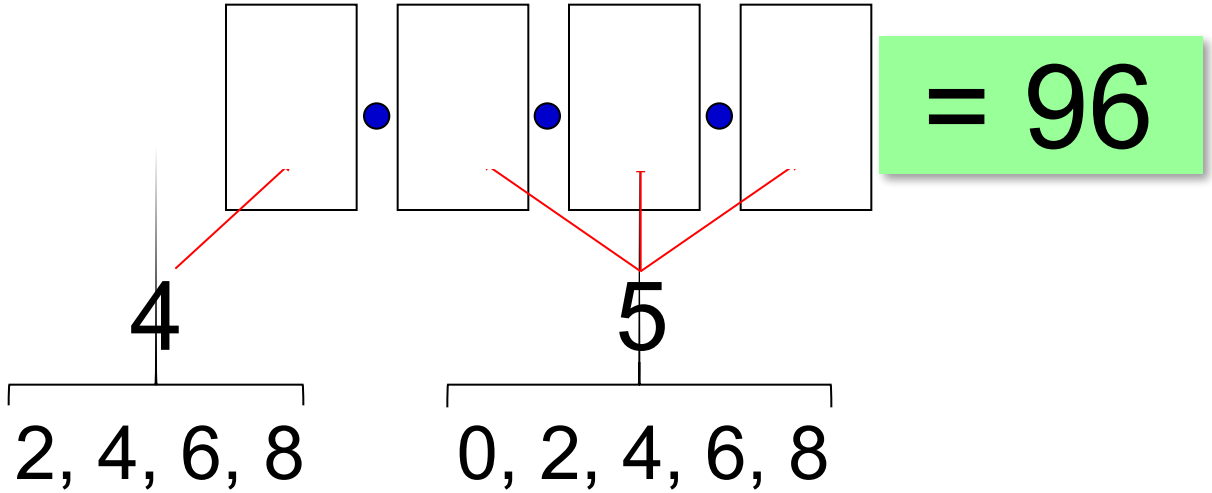


$$N = M_1 \cdot M_2 \cdot M_3 \cdot M_4$$

Правило умножения!

Правило умножения

Задача. Сколько существует четырёхзначных чисел, составленных из чётных цифр, в которых **цифры не повторяются**?



одна цифра уже использована!

Правило сложения

Задача. Сколько сообщений длиной от 2 до 5 символов можно записать с помощью алфавита {0, 1}?

$L = 2$: $N_2 = 2^2 = 4$ $L = 3$: $N_3 = 2^3 = 8$

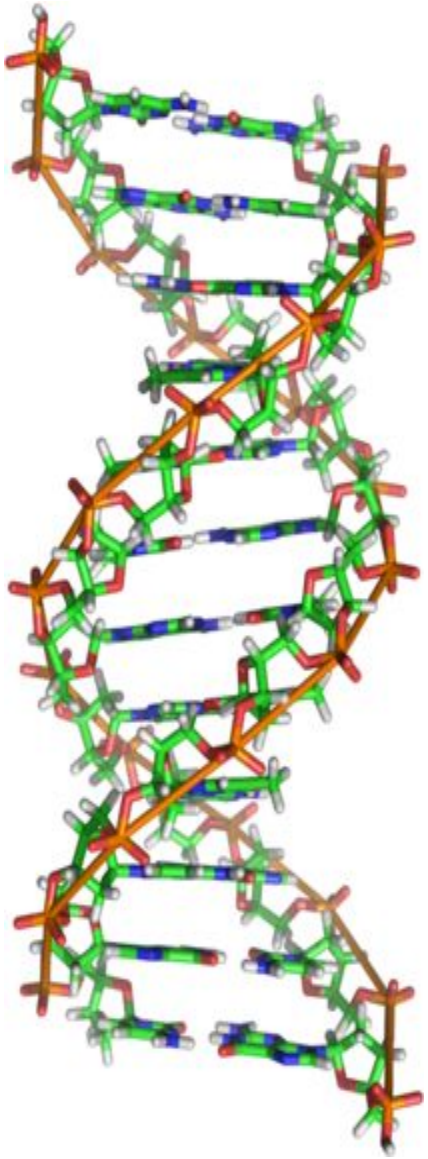
$L = 4$: $N_4 = 2^4 = 16$ $L = 5$: $N_5 = 2^5 = 32$

$$N = 4 + 8 + 16 + 32 = 60$$

$$N = N_2 + N_3 + N_4 + N_5$$

Правило сложения!

