



VILNIAUS GEDIMINO
TECHNIKOS UNIVERSITETAS

Unikodas/UTF-8

2 laboratorinis darbas

Dėstytojai: doc. dr. Pavel Stefanovič ir lekt. Rokas Štrimaitis

Įvadas į simbolių kodavimas

- Kiekvienam programuotojui dirbančiam su tekstu turi kilti klausimas kaip tas tekstas yra užkoduotas.
- Svarbu: internetiniai puslapiai, duomenų bazės, tinklo programavimas (angl. *network programming*), bet kokiam duomenų perdavime.
- Pasekmės: informacijos praradimas (netinkamas teksto apdorojimas), paieškos sistemų blogas indeksavimas ar net saugumo spragos.

Daugiau informacijos:

- <http://kunststube.net/encoding/>
- <https://www.baeldung.com/java-char-encoding>

Problemos pavyzdys (1)

```
<html>
<head>
</head>
<body>
A Ć E Ė Į Š Ų
</body>
</html>
```



Ä,, Ä Ä Ä– Ä® Å Å²

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
</head>
<body>
A Ć E Ė Į Š Ų
</body>
</html>
```



A Ć E Ė Į Š Ų

Problemos pavyzdys (2)

New

51703_8511922-1

ŽUVYS FILE

*Aukščiausios kokybės puikuotinė medvilnė,
išskirtinio dizaino ir ypatingai plona vyrų kojinių.*

*Kojinių mezgimui naudojamas medvilnės pluoštas
puikiai sugeria drėgmę, leidžia odai kvėpuoti.*



> SUDĖTIS

> DYDŽIAI

> PRIEŽIŪRA

Simbolių kodavimas

- Kompiuteriai supranta informaciją ir saugo ją atmintyje kaip dvejetainius skaičius (arba elektrinį impulsą – jis yra arba jo nėra).
- Norint keistis informacija tarp kompiuterių, būtina simbolių kodavimus standartizuoti ir suvienodinti.
- Vienas iš pirmųjų standartų buvo ASCII. Naudojami 7 iš 8 bitų (0-127).

ASCII lentelė

ASCII (1977/1986)

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_	NUL 0000 0	SOH 0001 1	STX 0002 2	ETX 0003 3	EOT 0004 4	ENQ 0005 5	ACK 0006 6	BEL 0007 7	BS 0008 8	HT 0009 9	LF 000A 10	VT 000B 11	FF 000C 12	CR 000D 13	SO 000E 14	SI 000F 15
