Az igazi Shannon

Halácsy Péter

2006. október

Szimbólumokkal kommunikálunk

- Az ember mindig szimbólumokkal kommunikál.
- Kezdetben: kiáltás, bólintás, kézmozdulat, füstjel, barlangrajz (dokumentum!)
- Jön az írás: ábécé = szimbólumkészlet.
- Manapság szinte fénysebességgel továbbítunk már szimbólumokat — szöveget, hangot képet, videót digitális kódolásban.

Kommunikáció modellje

kommunikációs rendszer elemei

- a jel forrástól (adó) származik
- az információs csatornán adott sebességgel tud haladni (sávszélesség, de vajon miért?)
- nem ugyanaz jön ki a csatornán, zaj van (recseg a telefon)
- a vevőnek dekódolni kell a jeleket

Shannon célja

- a kommunikáció matematikai formalizmusa,
- hogy mérnöki problémaként legyen kezelhető az adatátvitel
- nem volt célja: nyelvészeti, bölcsésztudományi, kommunikáció tudományi és PR szakmáknak, khm, metaforát adni

Shannon célja

- a kommunikáció matematikai formalizmusa,
- hogy mérnöki problémaként legyen kezelhető az adatátvitel
- nem volt célja: nyelvészeti, bölcsésztudományi, kommunikáció tudományi és PR szakmáknak, khm, metaforát adni
- fogadás: nyelvtantanárok többsége nem érti az információ elmélet alapjait

Az információ fogalma

Speciális jelentés.

- Telegráf (sms) üzenetekben bizonyos szavakat kihagynak, rövidítenek.
- Mégis kitalálható, h mi volt az eredeti üzenet. Megjósolhatóak a szimbólumok.
- Az üznet knyen dekdolhtó.
- Megjósolható szimbólumok, elhagyhatóak, redundánsak.

Információ bizonytalanságot csökkent

- Tfh. egy gép háromféle szimbólum A, B, C és D valamelyikét tudja kiadni
- Várjuk a következő szimbólumot, bizonytalanság van bennünk.
- Megjelenik egy szimbólum, a bizonytalanság csökkent, mert információt kaptunk.
- Hogyan mérhető az információ?
- Nevezzük négy szimbólumos bizonytalanságnak

Két forrás?

- Most vegyünk egy másik gépet, ez 1 és 2 szimbólumokat ad ki.
- két szimbólumos bizonytalansága van
- egymás után nézzük meg a kimenetet: 6 szimbólumos bizonytalanság csökkent

Két forrás?

- Most vegyünk egy másik gépet, ez 1 és 2 szimbólumokat ad ki.
- két szimbólumos bizonytalansága van
- egymás után nézzük meg a kimenetet: 6 szimbólumos bizonytalanság csökkent
- ha a kettőt összekombináljuk 8 lehetséges szimbólum lesz: $A1, A2, B1, B2, C1, C2, D1, D2$
- ennek már nyolc szimbólumos bizonytalansága van

Szeretnénk inkább összeadni az információ mértékét.

Logaritmus

- Megoldható: első gép $log(4)$ második $log(2)$ mértékű bizonytalanságot csökkent, akkor
- egymás után $log(4) + log(2) = log(8)$
- a kombinált gép szintén $log(8)$

Emlékeztető

$$log(ab) = log(a) + log(b)$$

Kettes alapú logaritmus

- Mindegy, hogy mi a logaritmus alapja.
- Ha $log_2()$-t használunk, bit-ben kapjuk a mértéket.
- Két szimbólum esetén $log(2) = 1$ bit a bizonytalanság csökkenés.
- $log_2(1) = 0$ bit. Ez mit jelent?

A barkochba játékunk

- 32 ember közül gondoltam egyre. Hány kérdésből található ki, hogy kire?
- És ha előre meg kell mondani a kérdéseket?

