

Poročilo prve laboratorijske vaje pri predmetu Informacija in kodi

Rok Prezelj

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: rp0067@student.uni-lj.si

Povzetek. Laboratorijska vaja je bila namenjena analizi različnih datotek, pri čemer se je določila verjetnost pojavljanja 8-bitnih znakov in na njihovi podlagi izračunali entropijo informacijskega vira pri različni dolžini nizov.

1 Uvod

Entropija je merilo nedoločenosti naključnih sistemov, večja kot je, bolj je sistem nepredvidljiv [1]. V informacijski teoriji predstavlja entropija povprečno količino informacije, ki jo vsebuje posamezen simbol vira, in nam omogoča oceno učinkovitosti kodiranja ter stopnje naključnosti v podatkih [2].

2 Metodologija

V okviru naloge smo implementirali program za analizo datotek in izračun entropije za različne dolžine nizov [3]. Program prebere binarno vsebino izbranih datotek in prešteje pojavljanja vseh možnih kombinacij n -bajtnih nizov. Nato izračuna verjetnost ter entropijo po formuli:

$$H_n(X) = -\frac{1}{n} \sum_i p_i \log_2 p_i, \quad (1)$$

kjer je p_i verjetnost posameznega niza, n dolžina nizov v bajtih za $n = [1, 5]$.

3 Rezultati

Z formulo (1) smo analizirali različne oblike podatkov in opazili, da se z višanjem dolžine niza entropija manjša.

3.1 Besedilne datoteke

Za datoteko `besedilo.txt` se entropija od $H_1 = 4.57$ zniža do $H_5 = 3.03$. To pomeni, da ima besedilo precejšnjo redundanco, torej določeni znaki in kombinacije se pojavljajo bistveno pogosteje kot drugi. Takšna odvisnost je posledica jezikovne strukture (črke, presledki, pogosti nizi znakov), kar omogoča zelo učinkovito brezgubno kompresijo (npr. ZIP [4], Huffman [6], LZW [5]).

3.2 Zvočni posnetki

Zvočne datoteke v surovih formatih (`.wav`, `.aiff`, `.raw`) imajo skoraj enake entropije:

$$H_1 \approx 6.35, \quad H_5 \approx 4.27.$$

To pomeni, da vse tri vsebinsko hranijo iste vzorce in da zaglavje (header) pri WAV ali AIFF nima bistvenega vpliva.

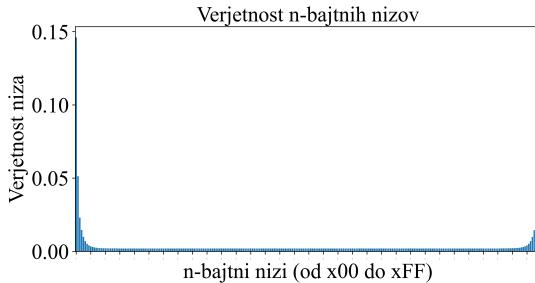
Pri kompresiranih oblikah (`.flac`, `.m4a`, `.mp3`, `.ogg`) pa opazimo naslednje:

| Format | H_1 | H_5 |
|------------------|-------|-------|
| WAV / RAW / AIFF | 6.35 | 4.27 |
| FLAC | 7.95 | 4.74 |
| M4A | 7.97 | 4.75 |
| MP3 | 7.93 | 4.46 |
| OGG | 7.98 | 4.22 |

Tabela 1: Entropije zvočnih posnetkov v različnih formatih.