1_	DLE 0010 16	DC1 0011 17	DC2 0012 18	DC3 0013 19	DC4 0014 20	NAK 0015 21	SYN 0016 22	ETB 0017 23	CAN 0018 24	EM 0019 25	SUB 001A 26	ESC 001B 27	FS 001C 28	GS 001D 29	RS 001E 30	US 001F 31
2_	SP 0020 32	! 0021 33	" 0022 34	# 0023 35	\$ 0024 36	% 0025 37	& 0026 38	' 0027 39	(0028 40) 0029 41	* 002A 42	+ 002B 43	, 002C 44	- 002D 45	. 002E 46	/ 002F 47
3_	0 0030 48	1 0031 49	2 0032 50	3 0033 51	4 0034 52	5 0035 53	6 0036 54	7 0037 55	8 0038 56	9 0039 57	: 003A 58	; 003B 59	< 003C 60	= 003D 61	> 003E 62	? 003F 63
4_	@ 0040 64	A 0041 65	B 0042 66	C 0043 67	D 0044 68	E 0045 69	F 0046 70	G 0047 71	H 0048 72	I 0049 73	J 004A 74	K 004B 75	L 004C 76	M 004D 77	N 004E 78	O 004F 79
5_	P 0050 80	Q 0051 81	R 0052 82	S 0053 83	T 0054 84	U 0055 85	V 0056 86	W 0057 87	X 0058 88	Y 0059 89	Z 005A 90	[005B 91	\ 005C 92] 005D 93	^ 005E 94	_ 005F 95
6_	` 0060 96	a 0061 97	b 0062 98	c 0063 99	d 0064 100	e 0065 101	f 0066 102	g 0067 103	h 0068 104	i 0069 105	j 006A 106	k 006B 107	l 006C 108	m 006D 109	n 006E 110	o 006F 111
7_	p 0070 112	q 0071 113	r 0072 114	s 0073 115	t 0074 116	u 0075 117	v 0076 118	w 0077 119	x 0078 120	y 0079 121	z 007A 122	{ 007B 123	 007C 124	} 007D 125	~ 007E 126	DEL 007F 127

☐ Letter
 ☐ Number
 ☐ Punctuation
 ☐ Symbol
 ☐ Other
 ☐ undefined
 ☐ Changed from 1963 version


Išplėstinės kodų lentelės

- Vis labiau populiarėjant kompiuteriams atsirado poreikis turėti daugiau standartinių kodų. Taip buvo panaudotas 8-tas bitas ir gauti nauji 128 kodai.
- Atsirado gausybė papildomų pritaikytų įvairioms šalims kodų lentelių (daugiau nei 220 DOS ir Windows išplėstinių lentelių).
- **PROBLEMA:** skirtingose koduotėse tas pats kodas gali būti priskirtas skirtingiems simboliams.
- Sparčiai gausėjančią kodų įvairovę iškilo būtinybė susisteminti.

ASCII išplėstinės lentelės

Code page 775


