Statistika za jezikoslovno istraživanje

Damir Ćavar
Sveučilište u Zadru
12. svibnja 2010.

Kodiranje u binarni oblik

- Ako imamo 10 znakova, koliko bitova nam treba za kodiranje svih znakova?
- $2^x = 10$
  - dva stanja po svakom bitu = 0 ili 1, znači osnova 2
  - tražimo x tako da 2 na x daje 10:
    - $\log_2(10)$ bitova = 3.321928 = 4 bita

Kodiranje u binarni oblik

- To vrijedi samo za 1/10 vjerojatnost pojave svakog pojedinačnog znaka.
    - Svaki znak ima istu vjerojatnost

Kodiranje u binarni oblik

- Simetrija:

$$log_2(10) = -log_2(\frac{1}{10})$$

- Negativni $\log_2$ od vjerojatnosti jednog znaka daje broj bitova za kodiranje svih znakova, ako su vjerojatnosti za sve znakove iste

Kodiranje u binarni oblik

- Ako su vjerojatnosti za sve znakove iste (njih 10), možemo izračunati broj bitova za kodiranje jednog znaka na način da množimo sve vjerojatnosti s vjerojatnosti jednog znaka:

$$-\frac{1}{10}log_2(\frac{1}{10})$$

Kodiranje u binarni oblik

- Ako zbrajamo broj bitova za sve znakove na osnovi njihove vjerojatnosti i množimo s -1, u R-u (označite i prebacite u R za testiranje):

  ```
  -1 * ( (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) + (log2(1/10) / 10) )
  ```

  dobijemo $\log_2(10)$ = broj bitova za kodiranje tih znakova u binarnom sustavu

Kodiranje u binarni oblik

- Što je isto kao:

$$-\sum_{n=1}^{10}\frac{1}{10}log_2(\frac{1}{10}) = log_2(10)$$

Kodiranje u binarni oblik

- Uzimamo $\log_2$ od broja koji je između 0 i 1 (relativna frekvencija ili vjerojatnost), što znači da će rezultat biti negativan. Ako zbrajamo negativne brojeve, dobijemo negativni broj. Kako bi dobili pozitivni broj, znači broj bitova za kodiranje znakova, množimo s -1.

Kodiranje u binarni oblik

- Što, ako se vjerojatnosti pojedinačnih znakova razlikuju?
- Npr. imamo podatke da se 10 znakova pojavljuje s nekom posebnom vjerojatnosti:

  a: 0.21, b: 0.15, c: 0.1, d: 0.05, e: 0.2, f: 0.06, g: 0.04, h: 0.09, i: 0.05, j: 0.05

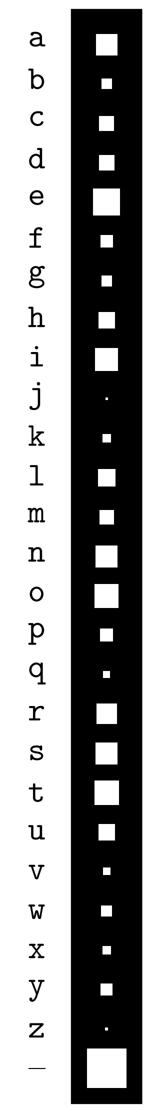
Kodiranje u binarni oblik

- Različite frekvencije znakova u engleskim tekstovima, MacKay (2003): slika 2.1, stranica 22
- Različita frekvencija kolokacije dvaju znakova u engleskim tekstovima, MacKay (2003): slika 2.2, stranica 23

