

Kontekstno-neodvisne gramatike za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

21. 11. 2022

Kodiranje in kod

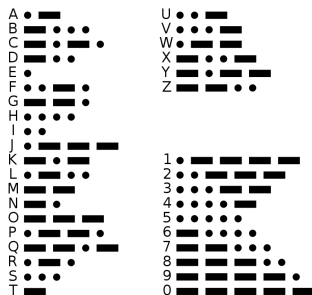
Zapis informacije v neki obliki ni primeren za vsakršno rabo. Spreminjanje zapisa sporočila pravimo *kodiranje*, sistemu pravil, po katerem se kodiranje opravi, pa *kod*.

Primer kodiranja

Morsejeva abeceda je kodiranje črk, števil in ločil s pomočjo zaporedja kratkih in dolgih signalov:

- Dolžina kratkega signala je ena enota.
- Dolgi signal je trikrat daljši od kratkega signala.
- Razmik med signali znotraj črke je tišina dolžine kratkega signala.
- Razmik med črkami je tišina dolga tri kratke signale oz. en dolgi signal.
- Presledek med besedami je tišina dolga sedmih kratkih signalov.

Morsejeva abeceda



Slika: Mednarodna Morsejeva abeceda

Abeceda in nizi na abecedi

Definicija

Abeceda je končna neprazna množica Σ . Elementom abecede pravimo črke. Množica vseh končnih nizov abecede Σ označimo z Σ^* in vključuje tudi prazen niz, ki ga označimo z ε . Dolžino niza w označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. Jezik na abecedi Σ je poljubna podmnožica množice Σ^* .

Abeceda in nizi na abecedi

Definicija

Abeceda je končna neprazna množica Σ . Elementom abecede pravimo črke. Množica vseh končnih nizov abecede Σ označimo z Σ^* in vključuje tudi prazen niz, ki ga označimo z ε . Dolžino niza w označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. Jezik na abecedi Σ je poljubna podmnožica množice Σ^* .

Primer

Naj bo $\Sigma = \{a, b, c\}$ abeceda, potem je

$$ab \in \Sigma^*$$

$$cababcccababcccab \in \Sigma^*$$

Kodiranje in dekodiranje

Definicija

Kodiranje nizov abecede Σ je injektivna funkcija $\kappa: \Sigma^ \rightarrow \Sigma_c^*$, kjer je Σ_c kodirna abeceda in $\kappa(w)$ imenujemo koda niza w . Dekodiranje kodiranja κ je funkcija $\kappa^{-1}: C \subseteq \Sigma_c^* \rightarrow \Sigma^*$, da velja*

$$\forall w \in \Sigma^*: \kappa^{-1}(\kappa(w)) = w$$

Formalizirzacija Morsejeve abecede

Abecedi sta

$$\Sigma = \{A, B, \dots, Z\} \cup \{0, 1, \dots, 9\} \cup \{-\}$$

$$\Sigma_c = \{., -, _\}$$

Definirajmo kodno funkcijo črk abecede $\kappa: \Sigma \rightarrow \Sigma_c^*$, ki vsakei črki iz abecede Σ priredi niz črk kodirne abecede Σ_c . Za niz $w = a_1 a_2 \dots a_n \in \Sigma^*$ definiramo kodirno funkcijo K po črkah

$$K(w) = \kappa(a_1)_- \kappa(a_2)_- \dots \kappa(a_n)$$

Formalizirzacija Morsejeve abecede

Vrednosti funkcije κ so določene s tabelo

A	• —	U	• • —
B	— • • •	V	• • — •
C	— • — •	W	• — • —
D	— • • •	X	— • — •
E	• — • •	Y	— • • —
F	• • — •	Z	— — • •
G	• — — •		
H	• • • •		
I	• • — —		
J	• — — —		
K	— • — •	1	• — — — —
L	— • • •	2	• • — — —
M	— — • •	3	• • • — —
N	— • — —	4	• • • • —
O	— — — •	5	• • • • •
P	• — — •	6	— • • • •
Q	• — • —	7	— • — • •
R	• • — —	8	— • — • •
S	• • • —	9	— • — • •
T	— • — —	0	— — — • —

