Poročilo druge laboratorijske vaje pri predmetu Informacija in kodi

Rok Prezelj

Univerza v Ljubljani, Fakulteta za elektrotehniko
E-pošta: rp0067@student.uni-lj.si

Povzetek. *Pri laboratorijski vaji smo analizirali kodne tabele, ki podpirajo slovenske šumnike (IBM-852, ISO-8859-2, Windows-1250, MacCE) ter Unicode kodiranja UTF-8, UTF-16LE in UTF-16BE. V drugem delu smo implementirali program, ki iz zaporedja Unicode kodnih točk ročno ustvari pravilno UTF-8 kodirano besedilo.*

1 Uvod

Kodiranje znakov omogoča shranjevanje in prenos besedilnih podatkov v digitalni obliki. V preteklosti so različna okolja uporabljala lastne 8-bitne kodne tabele (npr. IBM-852, ISO-8859-2, Windows-1250, MacCE), ki razširijo 7-bitni ASCII z nacionalnimi znaki. Ker isti bajt v različnih tabelah pogosto predstavlja različne znake, to povzroča težave pri izmenjavi podatkov.

Standard Unicode vsakemu znaku dodeli enolično kodno točko, kodirni sistemi UTF-8, UTF-16LE in UTF-16BE pa določajo pretvorbo v zaporedje bajtov [7]. Pri vaji smo analizirali kodiranje slovenskih šumnikov (Č, Š, Ž, č, š, ž) v izbranih kodnih tabelah, nato pa implementirali ročno pretvorbo iz Unicode kodnih točk v UTF-8.

2 Kodni standardi

V prvem delu vaje smo obravnavali naslednje kodne tabele, ki podpirajo slovenske šumnike:

- **IBM-852**: znan tudi kot DOS Central European, je standard za kodiranje znakov, ki ga je razvil IBM. Razvit je bil predvsem za srednjeevropske jezike, ki uporabljajo latinico [2].
- **ISO-8859-2**: del standarda ISO/IEC 8859, namenjen jezikom srednje Evrope. V zgornji polovici kode vsebuje znake, potrebne za slovenščino, hrvaščino, češčino in slovaščino [3].
- **Windows-1250**: je kodna tabela, ki se uporablja v operacijskem sistemu Microsoft Windows za prikazovanje besedil v srednjeevropskih in vzhodnoevropskih jezikih, ki uporabljajo latinico [4].
- **MacCE**: kodiranje Macintosh Central European se uporablja v računalnikih Apple Macintosh za prikazovanje besedil v srednjeevropskih in jugovzhodnoevropskih jezikih, ki uporabljajo latinico [5].
- **UTF-8**: kodiranje Unicode kodnih točk s spremenljivo dolžino. ASCII znaki so kodirani z enim bajtom, ostali z 2, 3 ali 4 bajti [8].
- **UTF-16LE/UTF-16BE**: 16-bitni kodiranji Unicode, ki za znake v osnovni večjezični ravnini uporabita eno 16-bitno enoto. Razlikujeta se po vrstnem redu bajtov (little-endian in big-endian) [6].

3 Metodologija

Pri izvedbi laboratorijske vaje smo uporabili programski jezik Python zaradi preproste uporabe.

3.1 Kodne tabele slovenskih znakov

Za 8-bitne kodne tabele smo uporabili Pythonovo funkcijo `encode()` na nizu `"ČŠŽčšž"` in iz dobljenih bajtov izračunali desetiške, šestnajstiške in dvojiške predstavitve.

3.2 Program za izračun kodnih točk v UTF-8

V drugem delu vaje smo implementirali program v Pythonu, ki iz vhodne datoteke prebere Unicode kodne točke, ločene z vejicami, in jih pretvori v UTF-8 besedilo. Program deluje po naslednjih korakih:

1. Prebere vsebino datoteke kot ASCII besedilo in jo pretvori v list z števili.
2. Za vsako število preveri, ali gre za veljavno Unicode kodno točko [`0`, `0010FFFF`], ter izključeno surrogate območje [`D800`, `DFFF`].
3. Na podlagi vrednosti kodne točke ročno izračuna UTF-8 predstavitev:
   - 1 bajt: `0xxxxxxx`,
   - 2 bajta: `110xxxxx 10xxxxxx`,
   - 3 bajti: `1110xxxx 10xxxxxx 10xxxxxx`,
   - 4 bajti: `11110xxx 10xxxxxx 10xxxxxx 10xxxxxx`.
4. UTF-8 bajte zapiše v izhodno datoteko v binarnem načinu (`"wb"`).

4 Rezultati

V nadaljevanju so prikazane kodne zamenjave šumnikov ter rezultat pretvorbe Unicode kodnih točk v UTF-8.

