Алгоритми стиснення інформації

Дисципліна

Інформаційні технології

Лекція 3

Базові відомості

Стиснення інформації дозволяє зменшити час передавання повідомлення і його об'єм

Трансляція аудіо файлу, оціфрованого за допомогою 16 розрядного АЦП (стерео запис) потребує швидкості 2*16*44100 біт/сек = **1 411 200 біт/сек**

Трансляція відео файлу, зображення в якому утворюється за допомогою технології RGB, потребує швидкості передавання

25*1280*1024*3*8 = **786 432 000 біт/сек**

Основна теорема Шеннона для каналу без завад

Повідомлення, складене з букв деякої абетки, можна закодувати таким чином, що середня кількість двійкових символів на букву буде наближена до ентропії джерела повідомлення, але не менша за цю величину

Ентропія Шеннона

- $$H = \sum_{i=1}^{N} p_i \, log_2 \frac{1}{p_i}$$

«A Mathematical Theory of Communication»

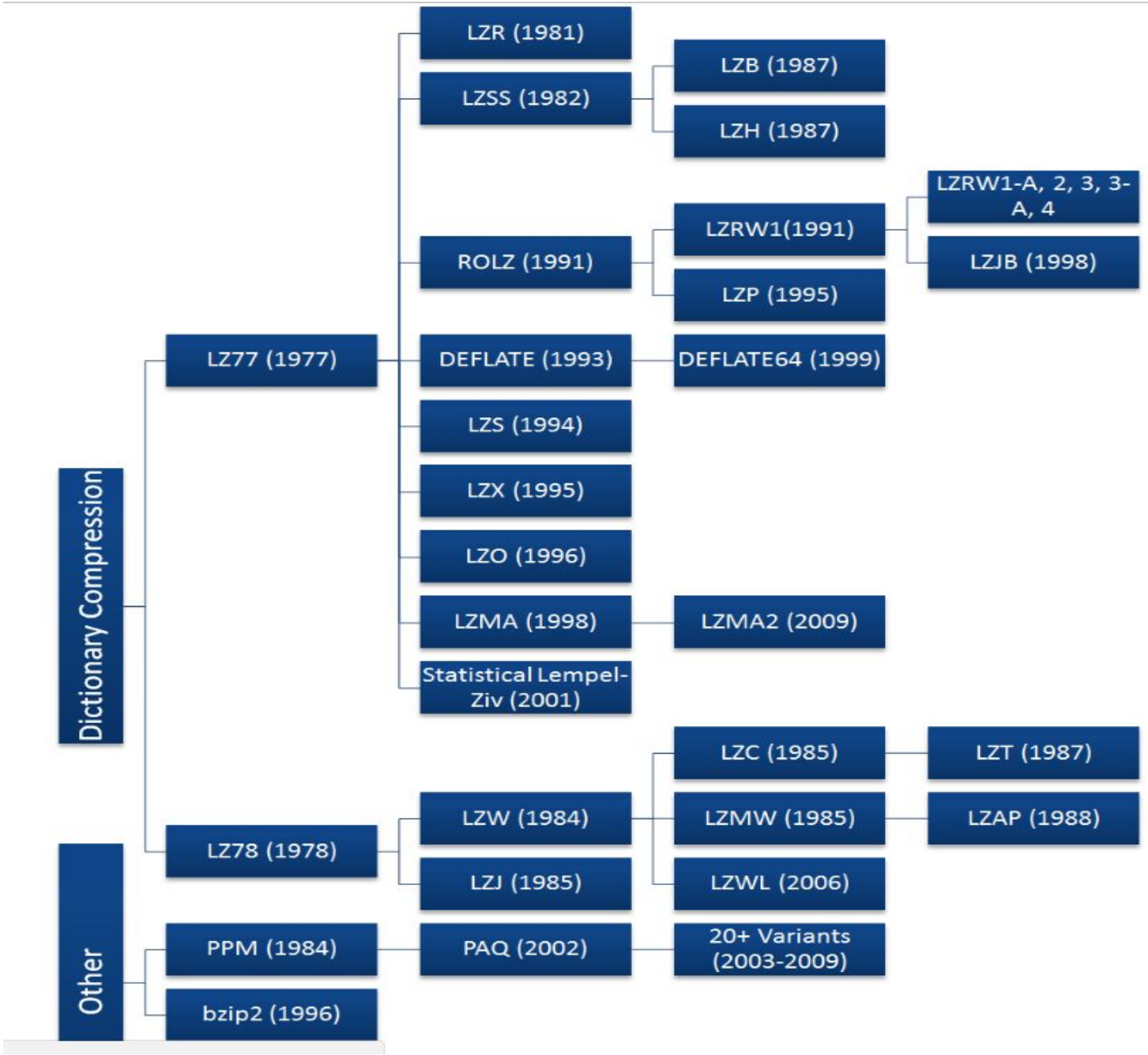
Bell System Technical Journal, 1948

Середнє число символів на букву

$$L = \sum_{i=1}^{N} p_i \, n_i$$

L-H надлишковість коду

(L-H)/L відносна надлишковість коду


Dictionary Compression
Other
LZ77 (1977)
LZ78 (1978)
PPM (1984)
bzip2 (1996)
LZR (1981)
LZSS (1982)
LZB (1987)
LZH (1987)
ROLZ (1991)
LZRW1(1991)
LZRW1-A, 2, 3, 3-A, 4
LZJB (1998)
LZP (1995)
DEFLATE (1993)
DEFLATE64 (1999)
LZS (1994)
LZX (1995)
LZO (1996)
LZMA (1998)
LZMA2 (2009)
Statistical Lempel-Ziv (2001)
LZW (1984)
LZJ (1985)
LZC (1985)
LZT (1987)
LZMW (1985)
LZAP (1988)
LZWL (2006)
PAQ (2002)
20+ Variants (2003-2009)

Методи стискання інформації

Стискання без втрат:

Шеннона-Фано (Shannon-Fano)

Хаффмена (Huffman)

Лемпела-Зіва-Велча (LZW)

Арифметичне стискання

