Poročilo za drugo laboratorijsko vajo predmeta Informacija in Kodi

Gašper Šavle[1]

[1] *Univerza v Ljubljani, Fakulteta za Elektrotehniko*
E-pošta: gaspersavle@yahoo.com

Abstract

This exercise explored the use of code tables for encoding text files, focusing on Slovenian characters with diacritical marks (Č, Š, Ž, č, š, ž). We studied key code tables like IBM-852, ISO-8859-2, Windows-1250, MacCE, UTF-8, UTF-16LE, and UTF-16BE and their use cases. This provided insights into how different standards handle Slovenian text encoding.

The second part of the exercise involved reading Unicode code points from a file, converting them to characters, and saving the result as a UTF-8 encoded text file.

Through this exercise, I gained practical knowledge of encoding standards and their implementation in programming. It demonstrated the complexity of text encoding in multilingual systems and highlighted the importance of encoding standards like UTF-8 for modern applications.

1 Uvod

Vaja raziskuje uporabo kodnih tabel za kodiranje besedilnih datotek, s poudarkom na slovenskih šumnikih. Analizirali smo kodne tabele IBM-852, ISO-8859-2, Windows-1250, MacCE, UTF-8, UTF-16LE in UTF-16BE ter s pomočjo programa v Python pridobili njihove binarne, decimalne in šestnajstiške zapise za izbrane znake. Tako smo spoznali, kako različni standardi pristopijo k kodiranju slovenskih besedil.

V drugem delu vaje smo prebrali Unicode kodne točke iz vhodne datoteke, jih pretvorili v znake in rezultat zapisali v datoteko, kodirano v UTF-8. Poleg tega smo generirali tabelo unikatnih znakov z njihovimi binarnimi, decimalnimi in šestnajstiškimi zapisi. Naloga je pokazala praktično uporabo kodnih tabel in standarda UTF-8, hkrati pa poudarila pomembnost razumevanja kodiranja v večjezičnih sistemih.

2 Metodologija

Pri reševanju vaje sem uporabil pristop, ki je s pomočjo Python knjižnice `codecs` zakodiral v razlicne kodirne standarde. Nato smo s pomočjo programa pridobili kodne zamenjave za izbrane znake in jih zapisali v različnih zapisih. Pripravljena je bila vhodna datoteka z Unicode kodnimi točkami, ki smo jih pretvorili v znake in zapisali v UTF-8 kodirani izhodni datoteki. Rezultate smo analizirali z izpisom unikatnih znakov in njihovih kodnih zamenjav.

Kodne tabele so sistemi za kodiranje znakov, ki omogočajo pretvorbo besedila v računalniško berljivo obliko. Vsaka kodna tabela določa nabor znakov in njihovo predstavitev v binarni obliki. Različni kodni standardi kodirajo besedila na različne načine zaradi zgodovinskih razlogov, jezikovne raznolikosti, platformne združljivosti, tehničnih omejitev in razlicnih stopenj razvoja standardizacije skozi čas. V prvih računalniških sistemih so bile kodne tabele omejene na 8-bitne formate, prilagojene lokalnim jezikom in potrebam, kot je na primer IBM-852 za srednjo Evropo. Različni jeziki uporabljajo specifične znake, kar je povzročilo razvoj tabel, kot sta Windows-1250 in MacCE, ki so zadostile regionalnim potrebam. Proizvajalci, kot so IBM, Microsoft in Apple, so ustvarili lastne standarde za združljivost svojih sistemov. Tehnične omejitve, kot so vrstni red bitov in optimizacija za prostor ali hitrost, so vplivale na zasnovo kod, kot sta UTF-16LE in UTF-8. Razvoj Unicode je rešil omejitve prejšnjih standardov z univerzalnim kodiranjem vseh znakov in združljivostjo z obstoječimi sistemi, zaradi česar je danes postal globalni standard. [1]