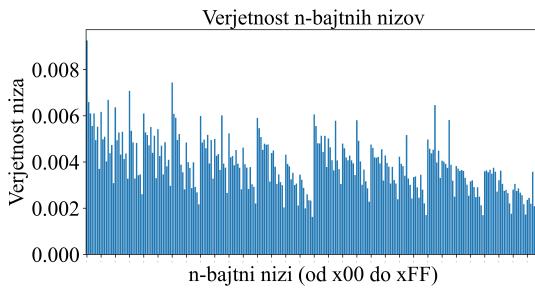
Pri formatu OGG, ki je definiran v standardu rfc3533 [7], opazimo, da ima H_1 skoraj maksimalno vrednost, kar potrjuje učinkovito odstranitev redundancy med posameznimi bajti. Kljub temu pa ima H_5 nekoliko nižjo vrednost kot nekompresirani formati, kar kaže na prisotnost ponavljajočih se struktur v kodnem zapisu (paketno kodiranje, zaglavja, metapodatki). Takšne periodične vzorce algoritmi kompresije vnesejo zaradi organizacije podatkov, ne zaradi vsebinske redundancy.

Na slikah 1 in 2 sta prikazani porazdelitvi verjetnosti posameznih bajtov v nekompresiranem (AIFF/RAW/WAV) in kompresiranem (MP3) zvočnem zapisu. Pri formatu AIFF je opazna značilna oblika črke U , saj prevladujeta vrednosti bajtov `0x00` in `0xFF`. To nakazuje, da v zapisu pogosto nastopajo vzorci z nizkimi ali visokimi amplitudami, kar je značilno za PCM-kodiranje [8], kjer vrednosti neposredno predstavljajo amplitudo signala.



Slika 1: Porazdelitev verjetnosti bajtov v nekompresiranem zapisu AIFF. Značilna oblika črke U kaže na pogosti vrednosti $0x00$ in $0xFF$.

Pri formatu MP3 je porazdelitev verjetnosti bajtov bolj enakomerna in ima žagasto obliko. To je posledica izgubne kompresije, kjer se signal pred shranjevanjem pretvori v frekvenčno obliko in del podatkov zavriče. Zaradi tega so bolj naključni, kar pomeni večjo entropijo in učinkovitejše kodiranje.



Slika 2: Porazdelitev verjetnosti bajtov v kompresiranem zapisu MP3. Žagasta oblika porazdelitve je posledica izgubne kompresije.

Kompresirani formati imajo pri H_1 višje vrednosti (bližje 8 bitov/simbol), kar pomeni, da so bajti po kodiranju statistično bolj enakomerno porazdeljeni, kar pomeni, da datoteka vsebuje manj ponavljajočih se vzorcev.

3.3 Slikovne datoteke

Pri analizi slike je bil uporabljen enak format (.jpg), vendar se je spremenjala velikost fotografije po ločljivosti. Za njih smo izračunali entropijo pri različni dolžini niza. Uporabljena fotografija je vidna na sliki 3.



Slika 3: Fotografija Mednarodne vesoljske postaje (ISS), uporabljena pri analizi entropije slikovnih podatkov.

| Datoteka | Ločljivost [px] | Velikost datoteke |
|--------------|-----------------|-------------------|
| iss_0480.jpg | 480 × 270 | 32 kB |
| iss_0960.jpg | 960 × 540 | 92 kB |
| iss_1920.jpg | 1920 × 1080 | 220 kB |
| iss_2560.jpg | 2560 × 1440 | 320 kB |
| iss_3840.jpg | 3840 × 2160 | 528 kB |
| iss_7680.jpg | 7680 × 4320 | 12 MB |

Tabela 2: Velikosti in ločljivosti JPEG slik ISS uporabljenih pri analizi entropije. Z večanjem ločljivosti raste tako število piklov kot velikost datoteke, saj vsebuje več informacijskih vzorcev in posledično višjo entropijo.

| Datoteka | H_1 | H_2 | H_3 | H_4 | H_5 |
|--------------|-------|-------|-------|-------|-------|
| iss_0480.jpg | 7.97 | 7.18 | 4.97 | 3.73 | 2.98 |
| iss_0960.jpg | 7.95 | 7.57 | 5.47 | 4.11 | 3.29 |
| iss_1920.jpg | 7.97 | 7.74 | 5.88 | 4.43 | 3.55 |
| iss_2560.jpg | 7.96 | 7.75 | 6.02 | 4.55 | 3.65 |
| iss_3840.jpg | 7.86 | 7.64 | 6.15 | 4.69 | 3.77 |
| iss_7680.jpg | 7.99 | 7.93 | 7.53 | 5.88 | 4.71 |

