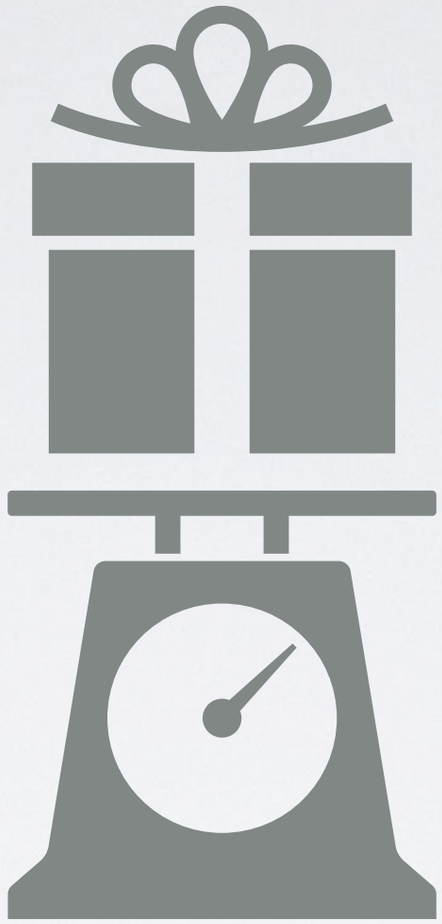


KAKO MERIMO PRESENEČENJE [ENTROPIJA IN RELATIVNA ENTROPIJA]

Žiga Virk

ZWI ZWE IGH OFON SAKOKISA?

[Kako si se imel za vikend?].

TAKA.

[Dobro.]

ZWA ZEA FLI ORI TADA LEME OKI-DOKI TADA LEN, PARA ISTI KALA MORI: TE DAKA MALILA SINTI GLANDI ?

[Je deževalo?]

KOLI.

[Malce je deževalo dopoldne. Ravno dovolj, da mi je opralo vesoljsko plovilo. Popoldne ob kosilu pa je že sijalo sonce.]