Dodatno presledek med besedami _ kodiramo v sedem kratkih enot tišine

$$\kappa(_) = \text{_____}$$

Stiskanje podatkov

Definicija

Stiskanje je kodiranje K za katerega velja

$$\exists n \in \mathbb{N} \forall w \in \Sigma^*: |w| \geq n \implies |K(w)| \ll |w|$$

Zgled stiskanja niza w

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in pogledjmo niz

$$w = cababcccababcccab$$

Zgled stiskanja niza w

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in pogledjmo niz

$$w = cababcccababcccab$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$.

$$w = cAABAABA$$

Zgled stiskanja niza w

Za abecedo vzemimo $\Sigma = \{a, b, c\}$ in pogledjmo niz

$$w = cababcccababcccab$$

Uvedemo novi spremenljivki $A = ab$ in $B = ccc$.

$$w = cAABAABA$$

Uvedemo novo spremenljivko $C = AAB$.

$$w = cCCA$$

Zgled stiskanja niza w

Prešnji postopek napišemo na sledeč način s pomočjo produkcijskih pravil

$$S \rightarrow cCCA,$$

$$A \rightarrow ab,$$

$$B \rightarrow ccc,$$

$$C \rightarrow AAB$$

Kontekstno-neodvisne gramatike

Definicija

Formalna gramatika G so pravila, ki nam iz abecede Σ tvorijo jezik, označimo ga z $L(G)$

Kontekstno-neodvisne gramatike

Definicija

Formalna gramatika G so pravila, ki nam iz abecede Σ tvorijo jezik, označimo ga z $L(G)$

Definicija

Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je V končna množica *spremenljivk*, abeceda Σ množica *končnih simbolov* tako, da $\Sigma \cap V = \emptyset$, $P \subseteq V \times (V \cup \Sigma)^*$ relacija, ki ji pravimo *produkcijsko pravilo* in $S \in V$ *začetna spremenljivka*.

Kontekstno-neodvisne gramatike

Definicija

Naj bo $G = (V, \Sigma, P, S)$ kontekstno-neodvisna gramatika. Naj bodo $\alpha, \beta, \gamma \in (V \cup \Sigma)^*$ nizi spremenljivk in končnih simbolov, $A \in V$ spremenljivka ter naj bo $(A, \beta) \in P$ produkcijsko pravilo, označimo ga z $A \rightarrow \beta$. Pravimo, da se $\alpha A \gamma$ *prepiše s pravilom* $A \rightarrow \beta$ v $\alpha \beta \gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha \beta \gamma$. Pravimo, da α *porodi* β , če je $\alpha = \beta$ ali če za $k \geq 0$ obstaja zaporedje $\alpha_1, \alpha_2, \dots, \alpha_n \in (V \cup \Sigma)^*$ tako, da

$$\alpha \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \dots \Rightarrow \alpha_n \Rightarrow \beta$$

in pišemo $\alpha \xRightarrow{*} \beta$.

Kontekstno-neodvisne gramatike

Posledica

Jezik kontekstno neodvisne gramatike G je

$$L(G) = \{w \in \Sigma^* \mid S \xRightarrow{*} w\}$$

Zgled stiskanja niza w z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$.

Zgled stiskanja niza w z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$. Označimo jo z

$G_w = (V, \Sigma, P, S)$, kjer je

$$V = \{S, A, B, C\},$$

$$\Sigma = \{a, b, c\},$$

$$P = \{S \rightarrow cCCA, A \rightarrow ab, B \rightarrow ccc, C \rightarrow AAB\},$$

$$S = S$$

Zgled stiskanja niza w z zapisom gramatike

Formalizirajmo gramatiko iz prejšnjega primera. Gramatiko smo generirali z nizom $w = cababcccababcccab$. Označimo jo z

$G_w = (V, \Sigma, P, S)$, kjer je

$$V = \{S, A, B, C\},$$

$$\Sigma = \{a, b, c\},$$

$$P = \{S \rightarrow cCCA, A \rightarrow ab, B \rightarrow ccc, C \rightarrow AAB\},$$

$$S = S$$

Vidimo, da G_w ustreza naši definiciji kontekstno-neodvisne gramatike in res kodira w , saj je

$$L(G_w) = \{w\}$$