PROGRAMIRANJE 1
ZAPIS TEKSTOVA U RAČUNARU

Staša Vujičić

v 1.1

ŠTA JE TO TEKST?

- Tekst (ili dokument) je
  - "informacija namenjena ljudskom sporazumevanju koja može biti prikazana u dvodimenzionalnom obliku... Tekst se sastoji od grafičkih elemenata kao što su karakteri, geometrijski ili fotografski elementi ili njihove kombinacije, koji čine sadržaj dokumenta." (ISO-definicija)

TEKST JE NIZ KARAKTERA

- Iako obično tekst zamišljamo kao dvodimenzioni objekat, u računarima se tekst predstavlja kao jednodimenzioni (linearni) niz karaktera.
- Potrebno je, dakle, uvesti specijalne karaktere koji označavaju prelazak u novi red, tabulator, kraj teksta i slično

ZAPIS KARAKTERA U RAČUNARU

- Računari su zasnovani na binarnoj aritmetici
- Cele brojeve je moguće predstaviti u binarnom sistemu
- Osnovna ideja je svakom karakteru pridružiti određeni ceo broj na unapred dogovoreni način
- Ove brojeve zovemo ***kodovima karaktera*** (character codes)

GRAFIČKA REPREZENTACIJA KARAKTERA

- Glif – grafička reprezentacija karaktera
- Font – skup glifova

KOLIKO KARAKTERA ŽELIMO DA PREDSTAVIMO U RAČUNARIMA?

- Tokom razvoja računarstva broj karaktera je postajao sve veći
- Pošto je u početku razvoja englesko govorno područje bilo dominantno osnovno je bilo predstaviti sledeće karaktere:

ENGLESKO GOVORNO PODRUČJE

- Velika slova engleskog alfabeta : A,B,...,Z
- Mala slova engleskog alfabeta : a,b,...,z
- Cifre : 0,1,...,9
- Interpunkcijske znake : .,:;'+*-_ i slično
- Specijalne znake : kraj reda, tabulator i slično

STANDARDNI KARAKTERSKI KODOVI

- Šezdesetih godina su se pojavile tabele standardnih karakterskih kodova dovoljne za zapis pomenutih karaktera
- Najpoznatiji su
- EBCDIC – IBM-ov standard, korišćen uglavnom na mainframe računarima, pogodan za bušene kartice
- ASCII – Standard iz koga se razvila većina današnjih standarda

ASCII

- **ASCII (*American Standard Code for Information Interchange*)**
- Uspostavljen od strane organizacije ANSI (American National Standard Institute)
- ASCII sedmobitan (broj karaktera je 128)

ASCII TABELA

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	STX	SOT	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

PRIMERI

- Karakter *A* se zapisuje kao $(41)_{16}$ tj. 0x41 što je $(65)_{10}$ tj. $(1000001)_2$
- Razmak se zapisuje kao $(20)_{16}$ što je $(32)_{10}$ tj. $(0100000)_2$

1. Zapišite cifru *3* u ASCII kodu
2. Zapišite tekst *Fakultet* u ASCII kodu

REŠENJE:

- Cifra: *3*
- Heksadekadno ASCII: *33*
- Dekadno ASCII: *51*
- Binarno ASCII: *0110011*

REŠENJE:

- Tekst: *Fakultet*
- Heksadekadno ASCII:

  46 61 6B 75 6C 74 65 74
- Dekadno ASCII:

  70 97 107 117 108 116 101 116
- Binarno ASCII:

  1000110 1100001 1101011 1110101 1101100 1110100 1100101 1110100

OZNAKA ZA KRAJ REDA

- Oznaka za kraj reda se ne zapisuje isto u svim operativnim sistemima
- Pod Windows ova se oznaka se zapisuje sa dva karaktera (CR LF - Carriage return Line feed), 0xD 0xA tj. 13 10 – istorijski razlozi (stari štampači)
- Unix koristi samo karakter CR tj. 0xD

ŠTA SA OSTALIM JEZICIMA?