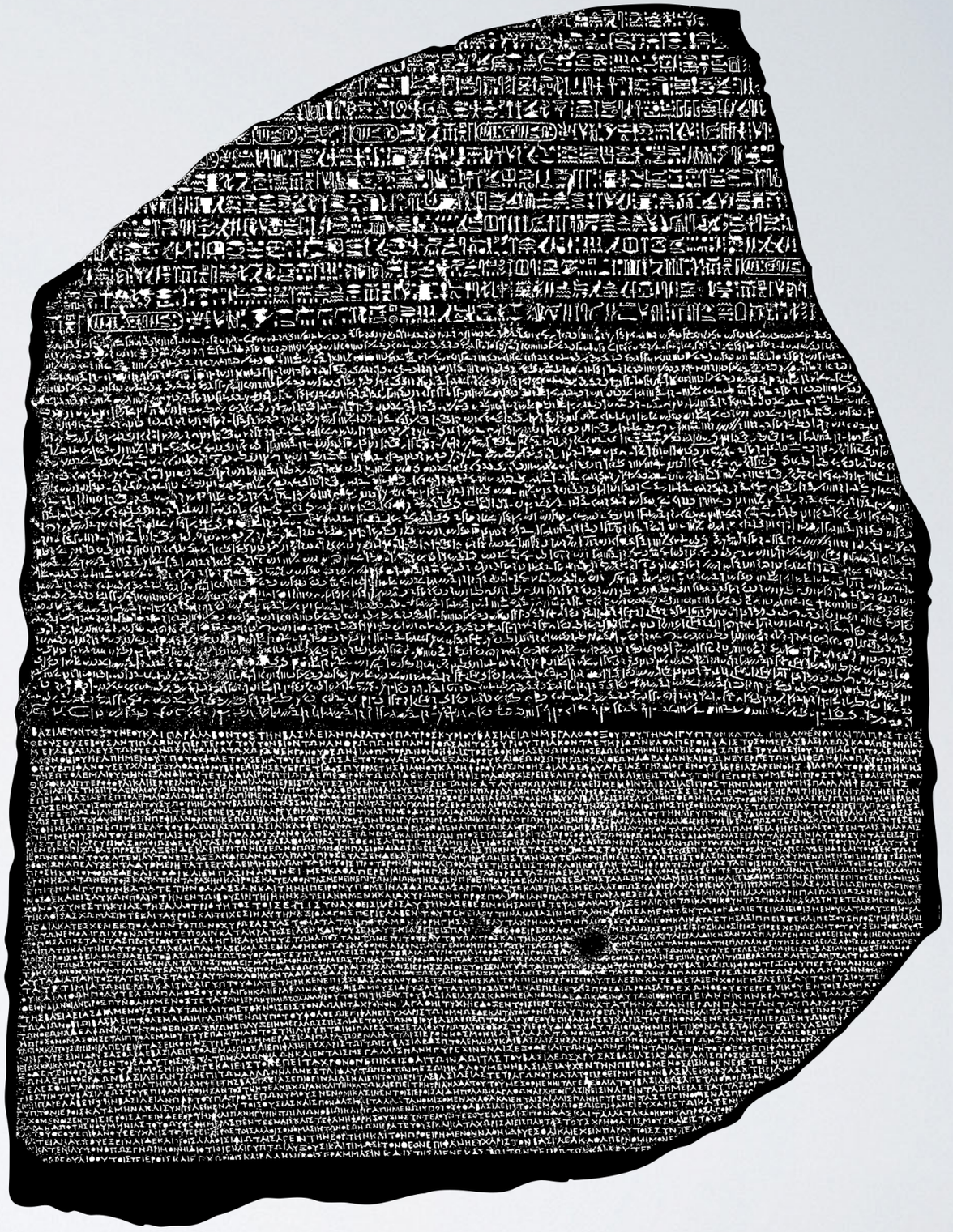
POGOVOR MED VESOLJCEMA

[Slovenski prevod v oklepajih]

Razlog za presenečenje:
Dolžina prevoda ne ustreza pričakovani dolžini glede na količino informacij.

<u>Obravnavali bomo naslednje:</u>

- Količina informacij
- Pričakovana dolžina zapisa (entropija)
- Presenečenje izhajajoče iz dejanske dolžine zapisa (relativna entropija)



File:Rosetta Stone.svg. (2020, October 20). Wikimedia Commons, the free media repository. Retrieved 12:15, February 1, 2022 from https://commons.wikimedia.org/w/index.php?title=File:Rosetta_Stone.svg&oldid=495427405.

KOLIČINA INFORMACIJ(E)

- Podatki ≠ informacije

- Spominska celica v računalniku ima vrednost 0 ali 1. **Mera podatkov**: vsaka celica ima 1 **bit** podatkov [→ byte, KB, MB, GB, …].

- Za opredelitev informacije potrebujemo kontekst (slučajno spremenljivko): izidi $x_1, x_2, \ldots, x_n$ in njihove verjetnosti (frekvence) $p_1, p_2, \ldots, p_n$.

$log_2 6$ bitov

- Informacija nam pomaga določiti izid:
  - 1 bit razlikuje med dvema izidoma.

  - n bitov razlikuje med 2^n izidi.

- **Def:** Izid x_i vsebuje $\log_2(1/p_i) = -\log_2 p_i$ informacije. Enota: bit oz. Shannon (NAT pri osnovi e, hartley in digit pri osnovi 10. Od sedaj naprej $\log = \log_2$).

