Digitalni zapis podataka

Stefan Mišković

8 Zapis teksta u računaru

8.1 Azbuka

Konačna azbuka V je konačni neprazni skup simbola. Elementi skupa V se nazivaju simboli ili slova. Konačne niske slova iz V se nazivaju reči. Prazna reč je reč koja ne sadrži nijedno slovo i najčešće se označava sa λ. Skup svih reči nad azbukom V se označava sa V^*, a skup svih nepraznih reči sa V^+. Važi da je $V^+ = V^* \setminus \{\lambda\}$. Jezik L je proizvoljan skup reči iz V^*, odnosno proizvoljan podskup tog skupa. Dužina reči p, u oznaci $|p|$, predstavlja broj simbola u reči p. Važi da je $|\lambda| = 0$. Oznaka p^i predstavlja i puta dopisanu reč p. Na primer, $p^3 = ppp$. Posebno je $p^0 = \lambda$. Ako azbuka ima n znakova, broj reči u njoj koji imaju dužinu d iznosi n^d.

Primeri nekih azbuka i jezika nad njima:

- Za azbuku $V = \{a, b\}$ jezik koji sadrži sve reči do dužine 1 je $L_1 = \{\lambda, a, b\}$, a jezik koji sadrži sve reči do dužine 2 je $L_2 = L_1 \cup \{aa, ab, ba, bb\}$.
- Za azbuku $V = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ jezik $L = V^+$ svih nepraznih reči te azbuke predstavlja cele brojeve u dekadnom sistemu sa eventualnim vodećim nulama.
- Za azbuku $V = \{0, 1\}$ jezik $L = V^+$ svih nepraznih reči te azbuke predstavlja cele brojeve u binarnom sistemu sa eventualnim vodećim nulama.

8.2 Kodovi

Neka su date dve azbuke V_1 i V_2 i dva jezika nad njima L_1 i L_2. Funkcija kodiranja je svaka funkcija $f : L_1 \to L_2$ koja slika jezik L_1 u jezik L_2. Ukoliko postoji inverz f^{-1}, može se definisati i funkcija dekodiranja $g = f^{-1}$. Funkcija dekodiranja $g : L_2 \to L_1$, nasuprot funkciji f, slika jezik L_2 u jezik L_1. Kodiranje predstavlja izračunavanje vrednosti $f(p)$ za neku reč $p \in L_1$, a dekodiranje izračunavanje vrednosti $g(q)$ za neku reč $q \in L_2$. Kod jezika L_1 je skup svih takvih vrednosti $f(p)$, $p \in L_1$, a može se označiti i za $f(L_1)$. Važi da je $f(L_1) \subseteq L_2$, ali ta dva skupa ne moraju da se poklapaju (ne moraju sve reči iz jezika L_2 da budu slike reči iz L_1). Kod može imati sledeće osobine:

- Kod je jednoznačan ako je funkcija f 1-1. Inače, kod je višeznačan.
- Kod je ravnomeran ako je dužina svih njegovih reči ista.
- Kod je potpun ako obuhvata sve reči određene dužine u jeziku L_2.

Primeri nekih kodova:

- Neka su definisane azbuke $V_1 = \{+, -, *, /\}$ i $V_2 = \{0, 1\}$ i jezici nad njima $L_1 = V_1 = \{+, -, *, /\}$ i $L_2 = \{00, 01, 10, 11\}$. Vidimo da se jezik L_1 poklapa sa azbukom V_1, a da je jezik L_2 skup svih reči dužine 2 iz azbuke V_2. U narednoj tabeli su prikazani primeri nekih funkcija kodiranja $f : L_1 \to L_2$. Svi dobijeni kodovi su jednoznačni, ravnomerni i potpuni.

Reč	Kod 1	Kod 2	Kod 3
+	00	10	11
−	01	11	00
*	10	01	10
/	11	00	01

- Neka su definisane azbuke $V_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ i $V_2 = \{0, 1\}$. Neka je $L_1 = V_1$, L_2 skup svih reči dužine 4 iz V_2 i neka je funkcijom kodiranja $f : L_1 \to L_2$ definisan kod koji odgovara 8421 zapisu (BCD kodiranje). Taj kod je jednoznačan i ravnomeran, ali nije potpun. Iako su sve vrednosti funkcije jednake dužine, nisu obuhvaćene sve reči dužine 4.

8.3 Kodne strane

Tekst se obično zamišlja kao dvodimenzionalni objekat, ali se u računaru predstavlja kao jednodimenzionalni niz karaktera. Pritom je potrebno uvesti i specijalne karaktere koji označavaju prelazak u novi red, tabulator, kraj teksta, itd. Ti specijalni karakteri se u računaru tretiraju na isti način kao slova, cifre i ostali karakteri. Osnovna ideja kod zapisivanja teksta u računaru je da se svakom karakteru pridruži odgovarajući ceo broj na unapred definisan način. Uređena lista karaktera predstavljena svojim kodovima se naziva kodna strana.

Primeri kodnih strana (jednobajtni strandard):

- ASCII. U ASCII kodu se može zapisati 128 različitih karaktera. Svakom karakteru se dodeljuje odgovarajuća sedmobitna niska. ASCII je jednobajtni standard, pa se vodeća cifra zapisa svakog karaktera ne koristi. Između ostalog, tu su predstavljeni kontrolni karakteri, velika slova engleske abecede, mala slova engleske abecede, dekadne cifre i neki interpunkcijski znakovi.

- ISO 8859-1 (Latin 1). Pomoću ove kodne strane se može zapisati 256 različitih karaktera i svakom karakteru se dodeljuje osmobitna niska. Na prvih 128 mesta se poklapa sa ASCII kodom. Pomoću nje se zapisuju svi znakovi zapadnoevropskih latinica.

- ISO 8859-2 (Latin 2). Ova kodna strana ima slične karakteristike kao Latin 1, s tim što se u ovom slučaju mogu zapisati svi znakovi istočnoevropskih latinica, uključujući srpsku.

- ISO 8859-5. Ova kodna strana ima slične karakteristike kao Latin 1, s tim što se u ovom slučaju mogu zapisati svi znakovi istočnoevropskih ćirilica, uključujući srpsku.

Primeri kodnih strana (višebajtni strandard):

- Unicode UCS-2. Ovaj standard svakom karakteru dodeljuje dvobajtni kod. Prvih 256 karaktera se poklapa sa Latin 1 standardom, a na preostalim mestima su, izmedu ostalih, obuhvaćeni karakteri današnjih jezika. U njemu mogu da se zapišu srpska latinična i ćirilična slova.

- Unicode UTF-8. UTF-8 algoritmom se svakom karakteru dodeljuje 1 ili više bajtova (najviše 4 bajta), s tim što su kodovi definisani tako da je onim karatkerima koji se češće javljaju dodeljeno manje bajtova, a onim koji se ređe javljaju više bajtova. Na primer, svi ASCII karakteri se zapisuju pomoću jednog bajta. U njemu takođe mogu biti zapisana sva slova srpske latinice i ćirilice.