Kapitola: Souborové formáty

Cvičení 6 – Rozpoznávání souborových formátů

Témata

Textové a binární souborové formáty a jejich rozpoznávání.
Obyčejný text, rozšířený text, konec řádku a jeho reprezentace v různých operačních systémech.
Jednobajtová a vícebajtová kódování národních znaků, jejich rozpoznání a změna.
Frekventované konkrétní souborové formáty.

Materiál

Programy: PSPad, 7zip, příkazový řádek v OS Unix a Windows, `file`, `enca`, `cstocs`
Soubory: pracovní archiv s 20 soubory různých formátů a různých kódování národních znaků
Ostatní: kódové tabulky jednobajtových kódování národních znaků

Co máme znát?

- Pojem *znakový kód* je předpis přiřazující danému znaku číselnou reprezentaci v počítači.
- Pojmem *textový soubor* označíme takový soubor, jehož všechny informace jsou vyjádřeny zobrazitelnými znaky. Může obsahovat konce řádků odpovídající operačnímu systému, kde soubor vznikl.
- Pojem *čistý text* (obyčejný, plain text) je textový soubor obsahující pouze zobrazitelné znaky ASCII.
- Pojem *rozšířený text* (extended text) je textový soubor obsahující kromě znaků ASCII ještě zobrazitelné znaky národních abeced, příp. typografické znaky.
- Operační systémy firmy Microsoft *asociují* aplikační program k datovému souboru. Orientují se pouze podle rozšíření jména souboru.

Úkoly

1. Archiv s pracovními soubory *rozbalte do předem připraveného adresáře*, jehož obsah lze sledovat jak z unixového, tak z windowsového prostředí (sdílený disk).
2. V průběhu cvičení budeme vyplňovat tabulku, která se nachází na další straně tohoto dokumentu.
3. U každého souboru zjistěte, zda je *textový*, nebo *netextový*. Použijte hexadecimální režim editoru PSPad (Soubor → Otevřít hexadecimálně). Výsledek zaneste do tabulky.
4. Je-li soubor textový:
   a) Zjistěte, zda jde o *čistý text*, nebo o *rozšířený text*, tedy zda obsahuje národní znaky. Pokud se jedná o rozšířený text, zjistěte dále v jakém *kódování* jsou národní znaky uloženy. Vyzkoušejte metodu ručního rozpoznání s využitím kódových tabulek, autodetekci v programu PSPad a příkaz `enca` v prostředí OS Unix. Výsledky zaneste do tabulky.
   b) Pomocí programu `cstocs` (přepínač `-i`, parametry vstupní a výstupní kódování – `il2`, `pc2`, `1250`, `utf8`, `kam` nebo `koi8` – a jméno souboru) nebo editoru PSPad proveďte změnu kódování vybraného rozšířeného textového souboru a ověřte, že změna proběhla korektně.
   c) Zjistěte, v jakém *operačním systému* byl soubor (pravděpodobně) vytvořen. Výsledek zaneste do tabulky.
5. Zjistěte *souborový formát* (správnou příponu, která souboru náleží). Použijte příkaz `file` v prostředí OS Unix. Výsledek zaneste do tabulky a soubor přejmenujte tak, aby měl správnou příponu.
6. Vyhledejte a doplňte do tabulky *asociovanou aplikaci*, kterou je možné soubor zpracovat.
7. Prohlédněte si soubor v příslušné aplikaci a doplňte do tabulky, *co se v souboru nachází*.

Jméno souboru	Je soubor textový?	Obsahuje nár. znaky?	V jakém je kódování?	Odkud asi pochází?	V jakém je formátu?	Čím jej zpracujeme?	Co se nachází uvnitř?
soubor.01							
soubor.02							
soubor.03							
soubor.04							
soubor.05							
soubor.06							
soubor.07							
soubor.08							
soubor.09							
soubor.10							
soubor.11							
soubor.12							
soubor.13							
soubor.14							
soubor.15							
soubor.16							
soubor.17							
soubor.18							
soubor.19							
soubor.20							

Kontrolní otázky

Následující otázky vyčerpávají problematiku tohoto cvičení a obsahují i poněkud širší kontext. Správné odpovědi na ně získáte z tohoto cvičení, z přednášek a z učebnice.