	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_	Ć	ü	é	ā	ă	ġ	ă	ć	ł	ē	Ŕ	ŕ	ī	ž	Ä	Å
128	0106	00FC	00E9	0101	00E4	0123	00E5	0107	0142	0113	0156	0157	012B	0179	00C4	00C5
9_	É	æ	Æ	ō	ö	Ġ	ċ	ś	ś	Ö	Ü	ø	£	∅	×	×
144	00C9	00E6	00C6	014D	00F6	0122	00A2	015A	015B	00D6	00DC	00F8	00A3	00D8	00D7	00A4
A_	Ā	Ī	Ó	Ž	ž	ž	“	ı	©	®	¬	½	¾	Ł	«	»
160	0100	012A	00F3	017B	017C	017A	201D	00A6	00A9	00AE	00AC	00BD	00BC	0141	00AB	00BB
B_	⌠	⌡	⌢	⌣	⌤	Ą	Č	Ę	Ê	⌥	⌦	⌧	⌨	〈	Š	Ų
176	2591	2592	2593	2502	2524	0104	010C	0118	0116	2563	2551	2557	255D	012E	0160	2510
C_	Ł	ł	Ť	ť	—	†	Ů	Ů	ℓ	ŕ	ℓ	ŕ	ℓ	ŕ	ŕ	ž
192	2514	2534	252C	251C	2500	253C	0172	016A	255A	2554	2569	2566	2560	2550	256C	017D
D_	ą	č	ę	ê	ı	š	ų	ū	ž	Ų	Ų	Ų	Ų	Ų	Ų	Ų
208	0105	010D	0119	0117	0123	0161	0173	016B	017E	2518	250C	2588	2584	258C	2590	2580
E_	Ó	ß	Ö	Ň	ö	Ö	µ	ñ	ķ	ķ	Ļ	Ļ	ņ	Ē	Ņ	'
224	00D3	00DF	014C	0143	00F5	00D5	00B5	0144	0136	0137	013B	013C	0146	0112	0145	2019
F_	SHY	±	“	¾	Ų	Ų	÷	”	°	°	°	°	°	°	■	NBSP
240	00AD	00B1	201C	00BE	00B6	00A7	00F7	201E	00B0	2219	00B7	00B9	00B3	00B2	25A0	00A0



 00F6

Code page 708

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
8_			é	â	†	à	†	ç	ê	ë	è	ï	î	ŋ	ŋ	†
128	2502	2524	00E9	00E2	2561	00E0	2562	00E7	00EA	00EB	00E8	00EF	00EE	2556	2555	2563
9_		ŋ	ŋ	ô	†	†	†	†	†	†	†	†	†	†	†	†
144	2551	2557	255D	00F4	255C	255B	00FB	00F9	2510	2514						
A_		†	†	†	†	†	†	†	†	†	†	†	†	†	†	†
160		2534	252C	251C	00A4	2500	253C	255E	255F	255A	2554	2569	060C	2566	00AB	00BB
B_	⌠	⌡	⌢	⌣	⌤	⌥	⌦	⌧	⌨	〈	〉	⌫	⌬	⌭	⌮	⌯
176	2591	2592	2593	2560	2550	256C	2567	2568	2564	2565	2559	061B	2558	2552	2553	061F
C_		ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
192		0621	0622	0623	0624	0625	0626	0627	0628	0629	062A	062B	062C	062D	062E	062F
D_	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م
208	0630	0631	0632	0633	0634	0635	0636	0637	0638	0639	063A	2588	2584	258C	2590	2580