PRIMERI INFORMACIJE

- Izid meta kovanca: $\log 2 = 1$ [bit]
- Izid meta kocke: $\log 6$
- Vržemo kocko in pogledamo sodost/lihost: 1
- Enakomerno izberemo število med 1 in 2^{10}: 10
- Danes je sobota: 0
- Ob 23:00 v Ljubljani ne bo sijalo sonce: 0
- Zbudimo se v letu 2022 in izvemo, da je datum 3.2.: $\log 365$
- Trenutno je ura med 11:00 in 12:00: 0
- Po dolgem spanju izvemo, da je četrtek popoldne: $\log 14 = 1 + \log 7$

KARAKTERIZACIJA INFORMACIJE

Definicija $-\log_a p$ je edina nenegativna količina φ, ki zadošča naslednjim trem pogojem:

- $p = 1 \implies \varphi = 0$

- φ je padajoča v p

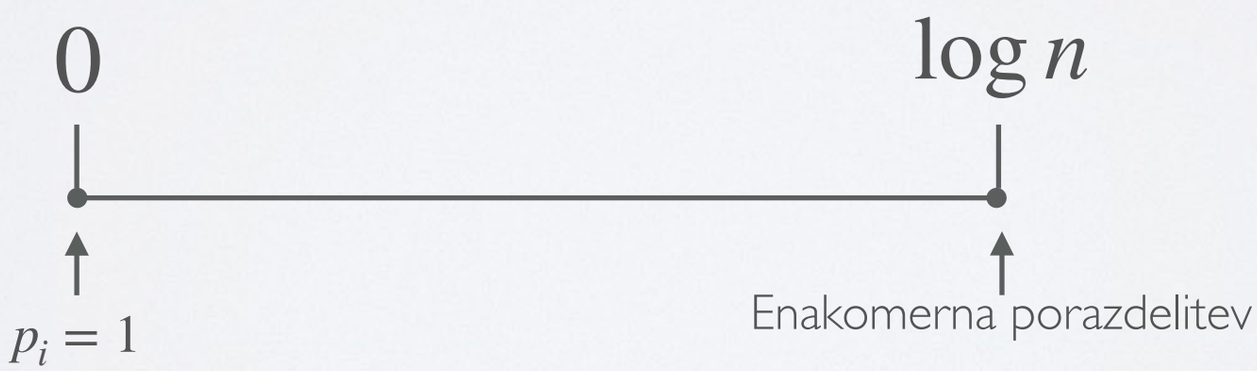
- Če sta A in B neodvisna, je
  $\varphi(A \cap B) = \varphi(A) + \varphi(B)$

ENTROPIJA: POVPREČNA KOLIČINA INFORMACIJE

Kontekst (slučajna spremenljivka X): izidi $x_1, x_2, \ldots, x_n$ in verjetnosti (frekvence) $p_1, p_2, \ldots, p_n$.

Def.: Entropija $H(X) = \sum_{i=1}^{n} p_i \log(1/p_i)$. $\quad [0 \log 0 = 0 \log(1/0) = 0]$

Interval entropije:

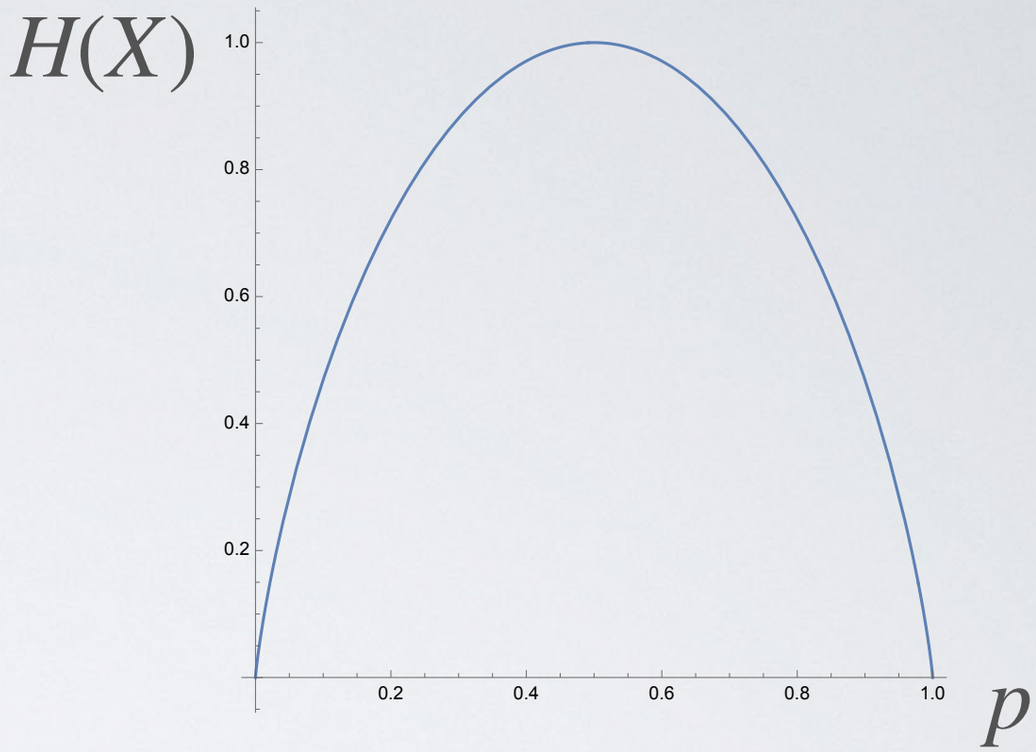


$log\ 2 = 1$

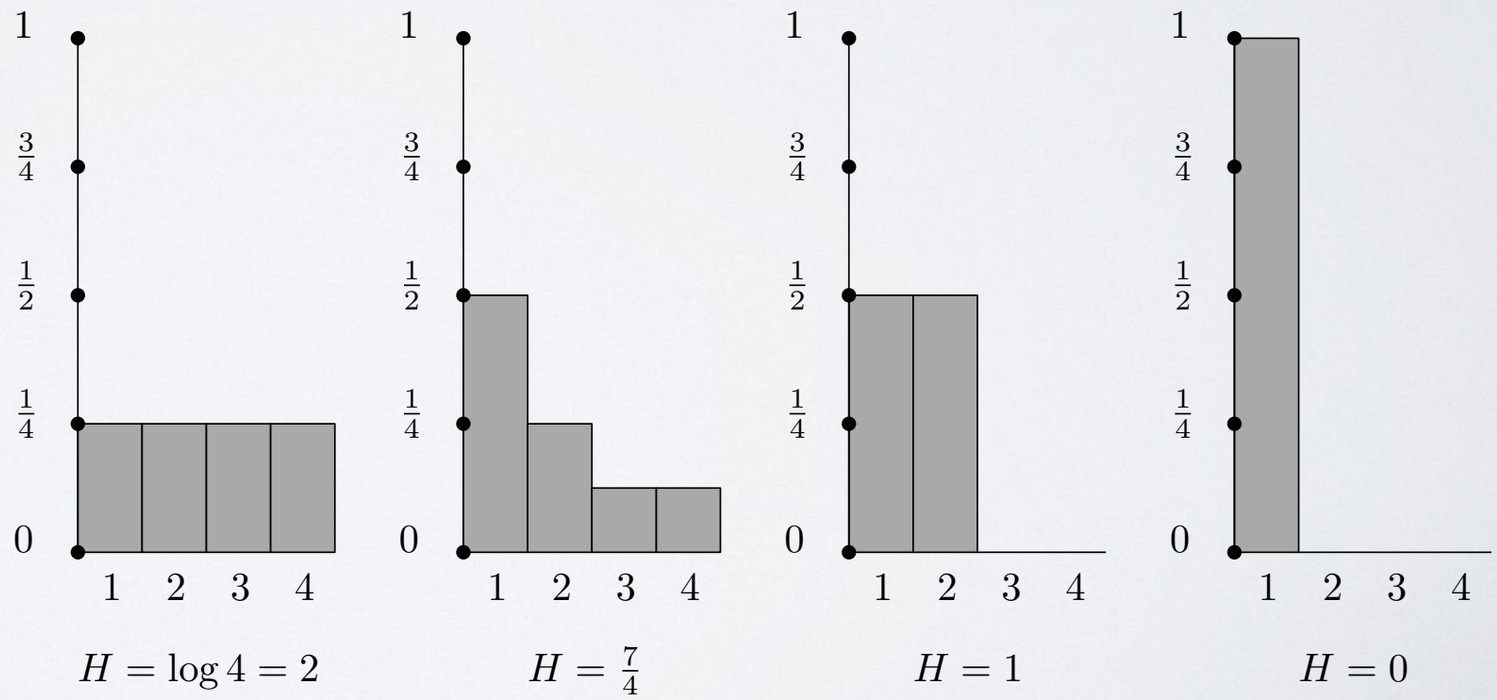
$log\ 6$

Slučajna spremenljivka X:

- Dva izida
- Verjetnosti p in $1 - p$.

