Kodiranje slovenskih besedil

Informacija in Kodi – Vaja 2

Asistent: as. dr. Klemen Grm

Kodiranje vira

- Vir informacije: besedilo (≈ 100 znakov)
- Komunikacijski kanal: računalniški pomnilnik oz. mrežna povezava (2 znaka)
- Kodiranje slovenskih besedil

Kodiranje vira

- Uporabni/neuporabni kodi
- Enakomerni/neenakomerni kodi
- Trenutni/netrenutni kodi

Kodiranje besedil

- Uporabni, trenutni kodi
- Enakomerni ali neenakomerni
- Prvi standard: tabela ASCII

ASCII

- 7-bitna kodna tabela – 128 znakov
- Primarni namen: angleška besedila
- Enakomerni kod
- Hex zapis za lažjo predstavo računalniškega pomnilnika

ASCII (1977/1986)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0	NUL 0000	SOH 0001	STX 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
1_ 16	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
2_ 32	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_ 48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_ 64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_ 80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6_ 96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F

Windows-1250 (CP1250)

- Razširitev ASCII za srednjeevropske jezike
- 8-bitni enakomerni kod, 256 znakov
- Poljščina, češčina, slovaščina, madžarščina, slovenščina, hrvaščina, srbščina (latinica), romunščina, albanščina

Windows-1250

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0	NUL 0000	SOH 0001	STX 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
1_ 16	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
2_ 32	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_ 48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_ 64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_ 80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6_ 96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F
8_ 128	€ 20AC		‚ 201A		„ 201E	… 2026	† 2020	‡ 2021		‰ 2030	Š 0160	‹ 2039	Ś 015A	Ť 0164	Ž 017D	Ź 0179
9_ 144		‘ 2018	’ 2019	“ 201C	” 201D	• 2022	– 2013	— 2014		™ 2122	š 0161	› 203A	ś 015B	ť 0165	ž 017E	ź 017A
A_ 160	NBSP 00A0	ˇ 02C7	˘ 02D8	Ł 0141	¤ 00A4	Ą 0104	¦ 00A6	§ 00A7	¨ 00A8	© 00A9	Ş 015E	« 00AB	¬ 00AC	SHY 00AD	® 00AE	Ż 017B
B_ 176	° 00B0	± 00B1	˛ 02DB	ł 0142	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	ą 0105	ş 015F	» 00BB	Ľ 013D	˝ 02DD	ľ 013E	ż 017C
C_ 192	Ŕ 0154	Á 00C1	Â 00C2	Ă 0102	Ä 00C4	Ĺ 0139	Ć 0106	Ç 00C7	Č 010C	É 00C9	Ę 0118	Ë 00CB	Ě 011A	Í 00CD	Î 00CE	Ď 010E
D_ 208	Đ 0110	Ń 0143	Ň 0147	Ó 00D3	Ô 00D4	Ő 0150	Ö 00D6	× 00D7	Ř 0158	Ů 016E	Ú 00DA	Ű 0170	Ü 00DC	Ý 00DD	Ţ 0162	ß 00DF
E_ 224	ŕ 0155	á 00E1	â 00E2	ă 0103	ä 00E4	ĺ 013A	ć 0107	ç 00E7	č 010D	é 00E9	ę 0119	ë 00EB	ě 011B	í 00ED	î 00EE	ď 010F
F_ 240	đ 0111	ń 0144	ň 0148	ó 00F3	ô 00F4	ő 0151	ö 00F6	÷ 00F7	ř 0159	ů 016F	ú 00FA	ű 0171	ü 00FC	ý 00FD	ţ 0163	˙ 02D9

IBM-852 (CP852)

- Še ena razširjena ASCII tabela
- Spodnjih 7 bitov: ASCII
- Isti jeziki kot Windows-1250

ASCII (1977/1986)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0	NUL 0000	SOH 0001	STX 0002	ETX 0003	EOT 0004	ENQ 0005	ACK 0006	BEL 0007	BS 0008	HT 0009	LF 000A	VT 000B	FF 000C	CR 000D	SO 000E	SI 000F
1_ 16	DLE 0010	DC1 0011	DC2 0012	DC3 0013	DC4 0014	NAK 0015	SYN 0016	ETB 0017	CAN 0018	EM 0019	SUB 001A	ESC 001B	FS 001C	GS 001D	RS 001E	US 001F
2_ 32	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_ 48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_ 64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_ 80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6_ 96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	DEL 007F