A barkochba játékunk

- 32 ember közül gondoltam egyre. Hány kérdésből található ki, hogy kire?
- És ha előre meg kell mondani a kérdéseket?
- Minden embert számozzunk meg és a sorszámát írjuk fel binárisan.
- $00000, 00001, 00010, 00011, 00100, 00101, 00110, 00111 \ldots$
- 5 kérdés az 5 számjegyre vonatkozik.

A barkochba játékunk

- 32 ember közül gondoltam egyre. Hány kérdésből található ki, hogy kire?
- És ha előre meg kell mondani a kérdéseket?
- Minden embert számozzunk meg és a sorszámát írjuk fel binárisan.
- $00000, 00001, 00010, 00011, 00100, 00101, 00110, 00111\ldots$
- 5 kérdés az 5 számjegyre vonatkozik.
- ZH kérdés: 23 binárisan?

Információ mértéke

- N elemű (szimbólum) halmazból egy kiválasztása
- $log_2(N)$ bit információt jelent
- Ügyesen kódolva a választás $log_2(N)$ biten leírható: sorba rendezzük és binárisan megszámozzuk

Szimbólumok nem egyforma valószínűséggel

- Tfh. tudjuk, hogy az első gép D-t soha nem adja ki. Mennyi bizonytalanságunk csökken? $log_2(3) < log_2(4)$

Emlékeztető

$$log(a) < log(b) \leftrightarrow a < b$$

Szimbólumok nem egyforma valószínűséggel

- Tfh. tudjuk, hogy az első gép D-t soha nem adja ki. Mennyi bizonytalanságunk csökken? $log_2(3) < log_2(4)$
- Néha kiadja D-t, de csak ritkán. Mit várunk?

$$log_2(N) = log_2(\frac{1}{\frac{1}{N}}) = log_2(\frac{1}{P})$$

Emlékeztető

$$log(a) < log(b) \leftrightarrow a < b$$

Információ mértéke

Egy X szimbólum vételekor kapott információ:

$$I(X) = log_2(\frac{1}{P}) = -log_2(P),$$

ahol P X valószínűsége.

- $I(fej) = 1$ bit, mert $P(fej) = \frac{1}{2}$
- Ha tudjuk, hogy csak fej lehet, akkor $I(fej) = 0$ bit.
- Valami minél ritkább, annál nagyobb az információ tartalma.

Emlékeztető

$P(X) = \frac{k}{N}$, azaz N esetből k-szor az X szimbólumot kapjuk (körülbelül).

Nem minden igen/nem kérdés 1 bites

- Kisasszony, szereti a sajtot?
- Kisasszony, hozzám jön feleségül?
- Kisasszony, nyert múlt héten a lottón?

Formálisabban, ha N elemű halmaz minden elemére $P = \frac{1}{N} \leftrightarrow log_2(N)$ bit egy kiválasztása.

Buta torpedó

Egy 8x8 táblán, elrejtek egy torpedót, bárhova tehettem, nincsen preferált cellám. Ki kell találni, hol van.

	1	2	3	4	5	6	7	8
A								
B								
C								
D								
E								
F								
G								
H								

Buta torpedó

- Mennyi információt jelent, ha megmondom hol van.
- Hány kérdésből lehet biztos kitalálni?
- Hány igen/nem kérdésből lehet biztos kitalálni?
- Mennyi az esélye, hogy elsőre kitalálod?

Mennyi információ?

Alakítsuk át a feladatot Shannon modelljének megfelelően.

- Valaki kérdéseket tesz fel, amire én válaszolok.
- Úgy, hogy előre rögzített szimbólum halmazból választhatok.
- Csatorna nem fontos, zaj nincs.

Formális és gyors megoldás

- 64 lehetséges helyre tehettem a torpedót
- Számozzuk meg a cellákat $0, 1, 2, \ldots 63$
- Kódoljuk le binárisan. $log_2(64) = 6$ bit szükséges
- Tehát, ha én minden cellát egyenlő valószínűséggel választom, akkor is 6 kérdésből ki lehet találni.
- (Megj: ha tudod, hogy inkább középre szoktam tenni a torpedót, akkor átlagban könnyebb kitalálnod.)