4.1 8-bitne kodne tabele

V preglednicah 1–4 so podane kodne zamenjave za slovenske šumnike v izbranih 8-bitnih kodnih tabelah. Vrednosti smo generirali z `.encode(…)`.

znak	desetiško	šestnajstiško	binarno
č	159	9F	10011111
š	231	E7	11100111
ž	167	A7	10100111
Č	172	AC	10101100
Š	230	E6	11100110
Ž	166	A6	10100110

Tabela 1: Kodne zamenjave šumnikov po tabeli IBM-852.

znak	desetiško	šestnajstiško	binarno
č	232	E8	11101000
š	185	B9	10111001
ž	190	BE	10111110
Č	200	C8	11001000
Š	169	A9	10101001
Ž	174	AE	10101110

Tabela 2: Kodne zamenjave šumnikov po tabeli ISO-8859-2.

znak	desetiško	šestnajstiško	binarno
č	232	E8	11101000
š	154	9A	10011010
ž	158	9E	10011110
Č	200	C8	11001000
Š	138	8A	10001010
Ž	142	8E	10001110

Tabela 3: Kodne zamenjave šumnikov po tabeli Windows-1250.

znak	desetiško	šestnajstiško	binarno
č	139	8B	10001011
š	228	E4	11100100
ž	236	EC	11101100
Č	137	89	10001001
Š	225	E1	11100001
Ž	235	EB	11101011

Tabela 4: Kodne zamenjave šumnikov po tabeli MacCE.

V UTF-8 se slovenski šumniki kodirajo z dvema bajtoma. V preglednici 5 je za vsak znak podana njegova Unicode vrednost in pripadajoča UTF-8 predstavitev.

znak	deset.	šest.	binarno
č	50317	C4 8D	11000100 10001101
š	50593	C5 A1	11000101 10100001
ž	50622	C5 BE	11000101 10111110
Č	50316	C4 8C	11000100 10001100
Š	50592	C5 A0	11000101 10100000
Ž	50621	C5 BD	11000101 10111101

Tabela 5: UTF-8 kodne zamenjave za slovenske šumnike (bajti interpretirani kot celo število).

V tabelah 6 in 7 je prikazana 16-bitna binarna predstavitev kodnih točk , pri UTF-16LE in UTF-16BE se bajta te 16-bitne vrednosti zamenjata.

znak	deset.	šest.	binarno
č	3329	0D 01	00001101 00000001
š	24833	61 01	01100001 00000001
ž	32257	7E 01	01111110 00000001
Č	3073	0C 01	00001100 00000001
Š	24577	60 01	01100000 00000001
Ž	32001	7D 01	01111101 00000001

Tabela 6: Kodne zamenjave šumnikov v kodiranju UTF-16LE (bajta interpretirana kot malo-endiško celo število).

znak	deset.	šest.	binarno
č	269	01 0D	00000001 00001101
š	353	01 61	00000001 01100001
ž	382	01 7E	00000001 01111110
Č	268	01 0C	00000001 00001100
Š	352	01 60	00000001 01100000
Ž	381	01 7D	00000001 01111101

Tabela 7: Kodne zamenjave šumnikov v kodiranju UTF-16BE (bajta interpretirana kot veliko-endiško celo število).

Iz preglednic je razvidno, da se šumniki v različnih kodnih tabelah nahajajo na različnih mestih, zato ista bajtna vrednost ne predstavlja nujno istega znaka.

4.2 Pretvorba datoteke v UTF-8

Vhodna datoteka vsebuje zaporedje Unicode kodnih točk, ki predstavljajo prvi stavek wikipedije o UTF-8 v več jezikih. Program je datoteko prebral kot ASCII besedilo ter ročno pretvoril v UTF-8 in rezultat zapisal v izhodno datoteko. Rezultati prvega dela naloge so vidni v prilogi A ter v tabelah 8-11.

5 Zaključek

V vaji smo primerjali kodne tabele za slovenske šumnike ter uspešno implementirali ročno pretvorbo Unicode kodnih točk v UTF-8, kar je potrdilo enoličnost in zanesljivost Unicode kodiranja.

Literatura

[1] N. Pavešić, *Informacija in kodi*, 2. spremenjena in dopolnjena izdaja. Ljubljana: Založba Fakultete za elektrotehniko in Fakultete za računalništvo in informatiko, 2010. ISBN 978-961-243-145-7.