Стискання з втратами:

MPEG (Moving Pictues Experts Group)

JPEG (Joint Photographic Expert Group)

Фрактальне стискання

Алгоритми стиснення без втрат

- Стиснення способом кодування серій

  (RLE - Run Length Encoding)

  44 44 44 11 11 11 11 11 01 33 FF 22 22

  вихідна послідовність

  03 **44** ***05*** **11** ***00*** ***03*** **01 33 FF** ***02*** **22**

  стисла послідовність

Дані методи, як правило, досить ефективні для стиснення растрових графічних зображень (BMP, PCX, TIF, GIF), тому що останні містять досить багато довгих серій повторюваних послідовностей байтів. Недоліком методу RLE є досить низька ступінь стиснення.

Стискання за алгоритмом Шеннона – Фано

Літерний текст (повідомлення) кодується у двійковому коді, як правило, при цьому кожній літері надається однакова довжина коду. Але одні літери зустрічаються частіше (вони популярні) ніж інші. Тому якщо популярним літерам приписати більш короткий код, ніж іншим, то сумарна довжина коду зменшиться.

Задача полягає у тому, щоб ефективно перекодувати текст - скласти таблицю кодів для літер повідомлення.

Алгоритм:

- - літери абетки повідомлення записуються в таблицю у порядку зменшення ймовірностей;
- - літери абетки розділяються на дві групи таким чином, щоб суми ймовірностей у кожній групі були по можливості однакові;
- - всім літерам верхньої половини в якості першого символу приписується одиниця, а всім нижним – нулі;
- - кожна з отриманих груп у свою чергу розділяється на дві підгрупи з однаковими сумарними ймовірностями і т.д., процес повторюється до того моменту, доки в кожній підгрупі залишиться по одній літері.