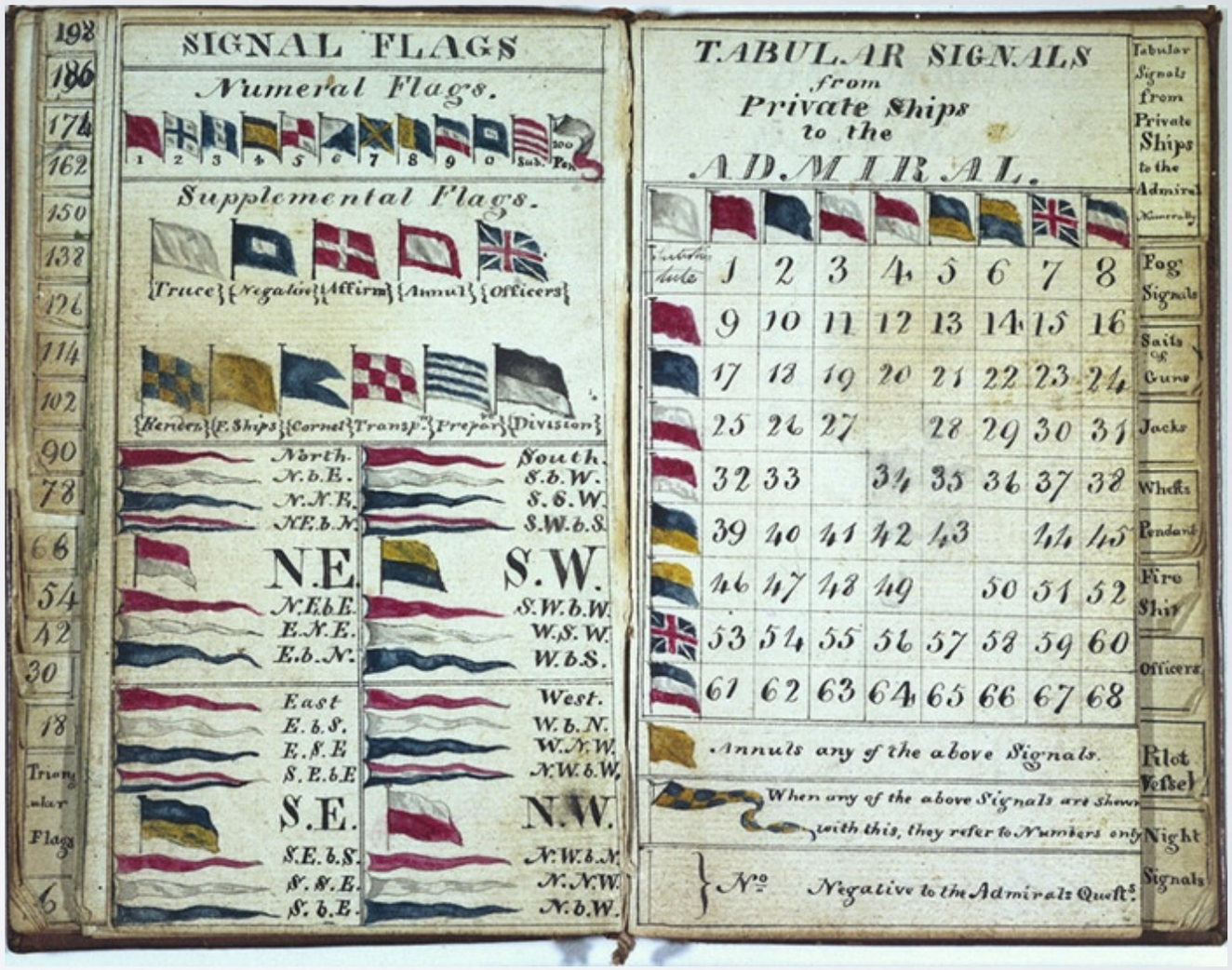


Entropija nekaterih porazdelitev na štirih točkah:



ENTROPIJA IN UČINKOVITI ZAPISI - KODIRANJA

- Mornariške zastave



Marryat, 1817

File:Marryat-Signalcode.jpg. (2015, May 3). Wikimedia Commons, the free media repository. Retrieved 20:54, February 1, 2022 from https://commons.wikimedia.org/w/index.php?title=File:Marryat-Signalcode.jpg&oldid=159589000.

ENTROPIJA IN UČINKOVITI ZAPISI - KODIRANJA

- Mornariške zastave
- Morseova koda

File:Dvorak keyboard layout diagram.png. (2020, September 17). Wikimedia Commons, the free media repository. Retrieved 21:02, February 1, 2022 from https://commons.wikimedia.org/w/index.php?title=File:Dvorak_keyboard_layout_diagram.png&oldid=462328034.

- Tipkovnice (Dvorak)
- Okrajšave (LOL), bližnjice (Ctrl-C), kratice (FMF)
- Matematični zapis: pet in tri je enako osem.
- *"Ne izrecite malo z veliko besedami, raje povejte veliko v le nekaj besedah."*

$A = \{x_1, x_2, \ldots, x_n\}$ abeceda s frekvencami (verjetnostmi) črk $p_1, p_2, \ldots, p_n$.