[2] Localizely, Character Encoding: CP 852 [Na spletu]. Dostopno na: `https://localizely.com/character-encodings/cp852/` [Dostopano: 22. november 2025]

[3] *ISO/IEC 8859-2:1999* Information technology — 8-bit single-byte coded graphic character sets — Part 2: Latin alphabet No. 2 [Na spletu]. Dostopno na: `https://www.iso.org/obp/ui/en/#iso:std:iso-iec:8859:-2:ed-1:v1:en` [Dostopano: 22. november 2025]

[4] *Wikipedia* Windows-1250 [Na spletu]. Dostopno na: `https://en.wikipedia.org/wiki/Windows-1250` [Dostopano: 22. november 2025]

[5] *Everything explained today* Mac OS Central European encoding explained [Na spletu]. Dostopno na: `https://everything.explained.today/Mac_OS_Central_European_encoding/` [Dostopano: 22. november 2025]

[6] The Unicode Consortium, *The Unicode Standard, Version 15.0.0*. Mountain View, CA: The Unicode Consortium, 2022. ISBN 978-1-936213-32-0.

[7] The Unicode Consortium, *The Unicode Standard*, v. 15.0. [Na spletu]. Dostopno na: `https://www.unicode.org/standard/standard.html`

[8] F. Yergeau, "UTF-8, a transformation format of ISO 10646," *IETF RFC 3629*, nov. 2003. [Na spletu]. Dostopno na: `https://www.rfc-editor.org/rfc/rfc3629`

[9] Python Software Foundation, *Python 3 Documentation*. [Na spletu]. Dostopno na: `https://docs.python.org/3/`

A Prevedno besedilo

- UTF-8 je eden izmed načinov kodiranja mednarodnega nabora znakov unicode, pri katerem znaki ASCII ostanejo enozložni, ostali znaki pa lahko zasedajo več zlogov.
- UTF-8（8-bit Unicode Transformation Format）是一種針對Unicode的可變長度字元編碼，也是一种前缀码。
- UTF-8은 유니코드를 위한 가변 길이 문자 인코딩 방식 중 하나로, 켄 톰프슨과 롭 파이크가 만들었다.
- Το UTF-8 (8-bit Unicode Transformation Format) είναι ένα μη-απωλεστικό σχήμα κωδικοποίησης χαρακτήρων μεταβλητού μήκους για το πρότυπο Unicode που δημιουργήθηκε από τους Ken Thompson και Rob Pike.

B Tabele

znak	deset.	šest.	binarno
'\n'	10	0A	00001010
'\r'	13	0D	00001101
' '	32	20	00100000
'('	40	28	00101000
')'	41	29	00101001
','	44	2C	00101100
'-'	45	2D	00101101
'.'	46	2E	00101110
'8'	56	38	00111000
'A'	65	41	01000001
'C'	67	43	01000011
'F'	70	46	01000110
'I'	73	49	01001001
'K'	75	4B	01001011
'P'	80	50	01010000
'R'	82	52	01010010
'S'	83	53	01010011
'T'	84	54	01010100
'U'	85	55	01010101
'a'	97	61	01100001
'b'	98	62	01100010
'c'	99	63	01100011
'd'	100	64	01100100
'e'	101	65	01100101
'f'	102	66	01100110
'g'	103	67	01100111
'h'	104	68	01101000
'i'	105	69	01101001
'j'	106	6A	01101010
'k'	107	6B	01101011
'l'	108	6C	01101100
'm'	109	6D	01101101
'n'	110	6E	01101110
'o'	111	6F	01101111
'p'	112	70	01110000
'r'	114	72	01110010
's'	115	73	01110011
't'	116	74	01110100
'u'	117	75	01110101
'v'	118	76	01110110
'z'	122	7A	01111010
'č'	50317	C4 8D	11000100 10001101
'ž'	50622	C5 BE	11000101 10111110

Tabela 8: Unikatni ASCII in slovenski znaki v prevedenem besedilu ter njihove kodne vrednosti.

znak	deset.	šest.	binarno
'Τ'	52900	CE A4	11001110 10100100
'έ'	52909	CE AD	11001110 10101101
'ή'	52910	CE AE	11001110 10101110
'ί'	52911	CE AF	11001110 10101111
'α'	52913	CE B1	11001110 10110001
'β'	52914	CE B2	11001110 10110010
'γ'	52915	CE B3	11001110 10110011
'δ'	52916	CE B4	11001110 10110100
'ε'	52917	CE B5	11001110 10110101
'η'	52919	CE B7	11001110 10110111
'θ'	52920	CE B8	11001110 10111000
'ι'	52921	CE B9	11001110 10111001
'κ'	52922	CE BA	11001110 10111010
'λ'	52923	CE BB	11001110 10111011
'μ'	52924	CE BC	11001110 10111100
'ν'	52925	CE BD	11001110 10111101
'ο'	52927	CE BF	11001110 10111111
'π'	53120	CF 80	11001111 10000000
'ρ'	53121	CF 81	11001111 10000001
'ς'	53122	CF 82	11001111 10000010
'σ'	53123	CF 83	11001111 10000011
'τ'	53124	CF 84	11001111 10000100
'υ'	53125	CF 85	11001111 10000101
'χ'	53127	CF 87	11001111 10000111
'ω'	53129	CF 89	11001111 10001001
'ό'	53132	CF 8C	11001111 10001100
'ύ'	53133	CF 8D	11001111 10001101