Formális és gyors megoldás

- 64 lehetséges helyre tehettem a torpedót
- Számozzuk meg a cellákat $0, 1, 2, \ldots 63$
- Kódoljuk le binárisan. $log_2(64) = 6$ bit szükséges
- Tehát, ha én minden cellát egyenlő valószínűséggel választom, akkor is 6 kérdésből ki lehet találni.
- (Megj: ha tudod, hogy inkább középre szoktam tenni a torpedót, akkor átlagban könnyebb kitalálnod.)

Megmutatom, hogy az információ mértéke ennél univerzálisabb.

Melyik a jobb stratégia?

1. Hol van?
   64 szimbólum valamelyike $A1$, $A2$, $A3 \dots H8$
2. Melyik oszlopban van? Melyik sorban van?
   Mindkét esetben 8 szimbólumom van: 1, 2 . . . 8, A, B . . . H, de két kérdés csak.
3. $G3$-n van? $B1$-n van? . . . $E5$-n van?
   Két szimbólum (igen, nem), de maximum 64 kérdés kell.
4. 1.,2.,3.,4. oszlopban van? A, B, C, D sorban van? Két szimbólum (igen, nem), hány kérdés?

Menny információ?

1. Hol van? $log_2(\frac{1}{\frac{1}{64}}) = 6bit$
2. Melyik oszlopban van? $log_2(\frac{1}{\frac{1}{8}}) = 3bit$ + Melyik sorban van? $log_2(\frac{1}{\frac{1}{8}}) = 3bit \longrightarrow 6bit$

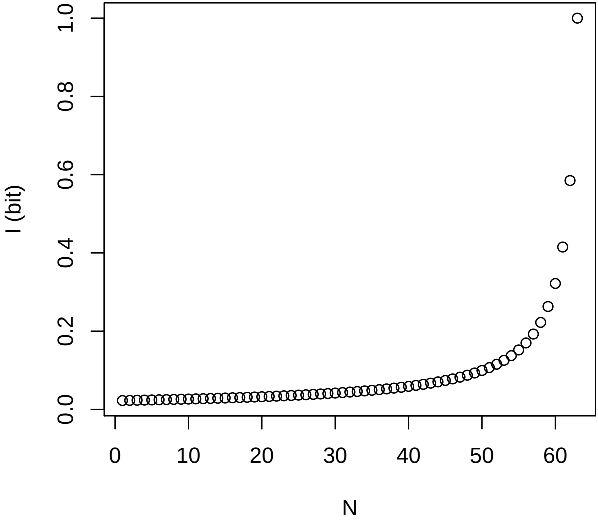
Menny információ?

$1, 1$-n van? $1, 2$-n van? $\ldots 8, 8$-n van?

- Mennyi a valószínűsége, hogy elsőre kitalálod?
  $P(1, igen) = \frac{1}{64}$
- Kicsi eséllyel kitalálható, de mennyire lepődünk ezen meg?
- Inkább: $P(1, nem) = \frac{63}{64} \to 0.0227$ bit
- Aztán: $P(2, nem) = \frac{62}{63} \to 0.0230$ bit

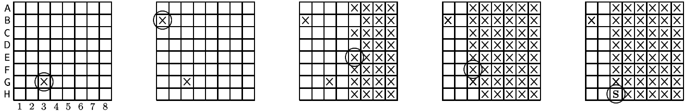
Beugrószintű ZH kérdés.

Már 62-t találgattál. Egyszer sem talált. Rákérdezel a 63. cellára. Mi az információ tartalma, hogy ,,nem, nem ott van" és az ,,ott van" válaszoknak?



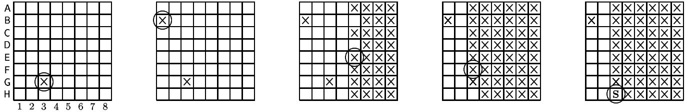
Az n. kérdésre kapott "nem" válasz információ tartalma.