- Co je to souborový formát dat? Jak lze pracovně definovat textový souborový formát?
- Jak lze poznat podle konce řádku operační systém, v němž textový formát pravděpodobně vznikl?
- Jakým způsobem lze zobrazit binární obsah souboru v prostředí systému Unix a Windows?
- Jakým způsobem lze zobrazit textový obsah souboru v prostředí systému Unix a Windows?
- Co je to znakový kód? Jaká jednobajtová kódování národních znaků znáte?
- Kolik znaků obsahuje základní kód ASCII?
- K čemu slouží řídicí a zobrazitelné znaky? Na kterých pozicích se v tabulce nacházejí?
- Co je to plain text (obyčejný text), extended text (rozšířený text)?
- Co je to dokument? Uveďte příklady souborových formátů, které chápeme jako dokumenty.
- Jaký je princip asociace formátu s aplikací v systému Windows?
- Jaký je možný postup zjištění souborového formátu v systému Unix a Windows?
- Lze pomocí uvedených prostředků rozpoznat každý binární formát? Pro své tvrzení uveďte příklady.
- Na jakém principu jsou založeny dokumentní formáty kancelářského balíku MS Office verze 2007 a novějších? Jak jsou reprezentována veškerá data v těchto formátech?
- Porovnejte možnosti, výhody a nevýhody různých způsobů zjišťování kódování národních znaků.
- Která kódování národních znaků jsou považována za perspektivní a proč?
- Která kódování národních znaků jsou ještě používána (ale jejich význam klesá) a proč?
- Popište výstižně souborové formáty BMP, JPG, PNG, GIF, TIFF, CDR, SVG; DOC, DOCX, RTF, XLS, XLSX, PPT, PPTX; HTML, MHT, XML; PS/EPS, PDF, CSV, TXT, PAS; MP3, AVI; ZIP.

Tabulka ASCII (ISO 646)

Hexa	0x	1x	2x	3x	4x	5x	6x	7x
x0	⟨NUL⟩ 0	⟨DLE⟩ 16	mezera 32	0 48	@ 64	P 80	‘ 96	p 112
x1	⟨SOH⟩ 1	⟨DC1⟩ 17	! 33	1 49	A 65	Q 81	a 97	q 113
x2	⟨STX⟩ 2	⟨DC2⟩ 18	” 34	2 50	B 66	R 82	b 98	r 114
x3	⟨ETX⟩ 3	⟨DC3⟩ 19	# 35	3 51	C 67	S 83	c 99	s 115
x4	⟨EOT⟩ 4	⟨DC4⟩ 20	$ 36	4 52	D 68	T 84	d 100	t 116
x5	⟨ENQ⟩ 5	⟨NAK⟩ 21	% 37	5 53	E 69	U 85	e 101	u 117
x6	⟨ACK⟩ 6	⟨SYN⟩ 22	& 38	6 54	F 70	V 86	f 102	v 118
x7	⟨BEL⟩ 7	⟨ETB⟩ 23	’ 39	7 55	G 71	W 87	g 103	w 119
x8	⟨BS⟩ 8	⟨CAN⟩ 24	(40	8 56	H 72	X 88	h 104	x 120
x9	⟨HT⟩ 9	⟨EM⟩ 25	) 41	9 57	I 73	Y 89	i 105	y 121
xA	⟨LF⟩ 10	⟨SUB⟩ 26	* 42	: 58	J 74	Z 90	j 106	z 122
xB	⟨VT⟩ 11	⟨ESC⟩ 27	+ 43	; 59	K 75	[91	k 107	{ 123
xC	⟨FF⟩ 12	⟨FS⟩ 28	, 44	< 60	L 76	\ 92	l 108	| 124
xD	⟨CR⟩ 13	⟨GS⟩ 29	- 45	= 61	M 77	] 93	m 109	} 125
xE	⟨SO⟩ 14	⟨RS⟩ 30	. 46	> 62	N 78	^ 94	n 110	~ 126
xF	⟨SI⟩ 15	⟨US⟩ 31	/ 47	? 63	O 79	_ 95	o 111	⟨DEL⟩ 127

Tabulka CP 895 (Kód bratří Kamenických)

