KONTEKSTNO-NEODVISNE GRAMATIKE ZA KODIRANJE IN STISKANJE PODATKOV

JANEZ PODLOGAR

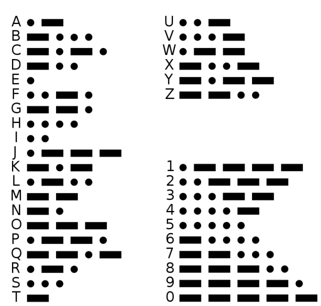
Kazalo

1. Kodiranje podatkov

Zapis informacije v neki obliki ni primeren za vsakršno rabo. Besedilo, zapisano z pismenkami, je neberljivo za slepe osebe, saj je komunikacijski kanal v tem primeru vid. Prav tako pisanega besedila v prvotni obliki ni mogoče poslati s telegrafom. V tem primeru je komunikacijski kanal žica in pismenke se po njej ne morejo sprehoditi. V obeh primerih je informacija, ki bi jo radi prenesli, zapisana v neprimerni obliki. V prvem primeru je potrebno besedilo zapisati z Braillovo pisavo. V drugem primeru pa je besedilo potrebno pretvoriti v električni signal. Spreminjanje zapisa sporočila imenujemo *kodiranje*, sistemu pravil, po katerem se kodiranje opravi, pa *kod*.

Primer 1.1. *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 oziroma en dolgi signal.
- Presledek med besedami je tišina dolga sedem kratkih signalov.



Slika 1. Mednarodna Morsejeva abeceda.

Date: 17. september 2023.

Prvotni namen Morsejeve abecede je komunikacija preko telegrama, saj komunikacijski kanal dovoljuje le električne signale in tišino med njimi. Kodiranje črk je takšno, da imajo črke z višjo frekvenco (v angleškem jeziku) krajši zapis. Tako se koda sporočila skrajša in posledično tudi čas njegovega prenosa.

$\diamondsuit$

Definicija 1.2. *Abeceda* je končna neprazna množica. Elementom abecede pravimo *črke*. Za abecedo Σ definiramo

$$\Sigma^0 = \{\varepsilon\}$$

in ε imenujemo *prazen niz*. Za vsak $\ell > 0$ rekurzivno definiramo *množico vseh nizov abecede* Σ *dolžine* $\ell + 1$

$$\Sigma^{\ell+1} = \{wa \mid w \in \Sigma^\ell \text{ in } a \in \Sigma\}.$$

Nadalnje, definiramo *množica vseh končnih nizov abecede* Σ

$$\Sigma^* = \bigcup_{\ell \geq 0} \Sigma^\ell$$

in *množica vseh končnih nizov abecede* Σ *brez praznega niza*

$$\Sigma^+ = \bigcup_{\ell > 0} \Sigma^\ell.$$

Jezik na abecedi Σ je poljubna podmnožica množice Σ^*.

Definicija 1.3. Naj bo Σ abeceda. Naj bo $*$ binarna operacija na množici vseh končnih nizov Σ^* tako, da je prazen niz ε nevtralni element in za niza $w, u \in \Sigma^*$ velja

$$w * u = w_1 w_2 \cdots w_n u_1 u_2 \cdots u_m,$$

kjer sta $w_1 w_2 \cdots w_n$ in $u_1 u_2 \cdots u_m$ predstavitvi nizov w in u s črkami abecede Σ. Operacijo $*$ imneujemo *stikanje* oziorma *konkatenacija*. Znak $*$ spustimo in krajše pišemo wu.

Opomba 1.4. Stikanje je asociativna operacija. $(\Sigma^*, *)$ je monoid in $(\Sigma^+, *)$ je grupoid.

Opomba 1.5. *Kleenejeva zvezdica* oziroma *Kleenejevo zaprtje* je enočlena operacija, ki abecedi Σ priredi najmanjšo nadmnožico Σ^*, ki vsebuje *prazen niz* ε in je zaprta za operacijo stikanje. Z drugimi besedami, Σ^* je množica vseh končnih nizov, ki jih lahko generiramo z stikanjem črk abecede Σ.