Генетический код



Типы звеньев (*нуклеотиды*)

- **A** – аденин (Adenine)
- **C** – цитозин (Cytosine)
- **G** – гуанин (Guanine)
- **T** – тимин (Thymine)

? Мощность алфавита? $M = 4$

3% – **гены** (информация о белка́х)

Белки ← 20 типов аминокислот

? Длина равномерного кода?

$4^2 < 20 < 4^3$ $L = 3$

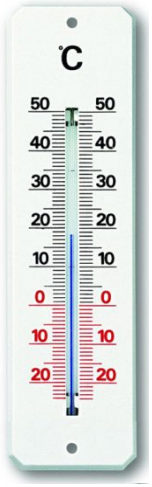
Интеллект-карта

Кодирование информации

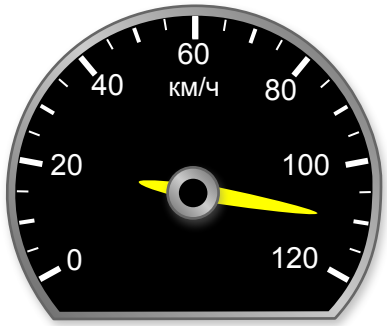
§ 5. Дискретное кодирование

Дискретизация

Дискретизация — это представление единого объекта в виде множества отдельных элементов.



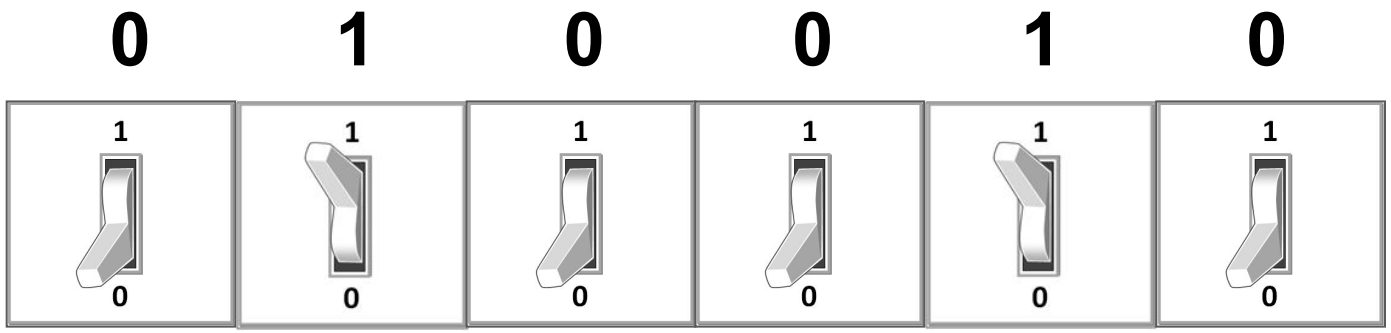
$t = 18°C$



110
км/ч

110,231 км/ч?

Хранение данных в компьютере



Компьютер — это **дискретное** устройство.

Двоичный код – это код, в котором используются два знака (0 и 1). Все данные в компьютере хранятся в двоичном коде.

Бит – это одна двоичная цифра (0 или 1).

010010 ? Сколько бит?

Двоичное кодирование

Кодовая таблица

А	Г	Р
000	010	100

кодовое слово

ГАГАРА: 010 000 010 000 100 000

Равномерный код — это код, в котором все кодовые слова имеют одинаковую длину.

? Сколько существует кодовых слов длиной L в двоичном коде?

2^L

Декодирование

Кодовая таблица

А	Г	Р
000	010	100

?: 100000010100000

Декодирование — это восстановление исходного сообщения из кода.

? Сколько символов было в сообщении? 5

? Как разбить на кодовые слова? по 3

100 000 010 100 000

Р А Г Р А

Как выбрать длину кодовых слов?

Задача. В сообщении встречаются 25 символов. Выберите минимальную длину кодовых слов, при которой все они могут получить разные коды.

1 бит: 2 варианта **< 25**

2 бита: 4 варианта **< 25**

3 бита: 8 вариантов **< 25**

4 бита: 16 вариантов **< 25**

5 битов: 32 варианта

$2^L \geq 25$

Выбор длины кодовых слов L: $M^L \geq M_0$, где M_0 — мощность алфавита исходного сообщения и M — мощность нового алфавита.