Tabela 3: Izračunane vrednosti entropij (H_1 – H_5) za JPEG slik ISS pri različnih ločljivostih. Z večanjem velikosti slike naraščajo tudi višje entropije (H_3 – H_5), kar nakazuje večjo kompleksnost in manjšo korelacijo med zaporednimi bajti.

Iz tabele je razvidno, da se pri vseh slikah vrednost H_1 nahaja blizu 8 bitov, kar pomeni, da so posamezni bajti v JPEG zapisu skoraj enakomerno porazdeljeni. Pri višjih členih (H_3 – H_5) pa entropija narašča z ločljivostjo slike: manjše slike (npr. iss_0480.jpg) imajo bolj izrazite ponavljajoče se vzorce in zato nižjo entropijo, medtem ko večje slike (npr. iss_7680.jpg) vsebujejo več detajlov in lokalnih variacij, kar vodi do višje entropije. S tem je potrjeno, da večja prostorska kompleksnost slike povzroči večjo informacijsko vsebino.

4 Povzetek

- Z naraščajočim n entropija pada pri vseh vrstah datotek, kar potrjuje prisotnost statistične odvisnosti.
- Besedilo ima najmanjšo entropijo zaradi jezikovne strukture in visoke redundance.
- Zvočni signali imajo srednjo entropijo, kompresija pa zmanjša redundanco.
- Slikovni podatki so najbližje naključnim (pri $H_1 \approx 8$), saj JPEG učinkovito porazdeli bite.
- Večja ločljivost slike poveča entropijo pri višjih n , ker raste količina unikatnih lokalnih vzorcev.

Literatura

- Wikipedia, *Entropy (information theory)*. [Na spletu]. Dostopno na: [https://en.wikipedia.org/wiki/Entropy_\(information_theory\)](https://en.wikipedia.org/wiki/Entropy_(information_theory)) [Dostopano: 8. november 2025].
- N. Pavešić, *Informacija in kodi*, 2. spremenjena in dopolnjena izdaja. Ljubljana: Založba Fakultete za elektrotehniko in Fakultete za računalništvo in informa-

- tiko, 2010. ISBN 978-961-243-145-7. [Na spletu]. Dostopno na: <https://plus.cobiss.net/cobiss/si/sl/data/cobib/250590208>
- [3] R. Prezelj, *GitHub repozitorij*, 2025. [Na spletu]. Dostopno na: <https://github.com/RokPre/informacija-in-kodi>
- [4] Spiceworks, "What Is a ZIP File? Meaning, Working, and Advantages", *Tech Encyclopedia*, 2024. [Na spletu]. Dostopno na: <https://www.spiceworks.com/tech/tech-general/articles/what-is-zip-file/>
- [5] T. A. Welch, "A Technique for High-Performance Data Compression", *Computer*, letn. 17, št. 6, str. 8–19, jun. 1984. doi: 10.1109/MC.1984.1659158
- [6] D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes", *Proceedings of the IRE*, letn. 40, št. 9, str. 1098–1101, sep. 1952. doi: 10.1109/JR-PROC.1952.273898
- [7] S. Pfeiffer, "The Ogg Encapsulation Format Version 0", *IETF RFC 3533*, maj 2003. [Na spletu]. Dostopno na: <https://www.rfc-editor.org/rfc/rfc3533>
- [8] Wikipedia, *Pulse-code modulation*. [Na spletu]. Dostopno na: https://en.wikipedia.org/wiki/Pulse-code_modulation [Dostopno: 8. november 2025].