Języki i metody programowania I

dr inż. Piotr Szwed
Katedra Informatyki Stosowanej
C2, pok. 403

e-mail: pszwed@agh.edu.pl

http://home.agh.edu.pl/~pszwed/

Aktualizacja: 2013-01-18

8. Łańcuchy znaków

Łańcuchy znaków – wprowadzenie (1)

- W języku C/C++ brak jest specjalnego typu danych dla reprezentacji napisów. Każdy napis jest traktowany jako ciąg znaków. Przyjętą reprezentacją napisu jest tablica znaków.
- Standardowo, znak jest reprezentowany przez jeden bajt. Takie założenie było przez długie lata wystarczające. Liczba symboli graficznych wymaganych w aplikacjach języku angielskim doskonale mieści się w zakresie od 0-127. Pozostałe znaki były używane do reprezentacji znaków specjalnych (np.: elementów ramek)
- Języki europejskie wymagają dodatkowych znaków, którym przydzielono kody powyżej 127. Niestety, układ symbole graficznych poszczególnych grup języków może ze sobą kolidować. (np.: zachodnioeuropejskich i środkowoeuropejskich).

Łańcuchy znaków – wprowadzenie (2)

- Odwzorowanie wartości bajtów w postać symboli graficznych uzależnione jest od używanej *strony kodowej* .
- Strony kodowe ISO 8859-2 (norma) oraz Windows 1250 (źródło: Wikipedia)

ISO/IEC 8859-2:1999

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*Znaki kontrolne*															
1x																
2x	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*Nieużywane*															
9x																
Ax	NBSP	Ą	˘	Ł	¤	Ľ	Ś	§	¨	Š	Ş	Ť	Ź	SHY	Ž	Ż
Bx	°	ą	˛	ł	´	ľ	ś	ˇ	¸	š	ş	ť	ź	˝	ž	ż
Cx	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
Dx	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
Ex	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
Fx	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

Windows-1250

	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1x	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2x	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
8x	€	*NZ*	‚	*NZ*	„	…	†	‡	*NZ*	‰	Š	‹	Ś	Ť	Ž	Ź
9x	*NZ*	‘	’	“	”	•	–	—	*NZ*	™	š	›	ś	ť	ž	ź
Ax	NBSP	ˇ	˘	Ł	¤	Ą	¦	§	¨	©	Ş	«	¬	SHY	®	Ż
Bx	°	±	˛	ł	´	µ	¶	·	¸	ą	ş	»	Ľ	˝	ľ	ż
Cx	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
Dx	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
Ex	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
Fx	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	˙

Łańcuchy znaków – wprowadzenie (3)

- W językach azjatyckich zawsze posługiwano się większą liczbą znaków. Z tego powodu używa się reprezentacji mieszanej napisów zawierającej zarówno znaki jedno i dwubajtowe..
- Nowszym standardem jest standard UNICODE. Każdy znak jest reprezentowany przez 16-bitową liczbę bez znaku. Standard UNICODE pokrywa symbole graficzne rozmaitych języków i pozwala na ich równoczesne użycie.

Hex	Znak	Unicode	Hex	Znak	Unicode
0xA1	Ą	U+0104	0xB1	ą	U+0105
0xC6	Ć	U+0106	0xE6	ć	U+0107
0xCA	Ę	U+0118	0xEA	ę	U+0119
0xA3	Ł	U+0141	0xB3	ł	U+0142
0xD1	Ń	U+0143	0xF1	ń	U+0144
0xD3	Ó	U+00D3	0xF3	ó	U+00F3
0xA6	Ś	U+015A	0xB6	ś	U+015B
0xAF	Ż	U+017B	0xBF	ż	U+017C
0xAC	Ź	U+0179	0xBC	ź	U+017A

Łańcuchy znaków – wprowadzenie (4)

- W wersji podstawowej, tablice znaków języka C/C++ są ciągami 8-bitowych wartości. Odwzorowanie kodów znaków w symbole graficzne pozostawione jest parametrom sterującym aplikacją (np.: użytemu fontowi, stronie kodowej przyjętej dla systemu operacyjnego).
- Znakiem szczególnym jest znak o zerowej wartości. Nigdy nie ma on reprezentacji graficznej i pełni funkcję znacznika specjalnego (ang. *sentinel*).
- W języku C/C++ napisy reprezentowane są jako ciągi 8-bitowych znaków zakończone dodatkowym znakiem zerowym (znacznikiem końca).

Przykłady

- Stała `"Tekst"` jest reprezentowana jako tablica znaków umieszczona w segmencie danych.

T	e	k	s	t	0

- Deklaracja tablicy znakowej z inicjalizacją

  `char text[]="Tekst";`

  Kompilator automatycznie przydzieli tablicy `text` 6 znaków odpowiednio ustawiając znaki.