Приклад. Розглянемо текст (**to be or not to be**). Випишемо за порядком літери, як вони зустрічаються у тексті і підрахуємо їх кількість: **t** – 3, **o** – 4, **b** – 2, **e** – 2, **r** – 1, **n** – 1. Всього нарахували 6 різних літер, а довжина фрази складає 13 літер (пробіли не враховуємо). Ймовірність для кожної літери розраховуємо таким чином: для t це 3/13, для o це 4/13 і т.д. Заповнимо таблицю кодів згідно алгоритму Шеннона-Фано

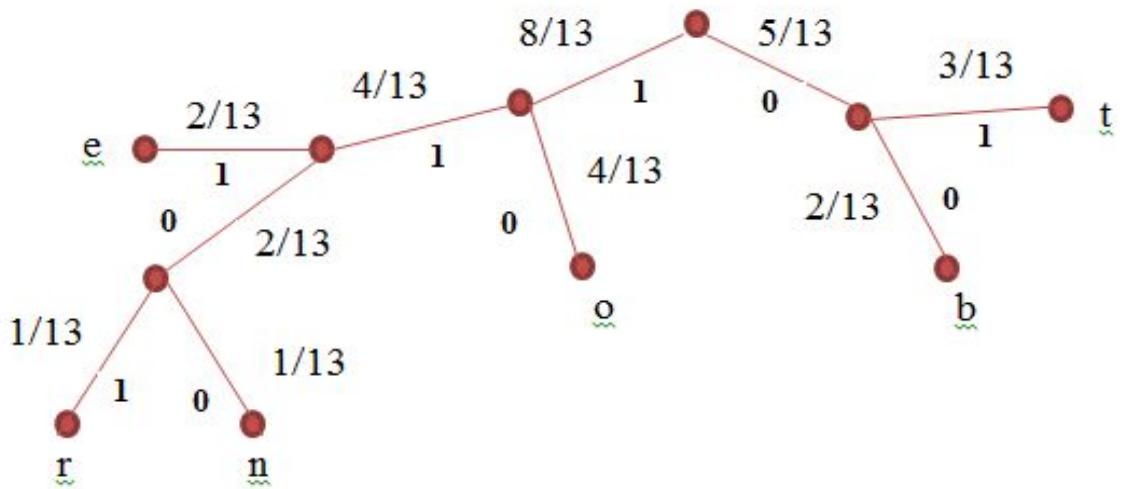
Літера	Ймовірність	Процес отримання коду			Код Шеннона-Фано
o	4/13	1	1		11
t	3/13	1	0		10
b	2/13	0	1	1	011
e	2/13	0	1	0	010
r	1/13	0	0	1	001
n	1/13	0	0	0	000

Стискання за алгоритмом Хаффмена

Алгоритм:

- літери повідомлення записуються в таблицю у порядку зменшення ймовірностей;
- дві останні літери об’єднуються в одну допоміжну літеру, для якої записують сумарну ймовірність;
- ймовірності літер, які не приймали участь у об’єднанні, і отримана сумарна ймовірність знову розташовуються у порядку зменшення ймовірностей, а дві останні об’єднують до тих пір, доки не отримують одну допоміжну літеру з ймовірністю одиниця;
- далі для побудови коду використовують бінарне дерево, у корені якого розташовується літера з ймовірністю одиниця, при розгалуженні гілці з більшою ймовірністю присвоюється код одиниця, а гілці з меншою ймовірністю – код нуль (або лівій гілці – один, правій – нуль).

Літера	Ймовірність	Допоміжні стовпці ймовірностей	Код Хаффмена
o	4/13	4/13 → 4/13 → 5/13 → 8/13 (1)	10
t	3/13	3/13 → 4/13 → 4/13 → 5/13 (1)	01
b	2/13	2/13 → 3/13 → 4/13 (1)	00
e	2/13	2/13 → 2/13 (1)	111
r	1/13	2/13 (1)	1101
n	1/13		1100



Арифметичне кодування

Розроблений у 1979 році в IBM. Досягає більшого ступеня стиснення, ніж Хаффмена, більш складний у порівнянні з попередніми алгоритмами. Замість розбиття ймовірностей по дереву алгоритм перетворює вхідні дані в одне раціональне число від 0 до 1.