49. kérdésre kitalálod. Mennyi információt kaptál?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

49. kérdésre kitalálod. Mennyi információt kaptál?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x=\mathtt{n}$	$x=\mathtt{n}$	$x=\mathtt{n}$	$x=\mathtt{n}$	$x=\mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

Őrület. Ez is 6 bit.

Akárhányadikra találod ki, mindig 6 bit. Már csak k maradt.

$$
\begin{aligned}
log_2 \frac{64}{63} + log_2 \frac{63}{62} + ... + log_2 \frac{k+1}{k} + log_2 \frac{k}{1} \\
= log_2 (\frac{64}{63} x \frac{63}{62} x ... x \frac{k+1}{k} x \frac{1}{k}) \\
= log_2(64)
\end{aligned}
$$

Felezgetős módszer

- Mindig olyan kérdést teszel fel, ami a lehetséges cellák számát pont felezi.
- Ez igen/nem kérdéssel feltehető.
- Minden esetben ugyanolyan valószínűséggel várod az igen-t és a nem-et.
- Minden válasz 1 bit információt ad.
- Hány kérdés kell?

Felezgetős módszer

- Mindig olyan kérdést teszel fel, ami a lehetséges cellák számát pont felezi.
- Ez igen/nem kérdéssel feltehető.
- Minden esetben ugyanolyan valószínűséggel várod az igen-t és a nem-et.
- Minden válasz 1 bit információt ad.
- Hány kérdés kell? 6 bit

Hol az információ?

- Mindegy, hogy milyen stratégiával kérdezel, összesen mindig 6 bit információt adok, mire kiderül, hol a torpedó.
- Minden cellára rákérdezés esetén a 32. nem után 1 bitet kaptál összesen. Ugyanannyit, mint a felezgetős módszer első kérdése után.
- A módszerek különböznek a szükséges kérdésszámban, de az információ nyereség mindig ugyanannyi (az a játék része).
- Ugyanannyi információt lehet a torpedóval lekódolni, mint a *gondoltam egy számra 1–64 között*

Hol a kommunikáció ?

- Tüntetők 64 lehetséges helyszín valamelyikét a tüntetés előtt választják ki véletlenszerűen, hogy a rendőrség ne tudjon előre felkészülni. Nem beszélhetnek egymással, nem telefonálhatnak.
- A vezető kiválasztja a helyszínt és szeretne 6 bit információt átküldeni mindenkinek.
- Előre megállapodtak egy kódrendszerben: egy szabadtéri saktáblán otthagynak egy befejezett sakjátszmát.
- Mindenki tudja, hogy melyik cella melyik helyszínt jelenti.
- A fekete király helye mutatja a tüntetés helyét.

Mi a kódolás?

Mindenkinek ismernie kell a 64 helyszínt és egy kódtáblázatot.

Oktogon	A1
Kossuth tér	A2
...	
Blaha	H8

És, hogy a sakktábla melyik cellája melyik sornak felel meg.

6 bitet átvinni zajos csatornán

Néha elkezdenek játszani a gyerekek a téren és arrébb teszik a fekete királyt. Jobb módszer kéne.

6 bitet átvinni zajos csatornán

Holnap 12 után minden órában megcsörgetlek egyszer, vagy nem. Ez 6 db. igen/nem (két szimbólumos) döntés.

- $P = 0.05$ a valószínűsége, hogy nincs térerő vagy véletlenül felhívlak.
- A zaj miatt minden bit csak $P = 0.95$ valószínűséggel megy át. 6 bit 73.5%, hogy hibátlanul átmegy.
- Ráadásul nem is tudsz arról, hogy hiba történt.
- Mit tehetünk?

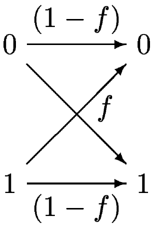
Átvitel zajos csatornán

Shannon ezzel foglalkozott és nem nyelvészettel.