  `char text[256]="Tekst";`

  Kompilator przydzieli tablicy `text` 256 znaków. Pierwszych sześć znaków zostanie zainicjowanych, pozostałe będą miały wartość zero. W tablicy można umieszczać teksty zawierające 255 znaków (należy zarezerwować miejsce na ostatni znak **0**).

Funkcje działające na tablicach znakowych

- Funkcje działające na tablicach znakowych zdefiniowane są w pliku nagłówkowym `<string.h>`
- Większość z nich zakłada, że łańcuchy znakowe są zakończone znakiem zerowym.

Funkcja strlen

```
size_t strlen( const char *string );
```

Funkcja oblicza długość łańcucha znakowego.

```
unsigned mystrlen( const char *string )
{
        unsigned i;
        for(i=0;string[i];i++);
        return i;
}
```

Funkcja strcpy

```
char *strcpy( char *dest, const char *source );
```

- Funkcja kopiuje zawartość łańcucha znakowego `source` do tablicy `dest`.
- Tablica docelowa musi mieć rozmiar ≥ `strlen(source)+1`.
- Działanie funkcji w przypadku nakładania się tablic `source` oraz `dest` jest nieokreślone.

```
char *mystrcpy(char *dest, const char *source )
{
        int i;
        for(i=0; source[i];i++){
                dest[i]= source[i];
        }
        dest[i]= 0;
        return dest;
}
```

Funkcja strcat

```
char *strcat( char *dest, const char *source );
```

Funkcja dodaje na końcu łańcucha `dest` łańcuch `source` . Tablica docelowa musi mieć rozmiar
≥ `strlen(source)+ strlen(dest)+1`.

```
char *mystrcat(char *dest, const char *source )
{
        char*pdest;
        const char*psrc;
        for(pdest=dest;*pdest; pdest++);
        for(psrc = source; * psrc; psrc ++,pdest++){
                *pdest = *psrc;
        }
        *pdest = 0;
        return dest;
}
```

Funkcja strcmp

```
int strcmp( const char *str1, const char *str2 );
```

Funkcja porównuje zawartość dwóch tablic `str1` oraz `str2`. Zwraca:

- Wartość < 0 - jeżeli `str1` < `str2`
- 0 - jeżeli łańcuchy są identyczne
- Wartość > 0 - jeżeli `str1` > `str2`

```
int mystrcmp( const char *str1, const char *str2 )
{
	for(;*str1 && *str2;str1++,str2++){
		if(*str1<*str2)return -1;
		if(*str1>*str2)return 1;
	}
	if(*str1 && ! *str2) return -1; // aX < a
	if(!*str1 && *str2) return 1; // a > aX
	return 0;
}
```

Funkcja strcoll

```
int strcoll( const char *str1, const char *str2 );
```

Funkcja – podobnie jak `strcmp` - porównuje zawartość dwóch tablic `str1` oraz `str2`, ale interpretuje teksty zgodnie z ustawieniami regionalnymi (locale). Dzięki temu możliwe jest np.: porównanie polskich tekstów.

Funkcja `setlocale` ustala wszystkie (`LC_ALL`) lub wybrane składniki informacji regionalnych (`LC_COLLATE, LC_CTYPE, LC_MONETARY, LC_NUMERIC, LC_TIME`)

```
int main(){
    setlocale(LC_ALL,""); // setlocale(LC_ALL,"C")
    char*s1= "kąt";
    char*s2="kbt";
    printf("%s %d %s",s1,strcoll(s1,s2),s2);
    return 0;
}
```

```
setlocale(LC_ALL,"");

kąt -1 kbt
```

```
//setlocale(LC_ALL,"C");

k╣t 1 kbt
```

Wydzielanie symboli (1)

- Bardzo często w programach występuje konieczność analizy łańcuchów znaków i wydzielenia z nich symboli składowych (np.: słów, słów kluczowych, liczb).
- W szczególnym prostym przypadku tekst może być traktowany jako ciąg symboli oddzielonych separatorami.
- Oznaczmy:
  - *A* –zbiór znaków ASCII {1..255}
  - *S* – zbiór separatorów
  - Symbolami będą dowolne podłańcuchy zawierające znaki ze zbioru $A \setminus S$.
- Funkcją pozwalającą na wydzielenie tak zdefiniowanych symboli (ang. *token*) jest funkcja `strtok()` .