- Razvojem računarstva se javlja potreba kodiranja tekstova i na drugim jezicima
- Kroz istoriju su postojala mnoga rešenja, od kojih su se neka zadržala, a neka su nestala

KODNE STRANE

- Pod *kodnom stranom* (*Code page*) tj. *skupom karaktera* (*Character set*, *charset*) podrazumevamo uređenu listu karaktera predstavljenih svojim karakterskim kodovima

KODNE STRANE

- Podaci se u računarima obično zapisuju bajt po bajt
- ASCII je sedmobitni standard
- ASCII karakteri se zapisuju tako što se u svakom bajtu bit najveće težine postavi na 0
- To ostavlja prostor za novih 128 karaktera čiji binarni zapis počinje sa 1

KODNE STRANE

- Ovaj prostor se može popuniti na razne načine
- Rešenje nije univerzalno, jer svakako na svetu postoji više od 256 različitih karaktera
- Postavljeni su razni standardi dopunjavanja ovih 128 karaktera
- Svim ovim kodnim stranama je zajedničko prvih 128 karaktera i oni se poklapaju sa ASCII

KODNE STRANE

- Ovako napravljene kodne strane obično omogućuju kodiranje tekstova na više srodnih jezika (obično i geografski bliskih)
- Nama su uglavnom važne kodne strane napravljene za centralno-evropske (Central European) latinice, kao i ćirilične kodne strane

NAJČEŠĆE KORIŠĆENE KODNE STRANE KOD NAS

- ISO 8859-2 (Latin2)
- ISO 8859-5 (Ćirilična)
- Windows 1250
- Windows 1251 (Ćirilična)

- Prve dve su delo međunarodne organizacije za standardizaciju (ISO - International Standard Organization), dok su naredne dve Microsoft-ovi standardi

LATIN 1

- Poželjno je poznavati i osnovnu kodnu stranu **ISO 8859-1 (Latin1)** jer je veoma često postavljena kao podrazumevana kodna strana. Ona se koristi za zapis tekstova na zapadno evropskim jezicima (Western European)

ISO 8859-1 (LATIN1)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	

ISO 8859-2

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	Ą	˘	Ł	¤	Ľ	Ś	§	¨	Š	Ş	Ť	Ź	SHY	Ž	Ż
B	°	ą	˛	ł	´	ľ	ś	ˇ	¸	š	ş	ť	ź	˝	ž	ż
C	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
D	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
E	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
F	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	

WINDOWS 1250

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	€		‚		„	…	†	‡		‰	Š	‹	Ś	Ť	Ž	Ź
9		‘	’	“	”	•	–	—		™	š	›	ś	ť	ž	ź
A		ˇ	˘	Ł	¤	Ą	¦	§	¨	©	Ş	«	¬		®	Ż
B	°	±	˛	ł	´	µ	¶	·	¸	ą	ş	»	Ľ	˝	ľ	ż
C	Ŕ	Á	Â	Ă	Ä	Ĺ	Ć	Ç	Č	É	Ę	Ë	Ě	Í	Î	Ď
D	Đ	Ń	Ň	Ó	Ô	Ő	Ö	×	Ř	Ů	Ú	Ű	Ü	Ý	Ţ	ß
E	ŕ	á	â	ă	ä	ĺ	ć	ç	č	é	ę	ë	ě	í	î	ď
F	đ	ń	ň	ó	ô	ő	ö	÷	ř	ů	ú	ű	ü	ý	ţ	

ISO-8859-5

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8																
9																
A	NBSP	Ё	Ђ	Ѓ	Є	S	I	Ї	J	Љ	Њ	Ћ	Ќ	SHY	Ў	Џ
B	A	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
C	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
D	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
E	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
F	№	ё	ђ	ѓ	є	s	i	ї	j	љ	њ	ћ	ќ	§	ў	

