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 pretvorbo kodnih točk v UTF-8

V drugem delu vaje smo implementirali program v Pythonu, ki iz vhodne datoteke prebere zaporedje 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.

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 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	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).

V UTF-16 se znaki iz osnovne večjezične ravnine predstavijo z eno 16-bitno enoto, enako Unicode vrednosti. V tabelah 6 in 7 je prikazana 16-bitna binarna predstavitev kodnih točk šumnikov; pri UTF-16LE in UTF-16BE se bajta te 16-bitne vrednosti zamenjata.

Iz preglednic je razvidno, da se šumniki v različnih 8-bitnih 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. Generirano datoteko smo uspešno odprli z urejevalnikom besedila, pri čemer so se vsi znaki prikazali pravilno.

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/