Code page 852

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_ 128	Ç 00C7	ü 00FC	é 00E9	â 00E2	ä 00E4	ů 016F	ć 0107	ç 00E7	ł 0142	ë 00EB	Ő 0150	ő 0151	î 00EE	Ź 0179	Ä 00C4	Ć 0106
9_ 144	É 00C9	Ĺ 0139	í 013A	ô 00F4	ö 00F6	ľ 013D	ľ 013E	ś 015A	ś 015B	Ö 00D6	Ü 00DC	Ť 0164	ť 0165	Ł 0141	× 00D7	č 010D
A_ 160	á 00E1	í 00ED	ó 00F3	ú 00FA	Ą 0104	ą 0105	Ž 017D	ž 017E	Ę 0118	ę 0119	¬ 00AC	ź 017A	Č 010C	ş 015F	« 00AB	» 00BB
B_ 176	░ 2591	▒ 2592	▓ 2593	| 2502	┤ 2524	Á 00C1	Â 00C2	Ě 011A	Ş 015E	╣ 2563	║ 2551	╗ 2557	╝ 255D	Ż 017B	ż 017C	┐ 2510
C_ 192	└ 2514	┴ 2534	┬ 252C	├ 251C	─ 2500	┼ 253C	Ă 0102	ă 0103	╚ 255A	╔ 2554	╩ 2569	╦ 2566	╠ 2560	═ 2550	╬ 256C	¤ 00A4
D_ 208	đ 0111	Đ 0110	Ď 010E	Ë 00CB	ď 010F	Ň 0147	Í 00CD	Î 00CE	ě 011B	┘ 2518	┌ 250C	█ 2588	▄ 2584	Ţ 0162	Ů 016E	▀ 2580
E_ 224	Ó 00D3	ß 00DF	Ô 00D4	Ń 0143	ń 0144	ň 0148	Š 0160	š 0161	Ŕ 0154	Ú 00DA	ŕ 0155	Ű 0170	ý 00FD	Ý 00DD	ţ 0163	´ 00B4
F_ 240	SHY 00AD	˝ 02DD	˛ 02DB	ˇ 02C7	˘ 02D8	§ 00A7	÷ 00F7	¸ 00B8	° 00B0	¨ 00A8	˙ 02D9	ű 0171	Ř 0158	ř 0159	■ 25A0	NBSP 00A0

ISO-8859-2

- Mednarodni standard za *prikaz* besedil
- Nova 8-bitna tabela
- Nima kontrolnih znakov
- Isti jeziki + lužiščina, turkmenščina

ISO/IEC 8859-2 (Latin-2)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0																
1_ 16																
2_ 32	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_ 48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_ 64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_ 80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6_ 96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	
8_ 128																
9_ 144																
A_ 160	NBSP 00A0	Ą 0104	˘ 02D8	Ł 0141	¤ 00A4	Ľ 013D	Ś 015A	§ 00A7	¨ 00A8	Š 0160	Ş 015E	Ť 0164	Ź 0179	SHY 00AD	Ž 017D	Ż 017B
B_ 176	° 00B0	ą 0105	˛ 02DB	ł 0142	´ 00B4	ľ 013E	ś 015B	ˇ 02C7	¸ 00B8	š 0161	ş 015F	ť 0165	ź 017A	˝ 02DD	ž 017E	ż 017C
C_ 192	Ŕ 0154	Á 00C1	Â 00C2	Ă 0102	Ä 00C4	Ĺ 0139	Ć 0106	Ç 00C7	Č 010C	É 00C9	Ę 0118	Ë 00CB	Ě 011A	Í 00CD	Î 00CE	Ď 010E
D_ 208	Đ 0110	Ń 0143	Ň 0147	Ó 00D3	Ô 00D4	Ő 0150	Ö 00D6	× 00D7	Ř 0158	Ů 016E	Ú 00DA	Ű 0170	Ü 00DC	Ý 00DD	Ţ 0162	ß 00DF
E_ 224	ŕ 0155	á 00E1	â 00E2	ă 0103	ä 00E4	ĺ 013A	ć 0107	ç 00E7	č 010D	é 00E9	ę 0119	ë 00EB	ě 011B	í 00ED	î 00EE	ď 010F
F_ 240	đ 0111	ń 0144	ň 0148	ó 00F3	ô 00F4	ő 0151	ö 00F6	÷ 00F7	ř 0159	ů 016F	ú 00FA	ű 0171	ü 00FC	ý 00FD	ţ 0163	˙ 02D9

MacCE
(Mac OS Central European)