Vprašanje: Kako učinkovito zakodirati črke z zaporedjem ničel in enic?

Primer:

$H = 3/2$

črka	frekv.	kodiranje 1	kodiranje 2
a	0.5	0 0	0
b	0.25	0 1	1 0
c	0.25	1 0	1 1
d	0	1 1	-
Povprečna dolžina zapisa črke		2	3/2

Izrek: Naj bo ℓ povprečna (glede na p_i) dolžina kode črke (zaporedja 0,1) za neko kodiranje. Tedaj:

- $\ell \geq H$
- Obstaja kodiranje z dolžinami kod $\lceil \log(1/p_1) \rceil, \lceil \log(1/p_2) \rceil, \ldots, \lceil \log(1/p_n) \rceil$. Pri tem je $\ell \leq H + 1$.
- Huffmanovo kodiranje: optimalna koda v zgornjem smislu (uporabno v .MP3, .JPEG, .ZIP, …)
- S kodiranjem nizov črk lahko za poljuben $\varepsilon > 0$ dosežemo $\ell \leq H + \varepsilon$.

Primer:

$H = 7/8$

črka	frekv.	kodiranje 1	kodiranje 2
a	1/2	0 0	0
b	1/4	0 1	1 0
c	1/8	1 0	1 1 0
d	1/8	1 1	1 1 1
Povprečna dolžina zapisa črke		2	7/8

- Shannon [1951]: Entropija frekvenc 26 črk v angleščini je približno 4.1.
- Teoretično: kot bi uporabljali 17 črk z enako frekvenco: $2^{4.1} \approx 17$.
- Z učinkovitim kodiranjem nizov bi lahko porabili le dobre 4 bite na črko (standardno se uporablja 8 bitov).

