Diplomski studij

Informacijska i komunikacijska tehnologija:

Obradba informacija
Telekomunikacije i informatika

Višemedijske komunikacije

3.

Informacijska svojstva i kodiranje jezika

Informacijska svojstva i kodiranje jezika

- Kolika je entropija prirodnog jezika?
- Kako kodirati tekst?

Kolika je entropija hrvatskog jezika?

- Osnovni simbol: slovo
- 27 slova (uključujući razmak):

$$H = \log 27 = 4{,}755 \text{ bit/simbol}$$

Vjerojatnost pojavljivanja pojedinih slova

znak	rel.frekv.	kod	znak	rel.frekv.	kod	znak	rel.frekv.	kod
razmak	0.1700	000	t	0.0367	10100	b	0.0155	11100
a	0.0960	001	u	0.0364	10101	z	0.0144	111010
e	0.0770	0100	d	0.0319	10110	š	0.0086	111011
o	0.0754	0101	m	0.0313	10111	č	0.0084	111100
i	0.0742	0110	v	0.0306	11000	c	0.0067	111101
n	0.0464	0111	l	0.0306	11001	h	0.0065	1111100
j	0.0435	1000	k	0.0298	11010	ž	0.0052	1111101
s	0.0420	10010	p	0.0204	110110	ć	0.0049	1111110
r	0.0382	10011	g	0.0166	110111	f	0.0011	1111111

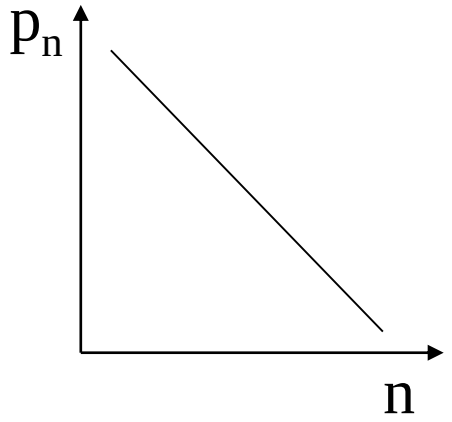
$$H = -\sum_{i=1}^{27} p(x_i) \log\, p(x_i) = 4{,}19$$

Korelacija među susjednim znakovima

- Npr. iza samoglasnika vjerojatniji suglasnik
- Neodređenost je smanjena
- Promatramo po dva susjedna znaka: H = 3,59
- Promatramo po tri susjedna znaka: H = 3,1

Entropija uz promatranje čitavih riječi

- Zipfov zakon: $p_n = \frac{P}{n}$
  - n: redni broj riječi, počevši od najčešćih
  - p_n: vjerojatnost pojavljivanja riječi n
  - P: konstanta



- Prosječni sadržaj informacije po riječi

$$\bar{I}_r = -\sum_{n=1}^{R} p_n \log p_n \left[bit / rije\check{c}\right]$$

- Dijelimo s prosječnim brojem slova po riječi; za engleski dobivamo 1,66 bit/simbol
- Korelacije među riječima, gramatika...: **0,6 – 1.3 bit/simbol**

Međutim...

- Metoda kodiranja zasnovana na ovim razmatranjima bila bi složena
- Gramatika, riječi, slova, vjerojatnosti pojave riječi i slova: sve ovisi o jeziku
- Jednostavnost i univerzalnost važnija od moguće uštede
- Stoga: ASCII (8 bit), Unicode (16 bit)