Hexa	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	Č 128	É 144	á 160	▒ 176	└ 192	╨ 208	α 224	≡ 240
x1	ü 129	ž 145	í 161	▓ 177	┴ 193	╤ 209	β 225	± 241
x2	é 130	Ž 146	ó 162	▓ 178	┬ 194	╥ 210	Γ 226	≥ 242
x3	ď 131	ô 147	ú 163	│ 179	├ 195	╙ 211	π 227	≤ 243
x4	ä 132	ö 148	ň 164	┤ 180	─ 196	╘ 212	Σ 228	⌠ 244
x5	Ď 133	Ó 149	Ň 165	╡ 181	┼ 197	╒ 213	σ 229	⌡ 245
x6	Ť 134	ů 150	Ů 166	╢ 182	╞ 198	╓ 214	μ 230	÷ 246
x7	č 135	Ú 151	Ô 167	╖ 183	╟ 199	╫ 215	τ 231	∼ 247
x8	ě 136	ý 152	š 168	╕ 184	╚ 200	╪ 216	Φ 232	° 248
x9	Ě 137	Ö 153	ř 169	╣ 185	╔ 201	┘ 217	Θ 233	· 249
xA	Ĺ 138	Ů 154	ŕ 170	║ 186	╩ 202	┌ 218	Ω 234	— 250
xB	Í 139	Š 155	Ŕ 171	╗ 187	╦ 203	▉ 219	δ 235	√ 251
xC	ľ 140	Ľ 156	¼ 172	╝ 188	╠ 204	▄ 220	∞ 236	^{n} 252
xD	Í 141	Ý 157	§ 173	╜ 189	═ 205	▌ 221	∅ 237	^{2} 253
xE	Ä 142	Ř 158	≪ 174	╛ 190	╬ 206	▐ 222	∈ 238	■ 254
xF	Á 143	ť 159	≫ 175	┐ 191	╧ 207	▀ 223	∩ 239	255

Tabulka CP 852 (PC Latin 2)

Hexa	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	Ç 128	É 144	á 160	░ 176	└ 192	ð 208	Ó 224	240
x1	ü 129	Ĺ 145	í 161	▒ 177	┴ 193	Đ 209	ß 225	˝ 241
x2	é 130	Í 146	ó 162	▓ 178	┬ 194	Ď 210	Ô 226	˛ 242
x3	â 131	ô 147	ú 163	│ 179	├ 195	Ë 211	Ń 227	ˇ 243
x4	ä 132	ö 148	Ą 164	┤ 180	─ 196	ď 212	ń 228	˘ 244
x5	ů 133	Ľ 149	ą 165	Á 181	┼ 197	Ň 213	ň 229	§ 245
x6	ć 134	ľ 150	Ž 166	Â 182	Ă 198	Í 214	Š 230	÷ 246
x7	ç 135	Ś 151	ž 167	Ě 183	ă 199	Î 215	š 231	¸ 247
x8	ł 136	ś 152	Ę 168	Ş 184	╚ 200	ě 216	Ř 232	° 248
x9	ë 137	Ö 153	ę 169	╣ 185	╔ 201	┘ 217	Ú 233	¨ 249
xA	Ő 138	Ü 154	170	║ 186	╩ 202	┌ 218	ŕ 234	· 250
xB	ő 139	Ť 155	ź 171	╗ 187	╦ 203	▮ 219	Ű 235	ű 251
xC	î 140	ť 156	Č 172	╝ 188	╠ 204	▬ 220	ý 236	Ř 252
xD	Ź 141	Ł 157	ş 173	Ż 189	═ 205	Ţ 221	Ý 237	ř 253
xE	Ä 142	× 158	≪ 174	ż 190	╬ 206	Ů 222	ţ 238	■ 254
xF	Ć 143	č 159	≫ 175	┐ 191	¤ 207	▬ 223	´ 239	255

Tabulka ISO 8859-2 (ISO Latin 2)

Hexa	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	nevyužito 128	nevyužito 144	⟨*NBSP*⟩ 160	° 176	Ŕ 192	Đ 208	ŕ 224	đ 240
x1	nevyužito 129	nevyužito 145	Ą 161	ą 177	Á 193	Ń 209	á 225	ń 241
x2	nevyužito 130	nevyužito 146	˘ 162	˛ 178	Â 194	Ň 210	â 226	ň 242
x3	nevyužito 131	nevyužito 147	Ł 163	ł 179	Ă 195	Ó 211	ă 227	ó 243
x4	nevyužito 132	nevyužito 148	¤ 164	´ 180	Ä 196	Ô 212	ä 228	ô 244
x5	nevyužito 133	nevyužito 149	Ľ 165	ľ 181	Ĺ 197	Ő 213	Í 229	ő 245
x6	nevyužito 134	nevyužito 150	Ś 166	ś 182	Ć 198	Ö 214	ć 230	ö 246
x7	nevyužito 135	nevyužito 151	§ 167	ˇ 183	Ç 199	× 215	ç 231	÷ 247
x8	nevyužito 136	nevyužito 152	¨ 168	¸ 184	Č 200	Ř 216	č 232	ř 248
x9	nevyužito 137	nevyužito 153	Š 169	š 185	É 201	Ů 217	é 233	ů 249
xA	nevyužito 138	nevyužito 154	Ş 170	ş 186	Ę 202	Ú 218	ę 234	ú 250
xB	nevyužito 139	nevyužito 155	Ť 171	ť 187	Ë 203	Ű 219	ë 235	ű 251
xC	nevyužito 140	nevyužito 156	Ź 172	ź 188	Ě 204	Ü 220	ě 236	ü 252
xD	nevyužito 141	nevyužito 157	⟨*SHY*⟩ 173	″ 189	Í 205	Ý 221	í 237	ý 253
xE	nevyužito 142	nevyužito 158	Ž 174	ž 190	Î 206	Ţ 222	î 238	ţ 254
xF	nevyužito 143	nevyužito 159	Ż 175	ż 191	Ď 207	ß 223	ď 239	˙ 255