E_	ن	ه	و	ي	ج	ي	ي	ي	ي	ي	ي	ي	ي	ي	ي	ي
224	0640	0641	0642	0643	0644	0645	0646	0647	0648	0649	064A	064B	064C	064D	064E	064F
F_	و	و	و									†	†	†	†	NBSP
240	0650	0651	0652									2561	2518	250C	00B5	00A0



 255C

SKIRTINGI SIMBOLIAI
 VIENODI NUMERIAI

Unikodas

- **Unikodas** (angl. *Unicode*) – standartas, apibrėžiantis beveik visų kalbų abėcėlių bei papildomų simbolių kodavimą kompiuteriuose.
- Pirmosios 256 pozicijos yra identiškos ISO 8859-1 kodavimo simboliams, kad būtų paprastesnis keitimas iš egzistuojančių Vakarų Europos kalbų tekstų.
- Unikodas numato tik pozicijas įvairiems simboliams, bet ne realų kodavimą fiziniame atmintyje.

Range: 0020– 007F

Number of characters: 96

type: alphabet

Languages: english, german, french,
italian, polish



<http://unicode-table.com/en/>

UTF-8

- UTF-8 yra Unikodo tipo koduotė, kurioje vienam simboliui skiriama nuo 1 baido iki 4 baidų.
 - **1 baidu** koduojami ASCII simboliai (angliškos abėcėlės raidės, skaičiai, skyrybos ženklai).
 - **2 baidais** – išplėstinės lotynų abėcėlės (tarp jų ir lietuvių), graikų, hebrajų ir t.t. raidės.
 - **3 baidais** – japonų, kinų ir kitų Azijos tautų raidės.
 - **4 baidais** – itin reti simboliai.

UTF-8

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	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

UNIKODAS → UTF-8

- Tarkime turime tekstinį failą kuriame pateiktas vienas simbolis Ä.
Unikodas: norint gauti unikodo reikšmę, tereikia pasiversti į šešioliktainę skaičiaus išraišką.
- Šiuo atveju jo **Unikodas** – **U+00C4**.



Latin Capital Letter a with
Diaeresis

Unicode number: U+00C4

HTML-code: Ä

UNIKODAS → UTF-8

- UTF-8:** kiek reikia baitų užkoduoti simbolį matome iš šešioliktainės skaičiaus išraiškos. Šiuo atveju C4 priklauso antram intervalui, todėl prireiks dviejų baitų. $0080 < 00C4 < 07FF$