- Svojevrstna razširitev ASCII tabele
- -Albanščina, hrvaščina, romunščina
- +Estonščina, litovščina, latvijščina
- Razširjena ločila, matematični simboli

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_ 128	Ä 00C4	Ā 0100	ā 0101	É 00C9	Ą 0104	Ö 00D6	Ü 00DC	á 00E1	ą 0105	Č 010C	ä 00E4	č 010D	Ć 0106	ć 0107	é 00E9	Ź 0179
9_ 144	ź 017A	Ď 010E	í 00ED	ď 010F	Ē 0112	ē 0113	Ė 0116	ó 00F3	ė 0117	ô 00F4	ö 00F6	õ 00F5	ú 00FA	Ě 011A	ě 011B	ü 00FC
A_ 160	† 2020	° 00B0	Ę 0118	£ 00A3	§ 00A7	• 2022	¶ 00B6	ß 00DF	® 00AE	© 00A9	™ 2122	ę 0119	¨ 00A8	≠ 2260	ģ 0123	Į 012E
B_ 176	į 012F	Ī 012A	≤ 2264	≥ 2265	ī 012B	Ķ 0136	∂ 2202	∑ 2211	ł 0142	Ļ 013B	ļ 013C	Ľ 013D	ľ 013E	Ĺ 0139	Í 013A	Ņ 0145
C_ 192	ņ 0146	Ń 0143	¬ 00AC	√ 221A	ń 0144	Ň 0147	Δ 2206	« 00AB	» 00BB	… 2026	NBSP 00A0	ň 0148	Ő 0150	Õ 00D5	ő 0151	Ō 014C
D_ 208	– 2013	— 2014	“ 201C	” 201D	‘ 2018	’ 2019	÷ 00F7	◊ 25CA	ō 014D	Ŕ 0154	ŕ 0155	Ř 0158	‹ 2039	› 203A	ř 0159	Ŗ 0156
E_ 224	ŗ 0157	Š 0160	‚ 201A	„ 201E	š 0161	Ś 015A	ś 015B	Á 00C1	Ť 0164	ť 0165	Í 00CD	Ž 017D	ž 017E	Ū 016A	Ó 00D3	Ô 00D4
F_ 240	ū 016B	Ů 016E	Ú 00DA	ů 016F	Ű 0170	ű 0171	Ų 0172	ų 0173	Ý 00DD	ý 00FD	ķ 0137	Ź 017B	Ł 0141	ż 017C	Ģ 0122	ˇ 02C7

Unicode

Situacija 1995:

- Ločene kodne tabele za različne jezike (skupine jezikov)
- Implementacije, specifične operacijskim sistemom in proizvajalcem
- Pomankljiva standardizacija

Unicode

Ideja:

- En kod za vsa besedila
- Enoten standard
- Neodvisnost od strojne, programske opreme

Unicode

- 154 abeced, do 10000 znakov / abecedo
- Potrebe:
  - Veliko število kodnih mest ($N > 10^6$)
  - Neenakomerno kodiranje za varčevanje s pomnilnikom
  - Trenutni kod
- Ena Unicode tabela, $1.112.064$ znakov
- Več različnih transformacijskih formatov (UTF)

UTF-16

- Neenakomeren kod, 1 ali 2 16-bitna znaka
- LE in BE izvedbe
- Uporaba: (interno) Windows, Java
- Odsvetovana uporaba na spletu
- Izrazito negospodarno kodiranje – vsaj 16 bitov/znak

UTF-8

- De-facto standard za spletno kodiranje besedila
- Uporaba:
  - Unix-like sistemi (Linux, Mac OS, Android,…)
  - Splet
  - Python
- Neenakomeren kod: 1, 2, 3 ali 4 8-bitni znaki
- ASCII znaki: identičen zapis
- Struktura predpon

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF[12]	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Laboratorijska vaja 2

- Naloga 1 – izpis kodnih zamenjav šumnikov slovenske abecede
  - Č,Š,Ž,č,š,ž
- IBM-852,
- ISO-8859-2
- Windows-1250,
- MacCE
- UTF-16LE, UTF-16BE
- UTF-8
- Decimalni, hex, binarni zapis

Laboratorijska vaja 2

- Naloga 2 – kodiranje in dekodiranje UTF-8
- Vhod: zaporedje kodnih mest
- Izhod: v utf-8 kodirana besedilna datoteka

1. Beri vhodno datoteko, pretvori kodna mesta v številske podatke
2. Pretvorba kodnih mest v Unicode znake
3. Zapis izhodne datoteke
4. Zapis unikatnih znakov in kodnih zamenjav