Tabulka KOI8čs

Hexa	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	128	144	160	176	á 192	ô 208	À 224	Ô 240
x1	129	145	161	177	á 193	ä 209	Á 225	Ä 241
x2	130	146	´ 162	178	ă 194	ř 210	Ă 226	Ř 242
x3	131	147	163	179	č 195	š 211	Č 227	Š 243
x4	132	148	¿ 164	180	ď 196	ť 212	Ď 228	Ť 244
x5	133	149	165	181	ě 197	ú 213	Ě 229	Ú 245
x6	134	150	˘ 166	182	ŕ 198	ë 214	Ŕ 230	Ë 246
x7	135	151	· 167	183	ch 199	é 215	CH 231	É 247
x8	136	152	¨ 168	184	ü 200	ű 216	Ü 232	Ű 248
x9	137	153	169	185	í 201	ý 217	Í 233	Ý 249
xA	138	154	° 170	186	ů 202	ž 218	Ů 234	Ž 250
xB	139	155	¸ 171	187	Í 203	219	Ĺ 235	251
xC	140	156	172	188	ľ 204	220	Ľ 236	252
xD	141	157	˝ 173	189	ö 205	ő 221	Ö 237	Ő 253
xE	142	158	˛ 174	190	ň 206	ė 222	Ň 238	Ė 254
xF	143	159	ˇ 175	191	ó 207	ß 223	Ó 239	255

Tabulka CP 1250 (Windows)

Hexa	8x	9x	Ax	Bx	Cx	Dx	Ex	Fx
x0	€ 128	nevyužito 144	⟨*NBSP*⟩ 160	° 176	Ŕ 192	Đ 208	ŕ 224	đ 240
x1	nevyužito 129	‘ 145	ˇ 161	± 177	Á 193	Ń 209	á 225	ń 241
x2	‚ 130	’ 146	˘ 162	˛ 178	Â 194	Ň 210	â 226	ň 242
x3	nevyužito 131	“ 147	Ł 163	ł 179	Ă 195	Ó 211	ă 227	ó 243
x4	„ 132	” 148	¤ 164	´ 180	Ä 196	Ô 212	ä 228	ô 244
x5	… 133	• 149	Ą 165	µ 181	Ĺ 197	Ő 213	Í 229	ő 245
x6	† 134	– 150	| 166	¶ 182	Ć 198	Ö 214	ć 230	ö 246
x7	‡ 135	— 151	§ 167	· 183	Ç 199	× 215	ç 231	÷ 247
x8	nevyužito 136	nevyužito 152	¨ 168	¸ 184	Č 200	Ř 216	č 232	ř 248
x9	‰ 137	TM 153	© 169	ą 185	É 201	Ů 217	é 233	ů 249
xA	Š 138	š 154	Ş 170	ş 186	Ę 202	Ú 218	ę 234	ú 250
xB	⟨ 139	⟩ 155	» 171	« 187	Ë 203	Ű 219	ë 235	ű 251
xC	Ś 140	ś 156	¬ 172	Ľ 188	Ě 204	Ü 220	ě 236	ü 252
xD	Ť 141	ť 157	⟨*SHY*⟩ 173	″ 189	Í 205	Ý 221	í 237	ý 253
xE	Ž 142	ž 158	® 174	ľ 190	Î 206	Ţ 222	î 238	ţ 254
xF	Ź 143	ź 159	Ż 175	ż 191	Ď 207	ß 223	ď 239	˙ 255