- Hibadetektálás: A hetedik alkalommal felhívlak, ha előtte páros esetben akartam telefonálni. Feltehetőleg észreveszed, ha baj volt.
  Figyelem: Ha nincs zaj és ezt Te is tudod, akkor a 7 hívás információ tartalma 0 bit lenne. (Ha tudod, hogy mit jelentenek a hívások!)
- Hibajavítás: minden órában háromszor telefonálok. Ezzel egy hibát tudunk javítani.

Képátvitel zajos csatornán

100x100 bitben kódolok egy fekete–fehér képet. A biteket átküldöm egy csatornán, de az néhány bitet invertál. $f = 0.1$



Milyen zaj lehet?

Igazából nincs zajmentes csatorna.

- Átrakják a fekete királyt.
- A disk néha bitek ront el.
- Telefonvonalat zavarja, ha bekapcsol valaki egy porszívót.
- Zajként értelmezhető a DNS átadás közbeni ,,hiba" is.

A kommunikáció maximális sebességét három dolog határozza meg:

- A fénysebesség.
- A csatorna zaja.
- Az elérni kívánt hibaszázalék.

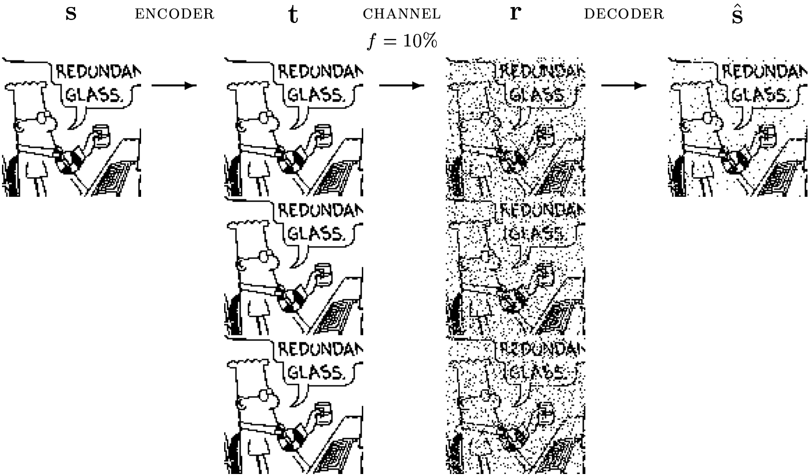
Hibatűrő kódolás

Minden bitet háromszor küldök el. Ha egy hármas blokkban csak egy bit vész el, akkor korrigálni tudom a hibát.

$\mathbf{s}$	0	0	1	0	1	1	0
$\mathbf{t}$	000	000	111	000	111	111	000
$\mathbf{n}$	000	001	000	000	101	000	000
$\mathbf{r}$	000	001	111	000	010	111	000
$\hat{\mathbf{s}}$	0	0	1	0	0	1	0
corrected errors		$\star$					
undetected errors					$\star$		

Nem minden hibát tudtam javítani

Az eredmény sokkal jobb, de nem 100%-os (elvileg nem is lehet).



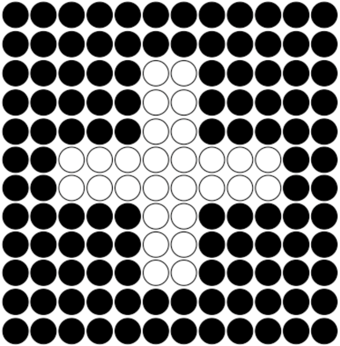
Redundancia

- A hibatűrő kódolásnak ára van. 30 000 bitet kellett használnom. 3x annyi idő átküldeni.
- 20 000 bit redundáns, ha nincs zaj.
- Fordítva gondolkozva: redundancia kiszűrése a tömörítés.