Wydzielanie symboli (2)

Funkcja strtok

```
char * strtok ( char* str, const char* delimiters );
```

- Wielokrotne wywołanie funkcji wydziela kolejne symbole z tesktu `str`.
- Podczas pierwszego wywołania do fukcji dostraczny jest wskaźnik `str`. Musi on wskazywać modyfikowalny tekst (funkcja umieszcza zera po kolejnych symbolach): `strtok(buf," \t\n.,")`
- Podczas kolejnych wywołań nie podaje się już adresu bufora (parametrem jest `NULL`): `strtok(NULL," \t\n.,")`
- Zbiory separatorów `delimiters` mogą być różne dla kolejnych wywołań.
- Funkcja zwraca `NULL` (`0`), jeżeli nie ma już więcej symboli do wydzielenia

Wydzielanie symboli (3)

Przykład

```
#include <string.h>
int main()
{
  char buf[255]; // bufor dla funkcji strtok
  const char text[]="To jest\ttekst. Słowa są"
  "oddzielone\nbiałymi znakami";
  char *ptr;
  const char sep1[]=" \n\t.,:!?";
  const char sep2[]=".?!";
  // Słowa
  strcpy(buf,text);  // strtok niszczy bufor!
  for( ptr=strtok(buf,sep1);       // inicjalizacja
       ptr;                        // czy wydzielono symbol?
       ptr=strtok(NULL,sep1))      // nast. symbol
       {
              printf("%s\n",ptr);
       }
```

Wydzielanie symboli (4)

Kontynuacja…

```
        // Zdania
        //  const char sep2[]=".?!";
        strcpy(buf,text);
        for(   ptr=strtok(buf,sep2);
                        ptr;
                        ptr=strtok(NULL,sep2)){
               printf("%s\n",ptr);
        }
        return 0;
}
```

Inne funkcje

- Istnieją wersje funkcji , które ograniczają porównanie, kopiowanie do określonej liczby znaków:
  - `int strncmp ( const char * str1, const char * str2, size_t num );`
  - `char * strncpy ( char * destination, const char * source, size_t num );`
- Funkcje do porównywania mogą ignorować duże małe litery: `stricmp()`. Standardowo, znaki diakrytyczne nie są poddawane translacji, stąd `stricmp("trąba","TRĄBA") ≠ 0`. To zachowanie może być zmienione przez ustawienie *locale*, zmiennej odpowiedzialnej za określenie rodzaju języka.
- Istnieje szereg funkcji, które działają na tablicach bajtów, ale nie zakładają, że są one tekstami zakończonymi znakiem `0`. Ich dodatkowym parametrem jest zawsze wielkość tablicy: `memXXX()`:
  - `void * memcpy ( void * destination, const void * source, size_t num );`
  - `int memcmp ( const void * ptr1, const void * ptr2, size_t num );`

Szerokie znaki (1)

- Programy w języku C mogą posługiwać się rozszerzoną reprezentacją znaków (w zasadzie zgodną ze standardem Unicode).
- W pliku nagłówkowym `<wchar.h>` zdefiniowano:
  - nowy typ znakowy `wchar_t` – odpowiadający `short`
  - oraz kilkadziesiąt funkcji obejmujących:
    - Formatowane wejście wyjście (`wprintf, putwchar, swscanf`)
    - Funkcje konwersji: `wcstod, wcstof, wcstol`
    - Funkcje manipulujące łańcuchami znaków (utrzymano zasadę, że zerowy szeroki znak kończy łańcuch).

Szerokie znaki (2)

- Przykład

```
int main(){
    wchar_t tab[100]=L"Ala ma kota";
    setlocale(LC_ALL,"");
    printf("Tablica zajmuje %d bajtów\n",sizeof(tab));
    wprintf(L"Zawarto\u015b\u0107 tablicy: %ls\n",tab);
    printf("Długość tablicy: %d\n",wcslen(tab));

    wcscat(tab, L" i  psa");
    wprintf(L"Zawarto\u015b\u0107 tablicy\
            po konkatencaji: %ls\n",tab);
    return 0;
}
```

```
Tablica zajmuje 200 bajtów
Zawartość tablicy: Ala ma kota
Długość tablicy: 11
Zawartość tablicy po konkatencaji: Ala ma kota i  psa
```

Typowe błędy

- Najczęściej spotykane błędy przy posługiwaniu się funkcjami działającymi na tablicach znakowych to użycie wskaźnika, który zawiera adres nieokreślony lub przekroczenie rozmiarów tablicy.
- Język C/C++ nie ma żadnych wbudowanych mechanizmów ochrony przed tego typu błędami. Rozmiary tablicy nie są sprawdzane w trakcie wykonania.

```
char *table1;
strcpy(table1,”Tekst”);

char *table2;
strcmp(table2, ”Tekst”);

char table3[]=”Ala ma ” ;
strcat(table3,”kota”);

char *table4 =”x=%d”;
strcpy(table4,”Txt”); /* błąd, modyfikujemy pamięć, do
której nie mamy dostępu do zapisu */
```