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	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Ä

Latin Capital Letter a with
Diaeresis

Unicode number: U+00C4

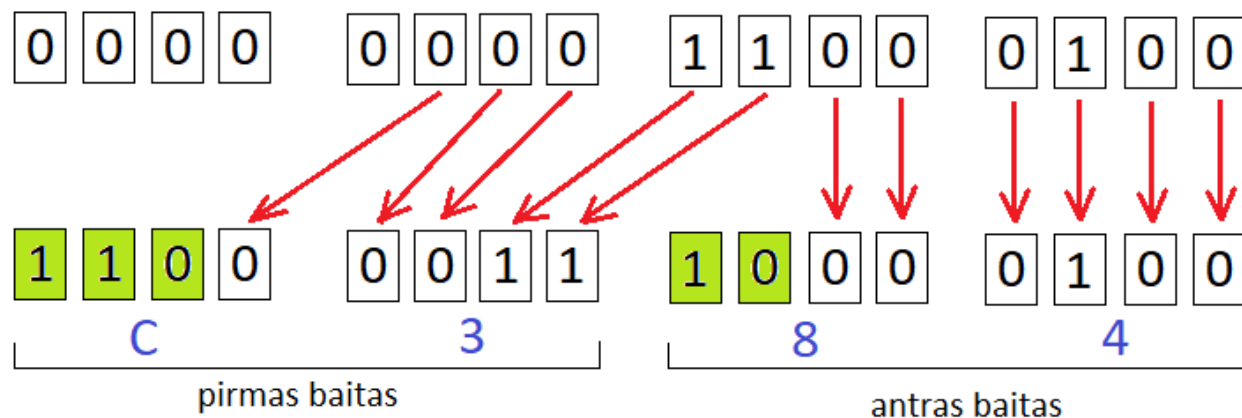
HTML-code: Ä

UNIKODAS → UTF-8

- Dvejetainė išraiška: 0000 0000 1100 0100
- Konvertuojame į UTF-8 pagal lentelėje pateiktą formą surašant skaitmenis iš dešinės į kairę.

2	11	U+0080	U+07FF	110xxxxx	10xxxxxx	
---	----	--------	--------	----------	----------	--

Originali reikšmė



Ä

Latin Capital Letter a with
Diaeresis

Unicode number: U+00C4

HTML-code: Ä

UTF-8
C3 84

UNIKODAS → UTF-8

Užduotis Nr. 1.: ar gaunate tokį pat unikodą ir utf-8?

ठ – $1001\ 0010\ 0000_2 = 2336_{10}$

Unikodas – U+0920

UTF-8 – E0 A4 A0

ठ

Devanagari Letter Ttha

Unicode number: U+0920

HTML-code: ठ

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	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

UNIKODAS → UTF-8

Užduotis Nr. 2.: ar gaunate tokį pat unikodą ir utf-8?

Њ – $0100\ 0111\ 1100_2 = 1148_{16}$

Unikodas – U+047C

UTF-8 – D1 BC



Cyrillic Capital Letter Omega
with Titlo

Unicode number: U+047C

HTML-code: Ѽ

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	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

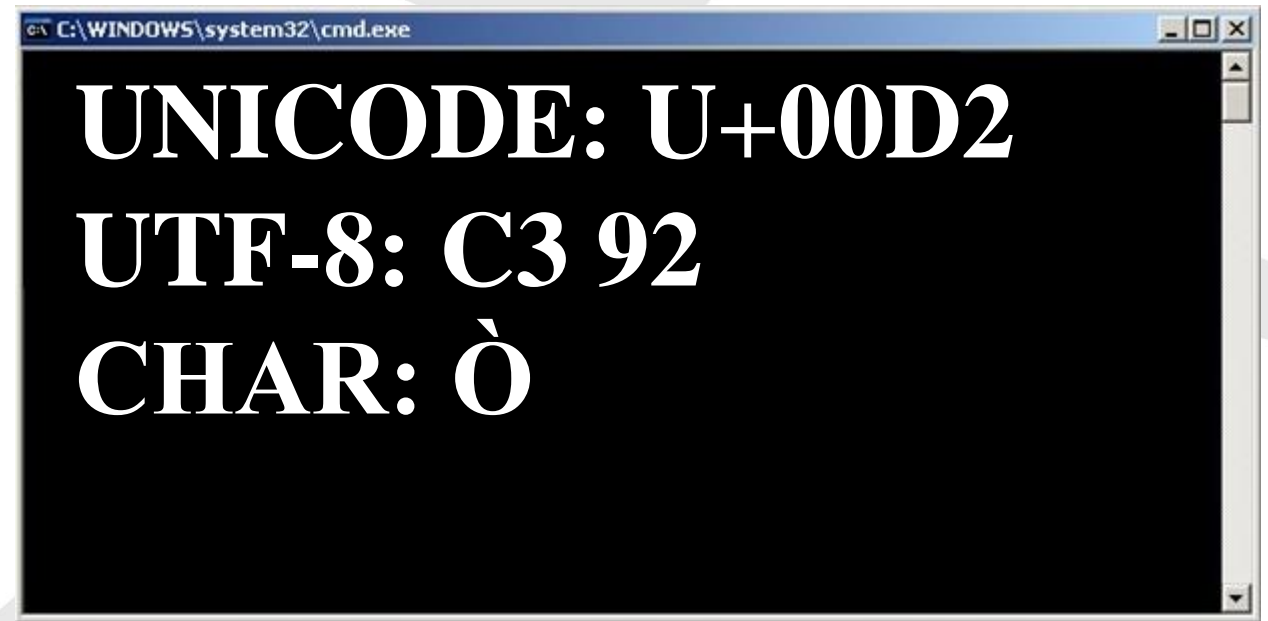
Laboratorinio darbo tikslas

- Išmokti konvertuoti bet kokį simbolį į dažniausiai šiuo metu naudojamą UTF-8 koduotę. Suprasti kas yra Unikodas, ką jis apibrėžia ir kokia jo sąsaja su UTF-8.
- Programavimo įgūdžių tobulinimas.

2 laboratorinis darbas (I dalis)

(iki 2019.10.25)

- Sukurkite programą, kurioje įrašomas dešimtainis skaičius, o rezultate yra išvedamas jo Unikodo numeris, UTF-8 kodas ir simbolis (veiks iki tam tikros reikšmės).

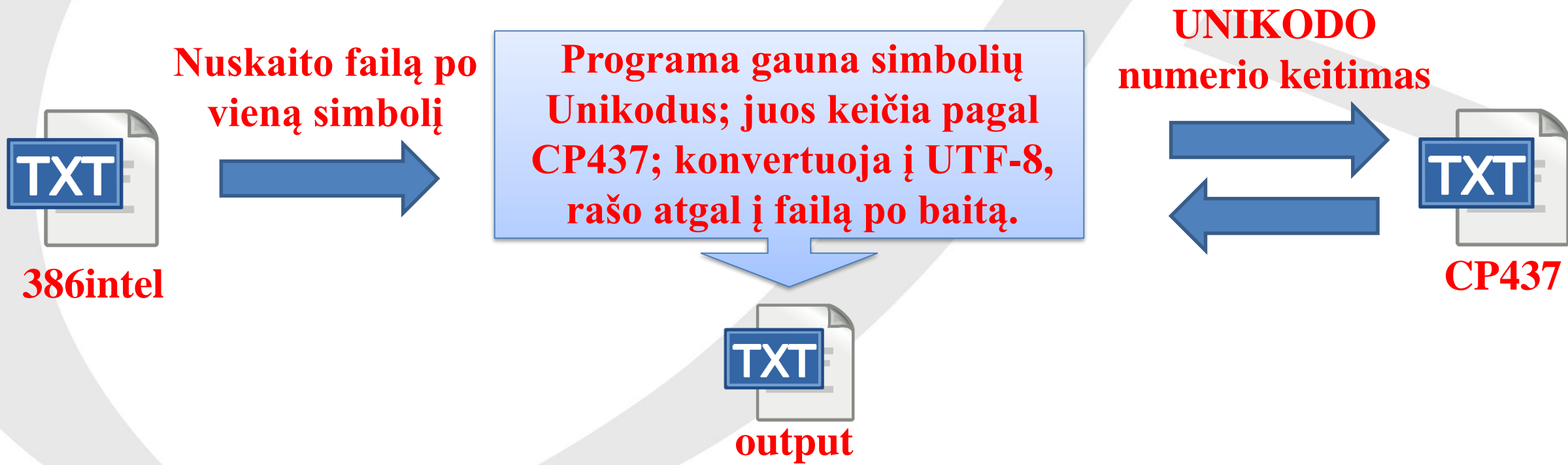