PRESENEČENJE: RELATIVNA ENTROPIJA

Skozi praktične primere si bomo ogledali kako merimo količino informacij, kaj je entropija in kako v vsakdanjem življenju uporabljamo relativno entropijo. V posebnem primeru, ~~katerega bomo seveda tudi obravnavali,~~ bo relativna entropija, ~~kot je omenjena v prejšnjem stavku,~~ predstavljala delež dolžine stavka, ~~potencialno podanega v povzetku kakšnega matematičnega predavanja~~, za katerega bi lahko stavek, ~~ki smo ga ravnokar omenili,~~ skrajšali, pri čemer bi, ~~kot je seveda pričakovati~~, hoteli s krajšim stavkom, ~~prav tako omenjenim zgoraj~~, podati enako količino informacij, kot s prvotno omenjenim ~~čudovitim a potencialno dolžinsko preambicioznim~~ stavkom.

PRESENEČENJE: RELATIVNA ENTROPIJA

- $\{x_1, x_2, \ldots, x_n\}$ abeceda s frekvencami $p_1, p_2, \ldots, p_n$ v besedilu P, s frekvencami $q_1, q_2, \ldots, q_n$ v besedilu Q.

- φ_P idealno kodiranje za P, φ_Q idealno kodiranje za Q.

- Povprečno število bitov porabljenih za črko v P, če uporabimo φ_P:

$$H(P) = \sum_{i=1}^{n} p_i \log(1/p_i)$$

- Povprečno število bitov porabljenih za črko v P, če uporabimo φ_Q:

$$H(P, Q) = \sum_{i=1}^{n} p_i \log(1/q_i)$$

- **Relativna entropija** je razlika zadnjih dveh izrazov:

$$D(P \parallel Q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{q_i}\right)$$

Nekaj lastnosti:

$$D(P \parallel Q) \in [0, \infty]$$

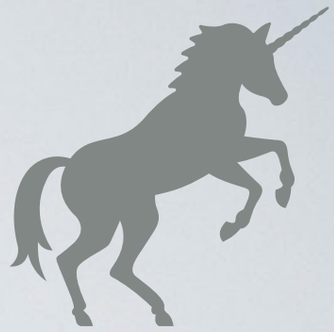
$$D(P \parallel Q) = 0 \Leftrightarrow P = Q$$

$$D(P \parallel Q) = \infty \Leftrightarrow \exists i : 0 = q_i < p_i$$

Ni simetrična

Predstavlja dodatno število bitov potrebnih zaradi uporabe "napačnega" kodiranja. (Oz. Dodatna količina administrativnega dela zaradi neoptimalnih postopkov.)

MALO ČAROVNIJE



- Izberite 5 števil $a_1, a_2, a_3, a_4, a_5 \in \mathbb{N} \cap [4,11]$.

- Pomnožite jih z dnevom (24), letom (2026) in številom mesecev (12).

- Izberite si skrivno neničelno števko rezultata in seštejte ostale števke.

$$24 * 2026 * 12 * a_1 * a_2 * a_3 * a_4 * a_5 = b_1 b_2 b_3 b_4 b_5 b_6 b_7$$

$$\Sigma = b_1 + b_2 + b_3 + b_5 + b_6 + b_7$$

NE-UČINKOVITI ZAPISI

- Včasih želimo na napake odporen zapis, t.j., da se kode razlikujejo vsaj za 3 napake.

- Razlog: Če se pri tipkanju "leva" zmotimo za 1 znak…

- Varen zapis: "lllleeeevvvaaaa". Ob eni napaki lahko rekonstruiramo besedo.

- Zadnja številka EMŠO je kontrolna.

Hvala za pozornost.