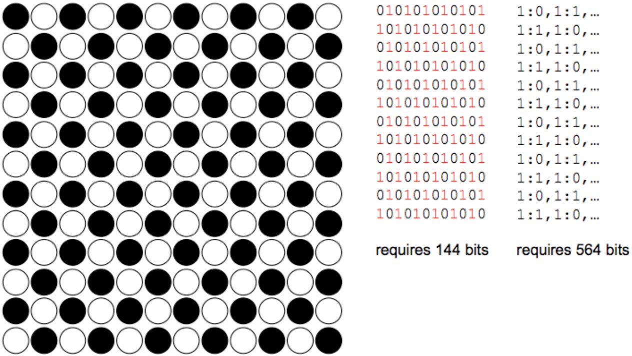
Tömörítsük a képet

Sok fehér folt van egymás mellett.

1. `ww`, `bw`, `wb`, `bb` szimbólumokat használok, és észreveszem, hogy sokkal több a `ww`, mint a `bb`
   ww - 0 wb - 10 bw - 110 bb - 111
2. Egymás mellett 12 fehér pont legyen 12:0 alakban kódolva. (köv ábra)



000000000000	12:0
000000000000	12:0
000001100000	5:0,2:1,5:0
000001100000	5:0,2:1,5:0
000001100000	5:0,2:1,5:0
001111111100	2:0,8:1,2:0
001111111100	2:0,8:1,2:0
000001100000	5:0,2:1,5:0
000001100000	5:0,2:1,5:0
000001100000	5:0,2:1,5:0
000000000000	12:0
000000000000	12:0
requires 144 bits	requires 104 bits
	30% compression



Van, amikor nem jó ez a kódolás.

Szövegtömörítés - Nincs zaj!

- Emlékszünk, a gyakori szavakat rövidítjük.
- Ezzel az össz karakter szám csökken.
- Vagy másképp: az átlagos karakter/szó igény.

Példa

- Kocsikat kell számlálni nyári munkában. Pontosabban egy elhaladó kocsiknak le kell írnod egymás után a színét.
- Észrevesszük, hogy nem egyenletes a színek eloszlása

fehér	50%
kék	25%
piros	25%

- Találjunk olyan kódolást, ami a lehető legrövidebb munkát adja!

Egyenletes kódolás

Minden szín kap egy bináris számot.

fehér	50%	00
kék	25%	01
piros	25%	10

Mennyi a várható bit szükséglet? $N = 100$ kocsi haladt el.

Egyenletes kódolás

Minden szín kap egy bináris számot.

fehér	50%	00
kék	25%	01
piros	25%	10

Mennyi a várható bit szükséglet? $N = 100$ kocsi haladt el.

$$L = 0.5 * 100 * 2 + 0.25 * 100 * 2 + 0.25 * 100 * 2 = 200 bit$$

Emlékeztető

p valószínűségű dolog N esetből körülbelül $p * N$ esetben fordul elő. U.n. nagyszámok törvénye.

Kicsit általánosabban

Van három szimbólum, p_1, p_2, p_3 valószínűséggel fordulnak elő és l_1, l_2, l_3 a hosszúságuk, akkor egy N szimbólumos üzenet várható hossza körülbelül

$$L = p_1 * N * l_1 + p_2 * N * l_2 + p_3 * N * l_3$$

Az átlagos kódolási bit szükséglet

$$\frac{L}{N} = p_1 * l_1 + p_2 * l_2 + p_3 * l_3$$

Mennyi információt kell lekódolnom.

Mennyi információt jelent, ha egy-egy színt látok?

fehér	50%	$log_2(2)$	1 bit
kék	25%	$log_2(4)$	2 bit
piros	25%	$log_2(4)$	2 bit

Mennyi információt kell lekódolnom.

Mennyi információt jelent, ha egy-egy színt látok?

fehér	50%	$log_2(2)$	1 bit
kék	25%	$log_2(4)$	2 bit
piros	25%	$log_2(4)$	2 bit

Mennyi információt fogok kapni $N = 100$ kocsi után?

$$h = 0.5*100*1+0.25*100*2+0.25*100*2 = 50+50+50 = 150 bit$$

200 biten kódoltam le 150 bit információt! Hol a bibi?