Неравномерные коды

Недостаток равномерных кодов – длинные закодированные сообщения.

Идея: часто встречающиеся символы должны иметь более короткие коды!

Код Морзе для русских букв и цифр

А	•–	О	– – –	Э	••–••
Б	–•••	П	•– –•	Ю	••– –
В	•– –	Р	•–•	Я	•–•–
Г	– –•	С	•••		
Д	–••	Т	–	1	•– – – –
Е	•	У	••–	2	••– – –
Ж	•••–	Ф	••–•	3	•••– –
З	– –••	Х	••••	4	••••–
И	••	Ц	–•–•	5	•••••

С. Морзе
(1791–1872)

Неравномерные коды

Кодовая таблица

А	Г	Р
0	1	10

ГАГАРА: 1 0 1 0 10 0

Неравномерный код — это код, в котором кодовые слова имеют разную длину.

Декодирование: 010010

АРАР АГААР АРАГА

АГААГА

Как выделить кодовые слова?

Не всегда однозначно!

Код Морзе

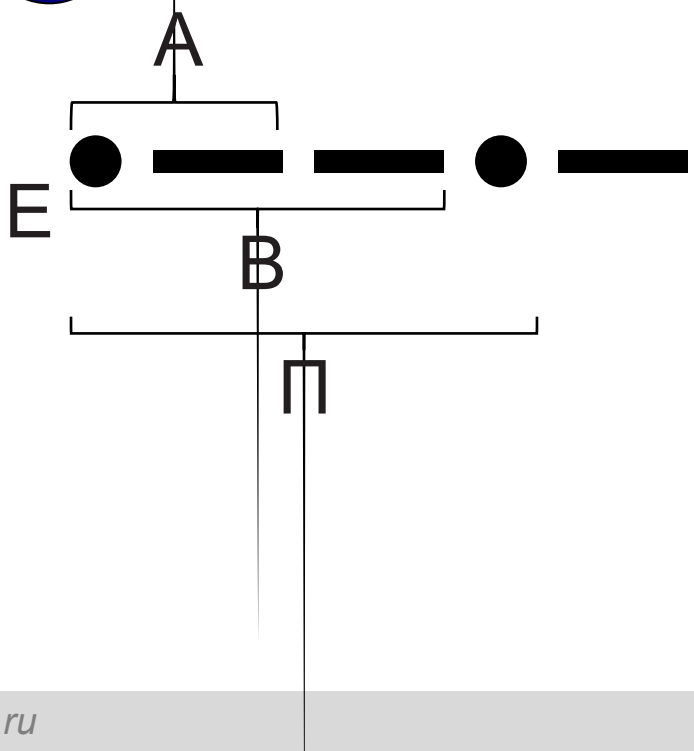
А	•–	О	– – –	Э	••–••
Б	–•••	П	•– –•	Ю	••– –
В	•– –	Р	•–•	Я	•–•–
Г	– –•	С	•••		
Д	–••	Т	–		
Е	•	У	••–		
Ж	•••–	Ф	••–•		
З	– –••	Х	••••		
И	••	Ц	–•–•		
Й	•– – –	Ч	– – –•		
К	–•–	Ш	– – – –		
Л	•–••	Щ	– –•–		
М	– –	Ь	–•• –		
Н	–•	Ы	–•– –		

Условие Фано?

НЕТ

Нужна пауза!

Как декодировать?