Алгоритм :

- Підраховують кількість унікальних символів на вході. Це число буде представляти основу для числення b (b = 2 - бінарне, і т.п.)
- Підраховують загальну довжину входу
- Призначають «коди» від 0 до (b-1) кожному з унікальних символів в порядку їх появи
- Замінюють символи кодами, отримуючи число в системі числення з основою b
- Записують отримане число в двійковій системі

Приклад. На вході рядок «ABCDAABD»

- 4 унікальних символи, основа = 4, довжина даних = 8
- призначаємо коди: A = 0, B = 1, C = 2, D = 3
- отримуємо число "0.01230013"
- перетворимо «0.01231123» з четверичной у двійкову систему: 0.01101100000111

Якщо маємо справу з восьмібітними символами, то на вході є 8x8 = 64 біта, а на виході є 15 біт.

MPEG: Загальна інформація

- Стандарт стиснення MPEG розроблений Експертної групою кінематографії (Moving Picture Experts Group - MPEG). MPEG це стандарт на стискання звукових і відео файлів в більш зручний для завантаження чи пересилання, наприклад через Інтернет, формат.

- Існують різні стандарти MPEG (як їх ще іноді називають фази - phase): MPEG-1, MPEG-2, MPEG-3, MPEG-4, MPEG-7.

- MPEG складається з трьох частин: Audio, Video, System (об'єднання і синхронізація двох інших).

Як працює MPEG

- Залежно від деяких причин кожен frame (кадр) в MPEG може бути наступного виду:

- **I (Intra)** frame - кодується як звичайна картинка.
- **P (Predicted)** frame - при кодуванні використовується інформація від попередніх I або P кадрів.
- **B (Bidirectional)** frame - при кодуванні використовується інформація від одного або двох I або P кадрів
- Послідовність кадрів може бути наприклад така: IBBPBBPBBPBBIBBPBBPB ...

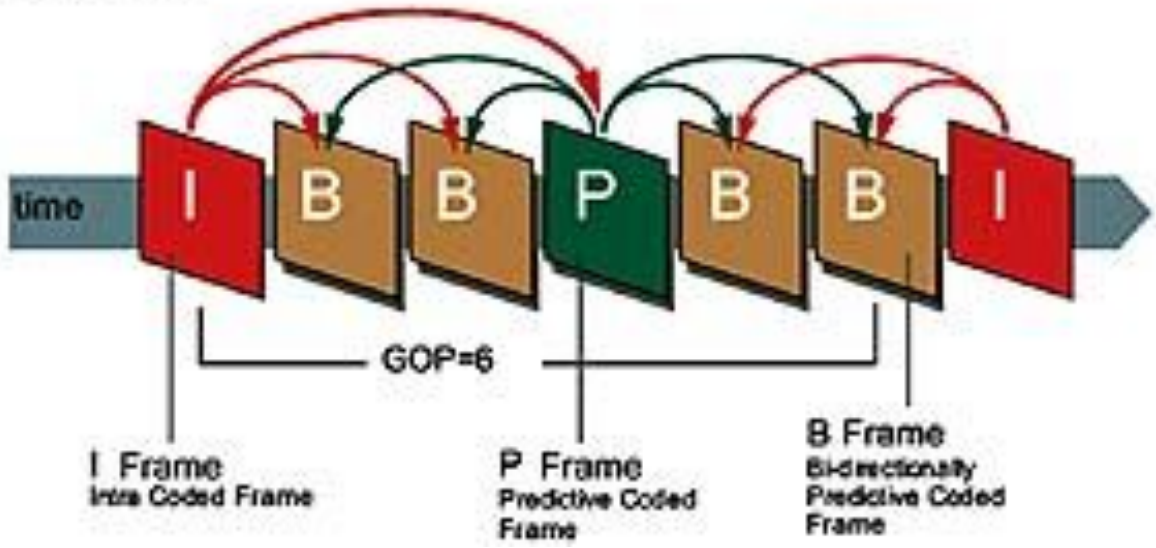
- Послідовність декодування: 0312645 ...