Kodiranje u binarni oblik

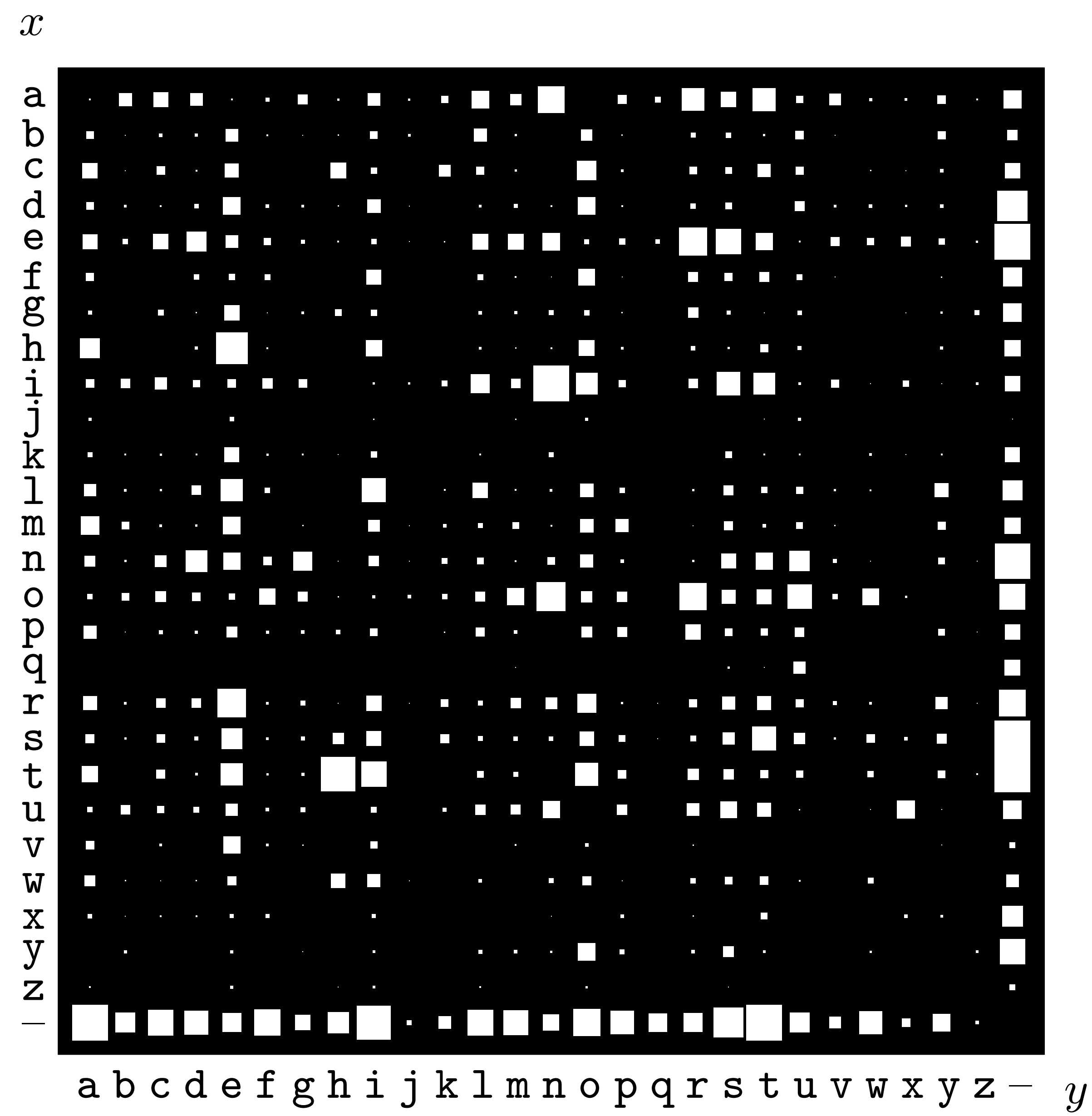
- MacKay (2003): slika 2.1, stranica 22

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928



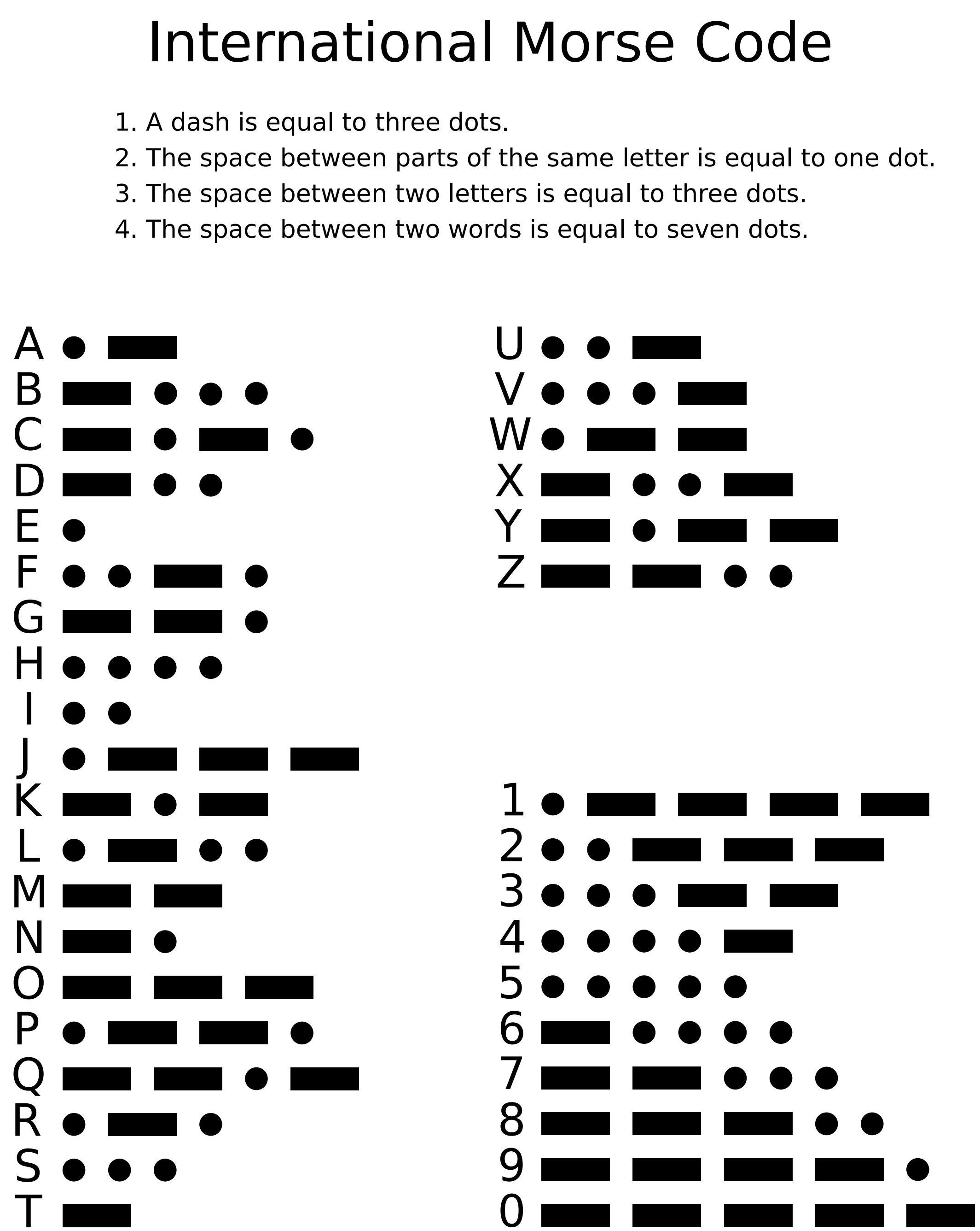
Kodiranje u binarni oblik

- MacKay (2003): slika 2.2, stranica 23



Kodiranje u binarni oblik

Morseov kod (Wikipedia)



Kodiranje u binarni oblik

- Ako je varijabla X niz vjerojatnosti:
  - c(0.21, 0.15, 0.10, 0.05, 0.20, 0.06, 0.04, 0.09, 0.05, 0.05)
  - i svaka vjerojatnost je dodjelena jednome znaku
- Jednadžba za taj niz vjerojatnosti je za n = broj mjera, a x konkretna vrijednost ili vjerojatnost:

$$-\sum_{i=1}^{n} x_i log_2(x_i)$$

Kodiranje u binarni oblik

- Ili općenito, **Entropija**, p(x) je vjerojatnost jednog znaka x:

$$-\sum_{i=1}^{n} p(x_i) log_2(p(x_i))$$

Kodiranje u binarni oblik

Primjer u R-u:

U R-u učitamo: `signfreq.txt`

```
podaci=read.csv(file=file.choose())
podaci
attach(podaci)
sum(vjerojatnost)
```

Kodiranje u binarni oblik

- Kodiranje entropije u R-u

$$-\sum_{i=1}^{n} x_i log_2(x_i)$$

- ako smo učitali CSV-tabelu i želimo korisiti podatke pod `vjerojatnost` ???

Kodiranje u binarni oblik

- Kod u R-u za entropiju:
  - `- sum( vjerojatnost * log2(vjerojatnost) )`

Kodiranje u binarni oblik

- Ako usporedimo entropiju distribucije i vjerojatnosti znakova koje smo učitali s uniformnom distribucijom:

```
c(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1,
0.1, 0.1, 0.1)
```

- što možemo zaključiti?

Kodiranje u binarni oblik

- Ako dodajemo male promjene u vjerojatnosti pojedinačnih znakova:

```
c(0.11, 0.09, 0.1, 0.1, 0.1, 0.1,
0.1, 0.1, 0.1, 0.1)
```

- kako se mijenja entropija?

Kodiranje u binarni oblik

- Koja je razlika između entropije za deset elemenata s istom vjerojatnosti, koja kada se vjerojatnosti razlikuju?
- Što možemo očekivati, hoće li ikada entropija biti veća u slučaju da nije vjerojatnost ista za sve znakove, tj. da nemamo uniformnu distribuciju?