Неравномерные коды

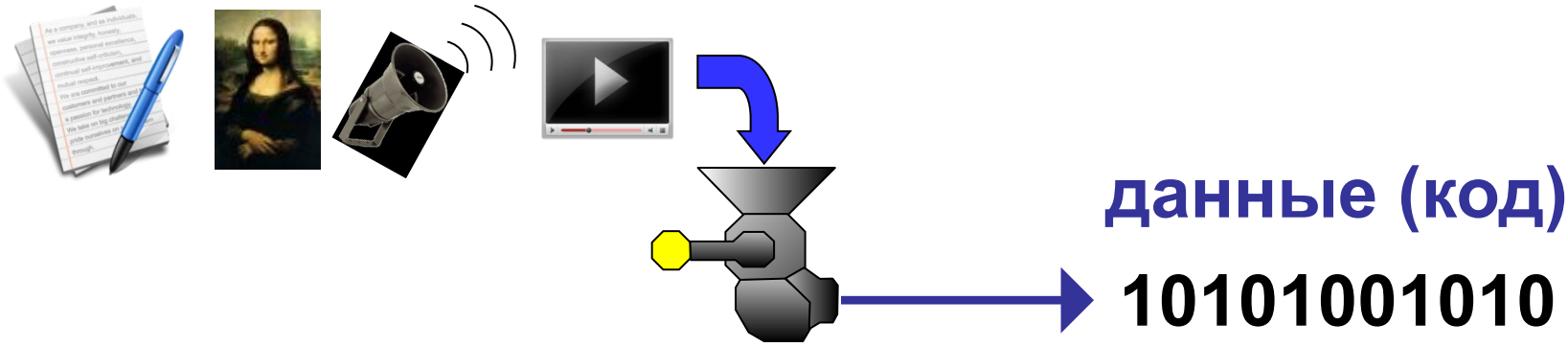
Кодовая таблица

А	Г	Р
0	10	11

Декодирование: 01001011 → **АГАГР**

Неравномерный код декодируется однозначно, если выполняется **условие Фано**: ни одно кодовое слово не совпадает с началом другого кодового слова.

Как измерить информацию?



Количество информации в битах определяется длиной сообщения в двоичном коде.

10101100 — 8 битов

Единицы измерения

2^{10}

1 **байт** = 8 бит

1 **Кбайт** (килобайт) = 1024 байта

1 **Мбайт** (мегабайт) = 1024 Кбайт

1 **Гбайт** (гигабайт) = 1024 Мбайт

1 **Тбайт** (терабайт) = 1024 Гбайт

1 **байт** = 2^3 бит

1 **Кбайт** = 2^{10} байта = $2^{10} \cdot 2^3$ бит = 2^{13} бит

1 **Мбайт** = 2^{10} Кбайт = $2^{10} \cdot 2^{13}$ бит = 2^{23} бит

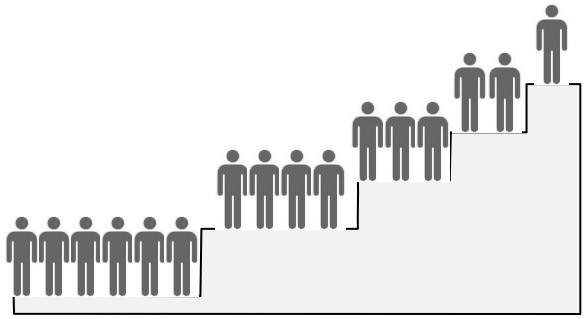
Перевод в другие единицы

2 Кбайт = 2 × (1 Кбайт) = 2 × 1024 байт
= 2048 байт
= 2048 × (1 байт) = 2048 × 8 бит
= 16 384 бита

Через степени числа 2:

2 Кбайт = 2×2^{10} байт = 2^{11} байт
= $2^{11} \times 2^3$ бит = 2^{14} бит.

Перевод в другие единицы



1 бит

× 8

1 байт

: 8

× 1024

1 Кбайт

: 1024

× 1024

1 Мбайт

: 1024

× 1024

1 Гбайт

: 1024

1 Тбайт

число уменьшается

1 байт = 8 бит

1 Кбайт = 1024 байта

число увеличивается

Алфавитный подход

Задача 1. Алфавит русского языка содержит 33 символа. Определите наименьшую длину кодовых слов при кодировании сообщений на русском языке с помощью равномерного кода.

$M = 33$	i бит $\rightarrow 2^i$ разных кодов $M \le 2^i$
$i = ?$	$2^5 < 33 \le 2^6$

6 бит на символ хватает!

5 бит на символ не хватает…

Ответ: $i = 6$ бит

? Если различать заглавные и строчные буквы?

$i = 7$ бит

Алфавитный подход

Задача 2. Текст длиной 160 символов записан с помощью алфавита из 26 символов. Определите количество информации в сообщении, закодированном с помощью равномерного кода наименьшей длины.

$L = 160$

$M = 26$

$I = ?$

$I = L \cdot i$

$2^4 < 26 \le 2^5$

$i = 5$

бит на символ

5 бит на символ хватает!

$I = 160 \cdot 5 = 800$ бит

В байтах?

$I = 800 : 8 = 100$ байт

Ответ: $I = 800$ бит $= 100$ байт

Алфавитный подход

Задача 3. Пароль длиной 8 символов может содержать английские буквы (заглавные и строчные), цифры и специальные знаки: @, #, $, %. Сколько бит памяти нужно выделить для хранения пароля?