- Потрібно зауважити, що перш ніж декодувати B кадр потрібно декодувати два I або P кадру. Існують різні стандарти на частоту, з якою мають слідувати I кадри, приблизно 1-2 в секунду, відповідні стандарти є і для P кадрів (кожен 3 кадр повинен бути P кадром).

DV

DV codes each frame a time eliminating redundant information inside each frame

MPEG-2



Техніка кодування

Для більшого стиснення в B і P кадрах використовується алгоритм передбачення руху (що дозволяє сильно зменшити розмір P і B кадрів) на виході якого є:

- Вектор зміщення (вектор руху) блоку який потрібно передбачити щодо базового блоку.
- Різниця між блоками (яка потім і кодується).
- Так як не будь-який блок можна передбачити на підставі інформації про попередні, то в P і B кадрах можуть бути I блоки (блоки без передбачення руху).

Метод кодування блоків (або різниці, одержуваної при методі передбачення руху) містить в собі:

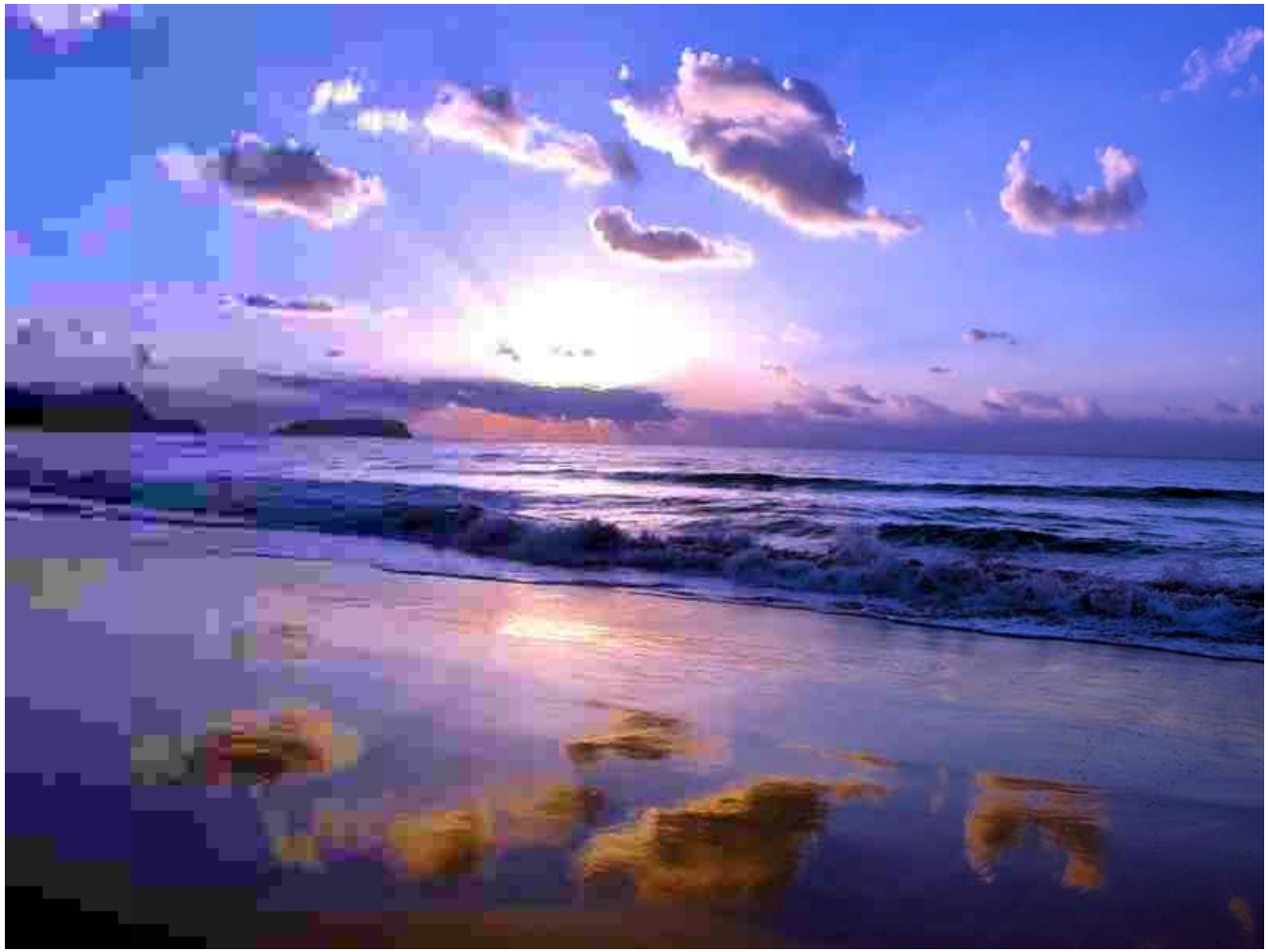
- Discrete Cosine Transforms (DCT - дискретне перетворення косинусів).
- Quantization (перетворення даних з неперервної форми в дискретну).
- Кодування отриманого блоку в послідовність.