2.1 IBM-852

Znana tudi kot CP852, je kodna tabela, ki jo je IBM razvila za kodiranje znakov srednjeevropskih jezikov v MS-DOS okolju. Poleg osnovnih ASCII znakov vključuje znake, potrebne za jezike, kot so poljščina, češčina, slovaščina, madžarščina in drugi. Uporabljala se je predvsem v operacijskih sistemih DOS in zgodnjih različicah Windows za pravilno prikazovanje besedil v teh jezikih. Je 8-bitna kodna tabela, ki omogoča kodiranje do 256 znakov, od katerih prvih 128 ustreza standardu ASCII, nadaljnih 128 pa ustreza diakritičnim znakom zgoraj navedenih srednjeevropskih jezikov. [2]

2.2 ISO-8859-2

Znana tudi kot Latin-2, je del standarda ISO/IEC 8859 in je namenjena kodiranju znakov srednje in vzhodnoevropskih jezikov, kot so poljščina, češčina, slovaščina, madžarščina, slovenščina in drugi. Omogoča kodiranje do 256 znakov, od katerih prvih 128 ustreza standardu ASCII, druga polovica pa vsebuje diakritične znake zgoraj omenjenih jezikov. Vključuje dodatne znake, ki niso

prisotni v ISO-8859-1 (Latin-1), in je bila široko uporabljena v Unix sistemih ter na spletu za prikazovanje besedil v teh jezikih. [3]

2.3 Windows-1250

Windows-1250 je kodna tabela, ki jo je Microsoft razvil za kodiranje znakov srednjeevropskih jezikov v operacijskih sistemih Windows. Vključuje znake za jezike, kot so poljščina, češčina, slovaščina, madžarščina, slovenščina in drugi. Uporabljala se je v Windows aplikacijah za pravilno prikazovanje besedil v teh jezikih. Kot prejsnje opisani kodni standardi, je tudi ta prav tako 8-bitni standard, od koder prvih 128 znakov ustreza naboru ASCII, naslednji pa ustrezajo zgoraj navedenim jezikom. Ta kodna tabela ključuje nekaj znakov, ki so specifični za Microsoftove aplikacije in se rahlo razlikuje od ISO-8859-2, kar lahko povzroči težave z združljivostjo v nekaterih primerih. [4]

2.4 MacCE

MacCE (Mac Central Europe) je kodna tabela, ki jo je Apple razvil za kodiranje znakov srednjeevropskih jezikov v operacijskem sistemu Macintosh. Vključuje znake za jezike, kot so poljščina, češčina, slovaščina, madžarščina, slovenščina in drugi. Uporabljala se je v Mac aplikacijah za pravilno prikazovanje besedil v teh jezikih. Prav tako, kot prejsnji primeri, je tudi ta tabela 8-bitna in je srednjeevropska razširitev osnovnega nabora znakov ASCII. [5]

2.5 UTF-8

UTF-8 je bil razvit leta 1993 kot del sistema Unicode za kodiranje besedila in je postal prevladujoč standard zaradi svoje združljivosti z ASCII in učinkovitega kodiranja. Unicode, ki je nastal leta 1991, si je prizadeval združiti obstoječe kodne standarde, da bi omogočil univerzalno kodiranje znakov za vse jezike. UTF-8 uporablja spremenljivo dolžino podatka (1 do 4 bajte) za kodiranje znakov, kar omogoča optimizacijo prostora za pogoste znake, kot so ASCII, medtem ko podpira celoten nabor Unicode. Zaradi fleksibilnosti in interoperabilnosti je UTF-8 postal osnova za večino sodobnih sistemov, vključno s spletom in aplikacijami. [6]

2.6 UTF-16LE in UTF-16BE

UTF-16 (16-bitni Unicode Transformation Format) je kodna tabela, ki uporablja 16-bitne enote za kodiranje znakov Unicode. Omogoča kodiranje celotnega nabora Unicode znakov in se uporablja v različnih aplikacijah ter protokolih, kjer je potrebna podpora za širok nabor znakov. [7] Obstajata dve različici:

- **Big-endian** V bajtu informacije shranjuje najmanj pomemben bit (najnižja utež) pred najbolj pomembnim (najvišja utež). Uporablja se v sistemih, ki uporabljajo little-endian arhitekturo.
- **Little-endian** V bajtu informacije shranjuje najbolj pomemben bit pred najmanj pomembnim. Uporablja se v sistemih, ki uporabljajo big-endian arhitekturo.