Ügyes kódolás.

Értelemszerűen a gyakoribb fehér színnek rövidebb kódot kéne választani, és fölösleges volt a 11 szimbólum. (emlékszünk? a gép, ami D-t sohasem adott ki, de mi ezt nem tudtuk)

fehér	50%	$log_2(2)$	1 bit	0
kék	25%	$log_2(4)$	2 bit	10
piros	25%	$log_2(4)$	2 bit	11

A kódolás jónak tűnik. Az üzenet hossza:

Ügyes kódolás.

Értelemszerűen a gyakoribb fehér színnek rövidebb kódot kéne választani, és fölösleges volt a 11 szimbólum. (emlékszünk? a gép, ami D-t sohasem adott ki, de mi ezt nem tudtuk)

fehér	50%	$log_2(2)$	1 bit	0
kék	25%	$log_2(4)$	2 bit	10
piros	25%	$log_2(4)$	2 bit	11

A kódolás jónak tűnik. Az üzenet hossza:

$$h = 0.5*100*1+0.25*100*2+0.25*100*2 = 50+50+50 = 150 bit$$

150 bit információt 150 biten kódoltam (két jelentése van a bit szónak?)

Mit is csináltunk általánosabban?

Ha van k szimbólumom, amik valószínűsége $p_1, p_2, ..., p_k$

- A k. szimbólum megjelenésének információ tartalma

$$h_k = log \frac{1}{p_k}$$

- N hosszú üzenet várható információ tartalma

$$h(M) = p_1 * N * h_1 + p_2 * N * h_2 ... p_k * N * h_k$$

- Egy-egy szimbólum észlelésének átlagos információ tartalma

$$H = p_1 * h_1 + p_2 * h_2 ... p_k * h_k$$

Entrópia

Ha van k szimbólumom, amik valószínűsége $p_1, p_2, ..., p_k$

$$H = \sum_{i}^{k} p_k log \frac{1}{p_k}$$

Az átlagos várható információ minden egyes szimbólum fogadásánál.

Entrópia

Igazából ez a fontos fogalom.

- Megadja mennyire nem egyenletesen jönnek egymás után a szimbólumok.
- Fej vagy írás entrópiája 1 bit.
- Buta torpedóban a naív kérdezés esetén elsőre mit várhatunk?

$$\frac{1}{64} * log_2 64 + \frac{63}{64} + log_2 \frac{64}{63} = 0.1160953125$$

Rosszul kérdeztünk, mert a várható információ kicsi.

Letter	Probability of Use in English Text	Code 1	Code 2	Code 3
(Space)	0.1859	00000	00	000
A	0.0642	00001	10100	0100
B	0.0127	00010	0111100	011111
C	0.0218	00011	01101100	11111
D	0.0317	00100	011100	01011
E	0.1031	00101	100	101
F	0.0208	00110	1101100	001100
G	0.0152	00111	0101100	011101
H	0.0467	01000	111100	1110
I	0.0575	01001	1100	1000
J	0.0008	01010	101010100	0111001110
K	0.0049	01011	0110100	01110010
L	0.0321	01110	01100	1001
M	0.0198	01101	010100	001101
N	0.0574	01110	01100	1001
O	0.0632	01111	01010100	0110
P	0.0152	10000	10101100	011110
Q	0.0008	10001	010110100	0111001101
R	0.0484	10010	101100	1101
S	0.0514	10011	11100	1100
T	0.0796	10100	0100	0010
U	0.0228	10101	1100100	11110
V	0.0083	10110	1110100	0111000
W	0.0175	10111	1010100	001110
X	0.0013	11000	01110100	0111001100
Y	0.0164	11001	011010100	00111
Z	0.0005	11010	01011100	0111001111