Tabela 9: Unikatni grški znaki v prevedenem besedilu in njihove kodne vrednosti.

znak	deset.	šest.	binarno
'。'	14909570	E3 80 82	11100011 10000000 10000010
'一'	14989440	E4 B8 80	11100100 10111000 10000000
'也'	14989727	E4 B9 9F	11100100 10111001 10011111
'元'	15041923	E5 85 83	11100101 10000101 10000011
'前'	15042957	E5 89 8D	11100101 10001001 10001101
'可'	15044527	E5 8F AF	11100101 10001111 10101111
'字'	15052183	E5 AD 97	11100101 10101101 10010111
'對'	15052941	E5 B0 8D	11100101 10110000 10001101
'度'	15055526	E5 BA A6	11100101 10111010 10100110
'是'	15112367	E6 98 AF	11100110 10011000 10101111
'的'	15178372	E7 9A 84	11100111 10011010 10000100
'码'	15179905	E7 A0 81	11100111 10100000 10000001
'碼'	15180476	E7 A2 BC	11100111 10100010 10111100
'种'	15181709	E7 A7 8D	11100111 10100111 10001101
'種'	15181998	E7 A8 AE	11100111 10101000 10101110
'編'	15185832	E7 B7 A8	11100111 10110111 10101000
'缀'	15187072	E7 BC 80	11100111 10111100 10000000
'變'	15249034	E8 AE 8A	11101000 10101110 10001010
'針'	15304605	E9 87 9D	11101001 10000111 10011101
'長'	15308215	E9 95 B7	11101001 10010101 10110111

Tabela 10: Unikatni kitajski znaki (CJK) v prevedenem besedilu in njihove kodne vrednosti.

znak	deset.	šest.	binarno
'가'	15380608	EA B0 80	11101010 10110000 10000000
'과'	15381436	EA B3 BC	11101010 10110011 10111100
'길'	15382712	EA B8 B8	11101010 10111000 10111000
'나'	15434392	EB 82 98	11101011 10000010 10011000
'니'	15436680	EB 8B 88	11101011 10001011 10001000
'다'	15436708	EB 8B A4	11101011 10001011 10100100
'드'	15438748	EB 93 9C	11101011 10010011 10011100
'들'	15438756	EB 93 A4	11101011 10010011 10100100
'딩'	15439017	EB 94 A9	11101011 10010100 10101001
'로'	15442332	EB A1 9C	11101011 10100001 10011100
'롭'	15442349	EB A1 AD	11101011 10100001 10101101
'를'	15443388	EB A5 BC	11101011 10100101 10111100
'만'	15443852	EB A7 8C	11101011 10100111 10001100
'문'	15445176	EB AC B8	11101011 10101100 10111000
'방'	15446185	EB B0 A9	11101011 10110000 10101001
'변'	15446912	EB B3 80	11101011 10110011 10000000
'슨'	15501992	EC 8A A8	11101100 10001010 10101000
'식'	15502237	EC 8B 9D	11101100 10001011 10011101
'었'	15505288	EC 97 88	11101100 10010111 10001000
'위'	15506564	EC 9C 84	11101100 10011100 10000100
'유'	15506592	EC 9C A0	11101100 10011100 10100000
'은'	15506816	EC 9D 80	11101100 10011101 10000000
'이'	15506868	EC 9D B4	11101100 10011101 10110100
'인'	15506872	EC 9D B8	11101100 10011101 10111000
'자'	15507088	EC 9E 90	11101100 10011110 10010000
'중'	15508625	EC A4 91	11101100 10100100 10010001
'켄'	15514756	EC BC 84	11101100 10111100 10000100
'코'	15515028	EC BD 94	11101100 10111101 10010100
'크'	15565228	ED 81 AC	11101101 10000001 10101100
'톰'	15566512	ED 86 B0	11101101 10000110 10110000
'파'	15568012	ED 8C 8C	11101101 10001100 10001100
'프'	15570052	ED 94 84	11101101 10010100 10000100
'하'	15570328	ED 95 98	11101101 10010101 10011000
'한'	15570332	ED 95 9C	11101101 10010101 10011100
' ('	15711368	EF BC 88	11101111 10111100 10001000
') '	15711369	EF BC 89	11101111 10111100 10001001
', '	15711372	EF BC 8C	11101111 10111100 10001100

Tabela 11: Unikatni korejski znaki (Hangul) v prevedenem besedilu in njihove kodne vrednosti.