VIŠEBAJTNI KARAKTERSKI KODOVI

- Iako navedene kodne strane omogućuju kodiranje tekstova koji nisu na engleskom jeziku nije moguće npr. u istom tekstu mešati ćirilicu i našu latinicu.
- Azijskim jezicima nije dovoljno 256 mesta za zapis svih karaktera.
- Zbog toga se uvode višebajtni karakterski kodovi

UCS, ISO 10646, UNICODE

- Kasnih osamdesetih, dve velike organizacije su pokušale standardizaciju tzv. Univerzalnog skupa karaktera (Universal Character Set - UCS)
- To su bili ISO, kroz standard 10646 i projekat UNICODE organizovan i finansiran uglavnom od strane američkih firmi koje su se bavile proizvodnjom višejezičkog softvera (Apple, Sun Microsystems, Microsoft,...).

ISO 10646

- ISO 10646 je zamišljen kao 4 bajtni standard. Pri tome se prvih 65536 (2^16) karaktera koriste kao osnovni višejezični skup karaktera dok je ostali prostor ostavljen kao proširenje za drevne jezike, celokupnu naučnu notaciju i slično.

UNICODE

- Početna verzija Unicode standarda podrazumevala je dvobajtni kod za svaki karakter (heksadekadno, $(0000)_{16}$-$(FFFF)_{16}$)
- Vremenom se shvatilo da 2^16=65536 karaktera nije dovoljno za zapis svih karaktera koji postoje, pa je odlučeno da se skup kodova proširi na 2^20=1048576 (heksadekadno, $(000000)_{16}$-$(10FFFF)_{16}$)
- Prošireni skup kodova podeljen je na 16 takozvanih ravni pri čemu svaka ravan sadrži 2^16=65536 karaktera

UNICODE

- Prvih 2^16=65536 karaktera kodiranih u opsegu $(0000)_{16}$-$(FFFF)_{16}$ čini takozvanu osnovnu višejezičku ravan, u koju spada većina danas korišćenih karaktera uključujući čak i često korišćene CJK karaktere:

- Vremenom su se ISO 10646 i Unicode združili i danas postoji izuzetno preklapanje između ova dva standarda

UNICODE

- Raspored određenih grupa karaktera u osnovnoj višejezičkoj ravni:

0200-007E – ASCII printable

00A0 – 00FF Latin-1

0100 – 017F – Latin Extended A (osnovno proširenje latinice, sadrži sve naše dijakritike)

0180-077F – Latin Extended B

...

0370-03FF Ćirilica

...

UCS-2

- Unicode standard u suštini predstavlja veliku tabelu koja svakom karakteru dodeljuje broj.
- Standardi koji opisuju kako se niske karaktera onda prevode u nizove bajtova se dodadno definišu
- ISO definiše UCS-2 standard koji jednostavno svaki UNICODE karakter osnovne višejezičke ravni prevodi u odgovarajuća dva bajta

UTF

- Latinični tekstovi kodirani preko UCS-2 standarda sadrže veliki broj nula, koje obično u operativnim sistemima poput UNIX-a i u programskom jeziku C imaju specijalno značenje.
- Iz istog razloga softver koji je razvijen za rad sa dokumentima u ASCII formatu ne može da radi bez izmena nad dokumentima kodiranim preko UCS-2 standarda

UTF

- A *Unicode transformation format* (UTF) algoritam koji svakom UNICODE karakteru dodeljuje određeni niz bajtova čija dužina varira od 1 do najviše 3.
- UTF je ASCII kompatibilan, što znači da se ASCII karakteri zapisuju pomoću jednog bajta, na standardni način.

UTF-8

raspon	binarno zapisan Unicode kôd	binarno zapisan UTF-8 kôd
0000-007F	00000000 0xxxxxxx	0xxxxxxx
0080-07FF	00000yyy yyxxxxxx	110yyyyy 10xxxxxx
0800-FFFF	zzzzyyyy yyxxxxxx	1110zzzz 10yyyyyy 10xxxxxx