$L = 8$

$M = 26 \cdot 2 + 10 + 4 = 66$

$I = ?$

$I = L \cdot i$

$2^6 < 66 \le 2^7$

$i = 7$

7 бит на символ хватает!

$I = 8 \cdot 7 = 56$ бит

В байтах?

$I = 56 : 8 = 7$ байт

Ответ: $I = 56$ бит $= 7$ байт

Алфавитный подход

Задача 4. Текст длиной 4096 символов занимает в памяти 4 Кбайта. Определите наибольшее возможное количество символов в алфавите.

$L = 4096$
$I = 4$ Кбайт

$M = ?$

i бит $\rightarrow 2^i$ разных кодов $M \le 2^i$

Как найти i?

$I = L \cdot i$ $i = I : L$

$i = 4 : 4096$

Все ли верно?

$i = 4 \cdot 1024 \cdot 8 : 4096 = 8$ бит

$M \le 2^8 = 256$

Ответ: $M = 256$

Алфавитный подход

Задача 5. Участники соревнований по бегу получили номера от 1 до 100. На финише автоматическое устройство записывает номер спортсмена. Сколько байт нужно для хранения номеров 80 спортсменов?

$M = 100$
$L = 80$
$I = ?$

i бит $\longrightarrow 2^i$ разных кодов $M \le 2^i$

$2^6 < 100 \le 2^7$

7 бит на символ хватает!

$I = L \cdot i$

$I = 80 \cdot 7 = 560$ бит

? В байтах?

$I = 560 : 8 = 70$ байт

Ответ: $I = 70$ байт

Кодирование информации

§ 7. Кодирование с обнаружением ошибок

© К.Ю. Поляков, Е.А. Ерёмин, 2018

Обнаружение ошибок

10010 ? Верно ли переданы данные?

Бит чётности:

00 01 10 11 ⇒ 000 011 101 110

теперь число единиц в каждом блоке чётное

Если в принятом блоке нечётное число «1» – **ошибка**!

принято: **010** 110 000 **111** 000

? Можно ли исправить?

Для файлов – контрольные суммы (хэш):

CRC = *Cyclic Redundancy Code*

MD5, SHA-1

Обнаружение ошибок

К	А	Р	П
00	01	10	11

к коду каждой буквы справа добавляется бит чётности

Получено (с битами чётности):

1011110001000110000

Разбиваем на группы по **3 бита**:

101 111 000 100 011 000

Помечаем блоки с ошибками:

101 * 000 * 011 000

Убираем бит чётности в каждом блоке (последний):

10 * 00 * 01 00

Декодируем по таблице:

Р * К * А К

? Сколько бит в блоке?

Исправление ошибок

10010

111 **0**00 **0**00 **1**11 **0**00 – утроение каждого бита

принято: **010**111000**101**000

исправлено: **000**111000**111**000

Обнаруживает 1 или 2 ошибки, исправляет 1 ошибку!

Помехоустойчивый код – это код, который позволяет исправлять ошибки, если их количество не превышает некоторого уровня.

Исправление ошибок

П	О	Р	Т
11111	11000	00100	00011

Каждое кодовое слово отличается от остальных не менее, чем в 3 битах!

```
П = 11111    11111    11111
    11000 3  00100 4  00011 3
      О        Р        Т
```

Расстояние Хэмминга

Как декодировать?

```
10011 11100 00000
00011 11000 11100
  Т     О     Р
```

Исправление ошибок

П	О	Р	Т
11111	11000	00100	00011

? Как декодировать?

```
10101 11001 01001
  ?   11000   ?
  *     O     *
```

Конец фильма

ПОЛЯКОВ Константин Юрьевич
д.т.н., учитель информатики
ГБОУ СОШ № 163, г. Санкт-Петербург
kpolyakov@mail.ru

ЕРЕМИН Евгений Александрович
к.ф.-м.н., доцент кафедры мультимедийной дидактики и ИТО ПГГПУ, г. Пермь
eremin@pspu.ac.ru

Источники иллюстраций

1. http://fpg.unc.edu
2. http://s1.iconbird.com
3. https://sandstorm.deviantart.com
4. http://http://compression.ru
5. http://ru.wikipedia.org
6. https://www.kns.ru
7. http://nix.ru
8. http://http://www.computer-services.ru
9. http://http://www.masterna4as.com
10. http://blendercontest.com
11. http://http://geeky-gadgets.com
12. авторские материалы