3 Rezultati

V tem poglavju so predstavljeni rezultati analiz kodiranja slovenskih šumnikov (Č, Š, Ž in njihove male oblike) v različnih kodnih standardih, kot so IBM-852, ISO-8859-2, Windows-1250, MacCE, UTF-16 in UTF-8. Poleg tega so obravnavani procesi kodiranja in dekodiranja Unicode znakov v UTF-8, vključno s pretvorbo kodnih mest, zapisom izhodnih datotek in identifikacijo unikatnih znakov. Rezultati prikazujejo podporo slovenskega jezika v teh standardih.

Tabela 1: Kodiranje slovenskih diakritičnih znakov s standardom IBM-852

IBM-852			
Znak	DEC	HEX	BIN
Č	172	0xac	10101100
Š	230	0xe6	11100110
Ž	166	0xa6	10100110
č	159	0x9f	10011111
š	231	0xe7	11100111
ž	167	0xa7	10100111

Tabela 2: Kodiranje slovenskih diakritičnih znakov s standardom ISO-8859-2

ISO-8859-2			
Znak	DEC	HEX	BIN
Č	200	0xc8	11001000
Š	169	0xa9	10101001
Ž	174	0xae	10101110
č	232	0xe8	11101000
š	185	0xb9	10111001
ž	190	0xbe	10111110

Tabela 3: Kodiranje slovenskih diakritičnih znakov s standardom Windows-1250,

Windows-1250			
Znak	DEC	HEX	BIN
Č	200	0xc8	11001000
Š	138	0x8a	10001010
Ž	142	0x8e	10001110
č	232	0xe8	11101000
š	154	0x9a	10011010
ž	158	0x9e	10011110

Tabela 4: Kodiranje slovenskih diakritičnih znakov s standardom MacCE,

MacCE			
Znak	DEC	HEX	BIN
Č	200	0xc8	11001000
Š	138	0x8a	10001010
Ž	142	0x8e	10001110
č	232	0xe8	11101000
š	154	0x9a	10011010
ž	158	0x9e	10011110

Tabela 5: Kodiranje slovenskih diakritičnih znakov s standardom UTF-8,

UTF-8			
Znak	DEC	HEX	BIN
Č	196, 140	0xc4, 0x8c	11000100 10001100
Š	197, 160	0xc5, 0xa0	11000101 10100000
Ž	197, 189	0xc5, 0xbd	11000101 10111101
č	196, 141	0xc4, 0x8d	11000100 10001101
š	197, 161	0xc5, 0xa1	11000101 10100001
ž	197, 190	0xc5, 0xbe	11000101 10111110

Tabela 6: Kodiranje slovenskih diakritičnih znakov s standardoma UTF-16LE in UTF-16BE.

UTF-16LE / UTF-16BE				
Endian	Znak	DEC	HEX	BIN
LE	Č	12, 1	0xc, 0x1	00001100 00000001
	Š	96, 1	0x60, 0x1	01100000 00000001
	Ž	125, 1	0x7d, 0x1	01111101 00000001
	č	13, 1	0xd, 0x1	00001101 00000001
	š	97, 1	0x61, 0x1	01100001 00000001
	ž	126, 1	0x7e, 0x1	01111110 00000001
BE	Č	1, 12	0x1, 0xc	00000001 00001100
	Š	1, 96	0x1, 0x60	00000001 01100000
	Ž	1, 125	0x1, 0x7d	00000001 01111101
	č	1, 13	0x1, 0xd	00000001 00001101
	š	1, 97	0x1, 0x61	00000001 01100001
	ž	1, 126	0x1, 0x7e	00000001 01111110

Literatura

[1] Wikipedia contributors, "Code page — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[2] Wikipedia contributors, "Code page 852 — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[3] Wikipedia contributors, "Iso/iec 8859-2 — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[4] Wikipedia contributors, "Windows-1250 — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[5] Wikipedia contributors, "Mac os central european encoding — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[6] Wikipedia contributors, "Utf-8 — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].

[7] Wikipedia contributors, "Utf-16 — Wikipedia, the free encyclopedia," 2024. [Online; accessed 20-November-2024].