```
C:\WINDOWS\system32\cmd.exe  
  
UNICODE: U+00D2  
UTF-8: C3 92  
CHAR: Ò
```

2 laboratorinis darbas (II dalis)

(iki 2019.10.25)

- **DUOTAS** „386intel“ tekstinis failas, kuris yra užkoduotas CP437 lentele.
- Sukurkite programą, kuri atkoduotų „386intel“ failą, t. y. faile esančių tam tikrų simbolių Unikodus pakeistų į CP437 atitikmenis ir išvestų rezultatą į naują failą.

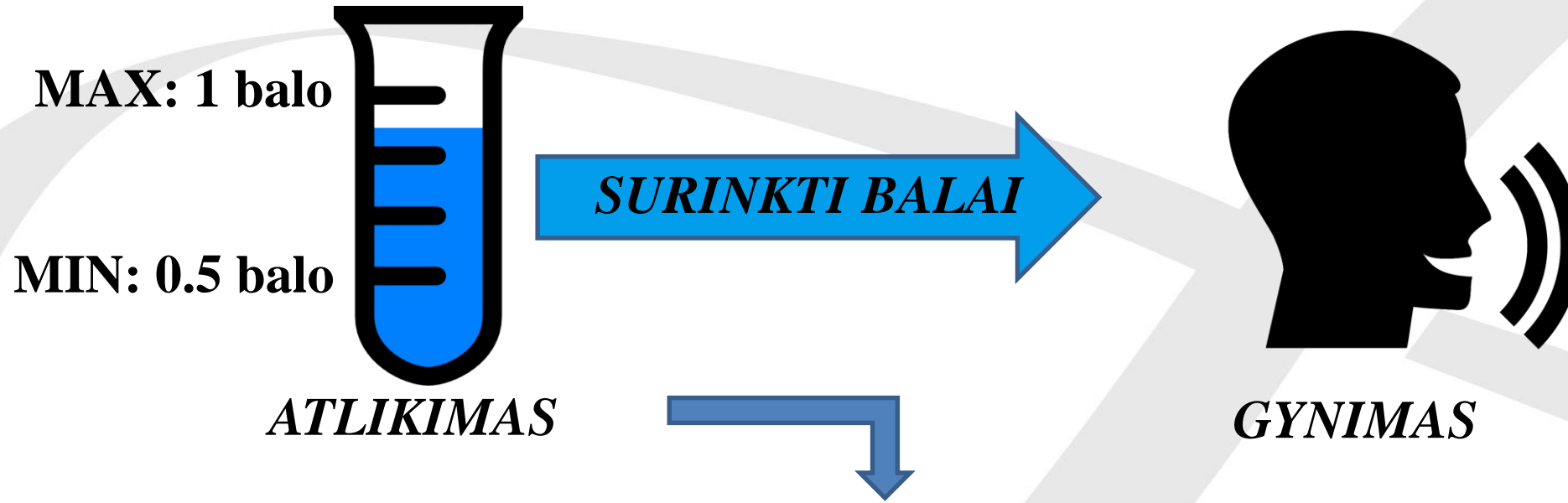


Code page 437

8_	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï	î	ì	Ä	Å
	00C7	00FC	00E9	00E2	00E4	00E0	00E5	00E7	00EA	00EB	00E8	00EF	00EE	00EC	00C4	00C5
	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
9_	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	¢	£	¥	ℳ	ƒ
	00C9	00E6	00C6	00F4	00F6	00F2	00FB	00F9	00FF	00D6	00DC	00A2	00A3	00A5	20A7	0192
	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159
A_	á	í	ó	ú	ñ	Ñ	ª	º	¿	¬	¬	½	¼	¿	«	»
	00E1	00ED	00F3	00FA	00F1	00D1	00AA	00BA	00BF	2310	00AC	00BD	00BC	00A1	00AB	00BB
	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
B_	⌘	⌘	⌘	[d]	†	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡	‡
	2591	2592	2593	2502	2524	2561	2562	2556	2555	2563	2551	2557	255D	255C	255B	2510
	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191
C_	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
	2514	2534	252C	251C	2500	253C	255E	255F	255A	2554	2569	2566	2560	2550	256C	2567
	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207
D_	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘	⌘
	2568	2564	2565	2559	2558	2552	2553	256B	256A	2518	250C	2588	2584	258C	2590	2580
	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223
E_	α	β[e]	Γ	π[ε]	Σ[σ]	σ	μ[h]	τ	Φ	Θ	Ω[ι]	δ[j]	∞	φ[k]	ε[l]	∩
	03B1	00DF	0393	03C0	03A3	03C3	00B5	03C4	03A6	0398	03A9	03B4	221E	03C6	03B5	2229
	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239
F_	≡	±	≥	≤	[m]	J	÷	≈	°	• [n]	•	√[o]	n	²	■	NBSP[P]
	2261	00B1	2265	2264	2320	2321	00F7	2248	00B0	2219	00B7	221A	207F	00B2	25A0	00A0
	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255

- CP437 yra simbolių rinkinys priklausantis IBM PC (asmeniniam kompiuteriui), arba DOS.

Vertinimas



I dalis (0.3)

1. Atsakymo pateikimas standartine forma (0.05).
2. Optimizuotas programos kodas (0.05).
3. Programa veikia teisingai (0.2)

II dalis (0.5 balo)

1. CP437 lentelė paimama iš failo (0.1).
2. Išvedimas į failą pagal UTF-8 (0.2).
3. Programa veikia teisingai (0.2)

Bendri reikalavimai (0.2 balo)

1. Praktiškumas:
 - Patogus/aiškus meniu;
 - Veikia neperkompiliavus;
 - Kiti...