Сжатие — это кодирование с уменьшением объема данных и возможностью однозначного декодирования.

Обратный процесс — декодирование — называется разжатие. Другие названия: *компрессия/декомпрессия*, *упаковка/распаковка*.

Эффективность алгоритма сжатия зависит от

- степени сжатия (отношение длины несжатых данных к длине соответствующих им сжатых данных);
- скорости сжатия и разжатия;
- объема памяти, необходимого для работы алгоритмов и т.д

- Сжатие без потерь (*lossless compression*) – собственно сжатие в смысле приведенного определения.

- Сжатие с потерями (*lossy compression*) – процесс, состоящих из двух этапов:

1. выделение сохраняемой части информации в зависимости от цели сжатия и особенностей приемника и источника;
2. собственно сжатие без потерь.

Кодирование длин повторов, ***Run Length Encoding*** (RLE, групповое кодирование)

- Один из наиболее старых методов сжатия, идея метода состоит в замене идущих подряд одинаковых символов (бит или байт) парой (количество, символ).
- В основном используется для кодирования растровых изображений.
- Характеристика: степень сжатия от 0,5 до 32.
- графические файлы jpeg, tiff

- **Групповой код А** задает количество нулевых и единичных значений в порядке их следования.
- **Групповой код В** задает индексы границ единичных участков.

0000 0000 1111 1000 0000 0000 0111 0000 0001 1111 1111 0000

A: 8(0) 5(1) 12(0) 3(1) 7(0) 9(1) 4(0)
B: (8,12) (25,27) (35,43)

Задание

- Построить коды A и B для изображения 011 110 000 111 011 111

Алгоритмы Зива-Лемпела (**LZ-методы**)

- сообщение кодируется не побуквенно (алфавитное кодирование), а по словам.
- Характеристики: степень сжатия в зависимости от данных, обычно 2-3;
- алгоритмы универсальны, но лучше всего подходят для сжатия текстов, рисованных картинок или других однородных данных
- архиваторы (форматы rar, zip, arj, cab, ace); графические файлы gif, tiff

010 001 011 001 010 001 101 011 00

Словарь:

{Λ, 0, 1, 00, 01, 011, 001, 010, 0011, 0101, 10}

0 1 2 3 4 5 6 7 8 9 10

0 1 00 01 011 001 010 0011 0101 10 0

(0, 0), (0, 1), (1, 0), (1, 1), (4, 1), (3, 1), (4, 0), (6, 1), (7, 1), (2, 0), (0, 0)

Задание

Закодируйте текст

1). 010 010 001

2). aba adb abc ecd ebc ea

3). 001 101 110 010 100 110 100 010 111 001 010 011 010 110 100

Задание

Раскодируйте текст

1) (0,0), (0,1), (2,0), (3,1), (2,1), (1,1), (4,1), (7,1), (6,0), (1,0), (9,1), (2,0)

2) (0,0), (0,1), (2,1), (2,0), (1,0), (3,0), (6,0), (7,1), (1,1), (9,1), (5,0),(0,1)

3) (0,γ), (0, α), (0,β), (1,γ), (2,β), (2,δ), (0,δ), (0,γ)

Арифметическое сжатие (**ARIC, Arithmetic Coding**)

- Характеристики: один из самых эффективных методов;
- степень сжатия от 1 до 8, т.е. не увеличивает размер данных в худшем случае;
- Не является алфавитным кодированием.
- Весь кодируемый текст представляется в виде дроби из [0, 1).

Пусть x = математика

y = ***мате***

символ	частота	вероятность	диапазон
а	3	0,3	[0; 0,3)
м	2	0,2	[0,3; 0,5)
т	2	0,2	[0,5; 0,7)
е	1	0,1	[0,7; 0,8)
и	1	0,1	[0,8; 0,9)
к	1	0,1	[0,9; 1)

текущий символ	рабочий интервал	длина интервала	1/10 длины
-	[0;1)	1	0,1
м [0,3; 0,5)	[0,3; 0,5)	0,2	0,02
а [0; 0,3)	[0,3; 0,36)	0,06	0,006
т [0,5; 0,7)	[0,33; 0,342)	0,012	0,0012
е [0,7; 0,8)	[0,3384; 0,3396)	0,0012	0,00012

$y^* = 0{,}339$

Задание

Выполнить декомпрессию кода y = 0.75, используя таблицу диапазонов, если известно, что длина сообщения 10 символов.

символ	частота	вероятность	диапазон
а	3	0,3	[0; 0,3)
м	2	0,2	[0,3; 0,5)
т	2	0,2	[0,5; 0,7)
е	1	0,1	[0,7; 0,8)
и	1	0,1	[0,8; 0,9)
к	1	0,1	[0,9; 1)

Задание

Закодировать первые четыре символа сообщения x = ”**ков.корова**”:

1) составить таблицу частот и диапазонов всех символов сообщения,
2) найти рабочий интервал для ”ков.” и выбрать число y – код слова,
3) найти рабочий интервал и код для слова ” кова” (использовать таблицу диапазонов из предыдущего задания),
4) рассмотреть процесс декомпрессии (восстановления слова ”ков.” по числу y).