Definicija 1.6. *Dolžino niza* w označimo z $|w|$ in je enaka številu črk v nizu $w \in \Sigma^*$. Natančneje, $|w| = \ell$ natanko tedaj, ko je $w \in \Sigma^\ell$.

Primer 1.7. Naj bo $\Sigma = \{a, b, c\}$ abeceda, potem so $ab \in \Sigma^2$, $ccc \in \Sigma^3$ in $cababcccababcccab \in \Sigma^{17}$ končni nizi abecede Σ in potemtakem elementi Σ^*.

$\diamondsuit$

Definicija 1.8. *Kodiranje nizov abecede* Σ je injektivna funkcija $\kappa \colon \Sigma^* \to \Sigma_c^*$, kjer je Σ_c^* neka abeceda, ki jo imenujemo Σ_c *kodna abeceda*, in $\kappa(w)$ imenujemo *koda niza* w. *Dekodiranje kodiranja* κ je funkcija $\delta \colon C \subseteq \Sigma_c^* \to \Sigma^*$, da velja

$$\forall w \in \Sigma^* \colon \delta(\kappa(w)) = w.$$

Opomba 1.9. Funkcijo κ imenujemo *kodna funkcija*, funkcijo δ pa *dekodna funckija*.

Opomba 1.10. Zožitev kodomene kodne funkcije κ na $C \subseteq \Sigma_c^*$ je bijektivna funkcija.

Primer 1.11. Formalizirajmo Morsejevo abecedo iz Primera 1.1. Abecedi sta

$$\Sigma = \{\mathrm{A}, \mathrm{B}, \ldots, \mathrm{Z}\} \cup \{0, 1, \ldots, 9\} \cup \{\textvisiblespace\}, \quad \Sigma_c = \{\cdot, -, \square\},$$

kjer je $\textvisiblespace$ presledek in $\square$ ena kratka enota tišine. Definirajmo kodno funkcijo črk abecede $\kappa_s \colon \Sigma \to \Sigma_c^*$, ki vsaki črki iz abecede Σ_s priredi niz črk kodne abecede Σ_c. Predpis funkcije κ_s je določen s tabelo iz Slike 1, dodatno presledek $\textvisiblespace$ kodiramo v tri kratkih enot tišine

$$\kappa_s(\textvisiblespace) = \square\square\square\square.$$

Za niz $w = a_1 a_2 \ldots a_n \in \Sigma^*$ definiramo kodno funkcijo K po črkah

$$\kappa(w) = \kappa_s(a_1)\square\square\square\kappa_s(a_2)\square\square\square \cdots \kappa_s(a_n).$$

Poglejmo si dva primera kodiranja v Morsejevi abecedi

$$\kappa(\mathrm{SOS}) = \,\cdot\square\cdot\square\cdot\square\square\square-\square-\square-\square\square\square\cdot\square\cdot\square\cdot\,,$$
$$\kappa(\mathrm{AD}\textvisiblespace\mathrm{HOC}) = \,\cdot\square-\square\square\square-\square\cdot\square\cdot\square\square\square\square\square\square\square\cdot\square\cdot\square\cdot\square\cdot\square\square\square-\square-\square-\square\square\square-\square\cdot\square-\square\cdot\,.$$

Recimo, da smo prejeli sporočilo, a se je pošiljatelj zmotil in je namesto kode, ki bi se dekodirala v

$$\delta(\,-\square-\square\cdot\square-\square\square\square\cdot\square\square\square-\square\cdot\square\cdot\,) = \mathrm{QED},$$

poslali kodo

$$-\square-\square\cdot\square-\square-\square\square\square\cdot\square\square\square-\square\cdot\square\cdot\,.$$

Sporočila ne znamo dekodirati, saj se ne nahaja v domeni C dekodne funkcije δ.

$\diamondsuit$