JPEG

- Файл з розширенням JPG - це те ж саме, що і JPEG. Термін JPEG насправді - це скорочення від «Спільна група експертів фотографії» (Joint Photographic Experts Group). JPEG являє собою стислий формат файлу зображення. JPEG зображення не обмежені певною кількістю кольорів, як GIF формат. Таким чином, формат JPEG краще підходить для стиснення фотографій.
- JPEG заснований на 24-бітної палітрі і підтримує 16,7 млн. кольорів. Це формат стиснення з втратами. Ступінь стиснення може бути в діапазоні від 10: 1 до 20: 1, і більшість графічних прикладних програм (наприклад, Adobe Photoshop) дозволяють вибрати ступінь стиснення.

- Формат JPEG файлів найкраще підходить для цифрової фотографії, де типова швидкість стиснення з дуже низьким рівнем втрати якості складає близько 10: 1. Як GIF, JPEG - це кросплатформений алгоритм, тобто той же файл буде виглядати так само, як на Mac так і PC.

Фотографія заходу сонця в форматі JPEG зі зменшенням ступеня стиснення зліва направо



JPEG - за і проти

- До недоліків стиснення за стандартом JPEG слід віднести появу на відновлених зображеннях при високих ступенях стиснення характерних артефактів: зображення розсипається на блоки розміром 8x8 пікселів (цей ефект особливо помітний на областях зображення з плавними змінами яскравості), в областях з високою просторовою частотою (наприклад, на контрастних контурах і границях зображення) виникають артефакти у вигляді шумових ореолів. Слід зазначити, що стандарт JPEG (ISO / IEC 10918-1, Annex K, п. K.8) передбачає використання спеціальних фільтрів для придушення блокових артефактів, але на практиці подібні фільтри, незважаючи на їх високу ефективність, практично не використовуються.

- Однак, незважаючи на недоліки, JPEG отримав дуже широке поширення через досить високий (щодо існуючих під час його появи альтернатив) ступень стиснення, підтримки стиснення повнокольорових зображень і відносно невисокої обчислювальної складності.

Перевірка знань

- Записати власне прізвище, ім'я та по-батькові одним рядком без пробілів
- Скласти таблицю кодів (закодувати кожну літеру) за алгоритмом Шеннона-Фано.