

Infroformacija in kodi

Poročilo druge laboratorijske vaje

Avtor: Bor Starčič

Mentorja: izr.prof.dr.Simon Dobrišek, as.dr.Klemen Grm

Datum: 21. april 2024

Kazalo

1	Prva Naloga	3
1.1	Kodne tabele	3
2	Druga Naloga	5

1 Prva Naloga

V prvi nalogi smo se poglobili v raziskovanje različnih znakovnih kodnih tabel, ki omogočajo, da se znaki iz človeškega jezika pretvorijo v digitalne zapise. Osredotočili smo se na šumnike Č, Š, Ž, č, š, ž, ki so pogosto uporabljeni v slovenskem jeziku. Preučili smo kako so ti znaki kodirani v kodnih tabelah: IBM-852, ISO-8859-2, Windows-1250, Mac Croatian, UTF-8, UTF-16LE in UTF-16BE.

Za vsako od teh kodnih tabel smo izvedli kodiranje šumnikov in rezultate prikazali v binarni, decimalni in hexadecimalni obliki. Tako smo lahko primerjali kako kodne tabele obravnavajo posamezne znake. Uporabili smo funkcijo encode, ki omogoča pretvorbo stringov v določeno kodiranje.

1.1 Kodne tabele

- **IBM-852:** Znan tudi kot CP 852, se uporablja za prikazovanje besedil v srednjeevropskih jezikih, kot so češčina, slovenščina, srbo-hrvaščina, poljščina, romunščina in madžarščina.
- **ISO-8859-2:** Podpira srednjeevropske jezike in je del serije standardov ISO-8859, vključno s češčino, slovaščino, poljščino in madžarščino.
- **Windows-1250:** Razvila ga je Microsoft, združljiva s ISO-8859-2, se uporablja za prikaz srednjeevropskih jezikov.
- **Mac-croatian):** Kodna tabela za Macintosh računalnike, namenjena podpori hrvaškega jezika in je najbližji MacCE, ki ga v pythonu ni.
- **UTF-8:** Univerzalna kodna tabela variabilne dolžine, ki podpira skoraj vse znake svetovnih jezikov in zmore predstavljati kateri koli znak v Unicode standardu.
- **UTF-16LE:** Kodira Unicode znake z 16-bitnimi kodnimi enotami v formatu Little Endian, pogosto uporabljena v Windows okoljih.
- **UTF-16BE:** Kodira Unicode znake z 16-bitnimi kodnimi enotami v formatu Big Endian, uporabljena predvsem v Unix in Linux sistemih.

CP852:

Č: BIN(10101100) DEC(172) HEX(AC)
Š: BIN(11100110) DEC(230) HEX(E6)
Ž: BIN(10100110) DEC(166) HEX(A6)
č: BIN(10011111) DEC(159) HEX(9F)
š: BIN(11100111) DEC(231) HEX(E7)
ž: BIN(10100111) DEC(167) HEX(A7)

ISO-8859-2:

Č: BIN(11001000) DEC(200) HEX(C8)
Š: BIN(10101001) DEC(169) HEX(A9)
Ž: BIN(10101110) DEC(174) HEX(AE)
č: BIN(11101000) DEC(232) HEX(E8)
š: BIN(10111001) DEC(185) HEX(B9)
ž: BIN(10111110) DEC(190) HEX(BE)

WINDOWS-1250:

Č: BIN(11001000) DEC(200) HEX(C8)
Š: BIN(10001010) DEC(138) HEX(8A)
Ž: BIN(10001110) DEC(142) HEX(8E)
č: BIN(11101000) DEC(232) HEX(E8)
š: BIN(10011010) DEC(154) HEX(9A)
ž: BIN(10011110) DEC(158) HEX(9E)

MAC-CROATIAN:

Č: BIN(11001000) DEC(200) HEX(C8)
Š: BIN(10101001) DEC(169) HEX(A9)
Ž: BIN(10101110) DEC(174) HEX(AE)
č: BIN(11101000) DEC(232) HEX(E8)
š: BIN(10111001) DEC(185) HEX(B9)
ž: BIN(10111110) DEC(190) HEX(BE)

Slika 1: Zapisi v kodnih tabelah: IBM852, ISO 8859 2, Windows 1250 in MAC Croatian

```

UTF-8:
Č: BIN(1100010010001100) DEC(50316) HEX(C48C)
Š: BIN(1100010110100000) DEC(50592) HEX(C5A0)
Ž: BIN(1100010110111101) DEC(50621) HEX(C5BD)
č: BIN(1100010010001101) DEC(50317) HEX(C48D)
š: BIN(1100010110100001) DEC(50593) HEX(C5A1)
ž: BIN(1100010110111110) DEC(50622) HEX(C5BE)

UTF-16LE:
Č: BIN(1100000000001) DEC(3073) HEX(0C01)
Š: BIN(1100000000000001) DEC(24577) HEX(6001)
Ž: BIN(1111101000000001) DEC(32001) HEX(7D01)
č: BIN(110100000001) DEC(3329) HEX(0D01)
š: BIN(1100001000000001) DEC(24833) HEX(6101)
ž: BIN(1111110000000001) DEC(32257) HEX(7E01)

UTF-16BE:
Č: BIN(100001100) DEC(268) HEX(010C)
Š: BIN(101100000) DEC(352) HEX(0160)
Ž: BIN(101111101) DEC(381) HEX(017D)
č: BIN(100001101) DEC(269) HEX(010D)
š: BIN(101100001) DEC(353) HEX(0161)
ž: BIN(101111110) DEC(382) HEX(017E)

```

Slika 2: Zapisi v kodnih tabelah: UTF 8, UTF 16LE in UTF 16BE

Na slikah so prikazani rezultati prve naloge, ki prikazujejo zapis šumnikov v različnih kodnih tabelah v binarnem, decimalnem in hexadecimalnem zapisu.

2 Druga Naloga

Glavni cilj naloge je bil izvesti pretvorbo seznama decimalnih števil iz datoteke `kodne_tocke.txt` v njihove ustrezne Unicode znake. Ti znaki so bili zapisani v izhodno datoteko z uporabo UTF-8 kodiranja pod imenom `generated_text.txt`.

Decimalne Unicode točke smo prebrali iz datoteke `kodne_tocke.txt` in jih pretvorili v heksadecimalne zapise. Te zapise smo nato uporabili za dekodiranje v ustrezne Unicode znake. Pretvorjene znake smo zapisali v datoteko `generated_text.txt`, pri čemer smo uporabili UTF-8 kodiranje. S pomočjo knjižnice `openpyxl` smo generirali Excel tabelo, ki vsebuje unikatne znake. Stolpci v tabeli so organizirani tako,

da prikazujejo binarne, decimalne in heksadecimalne kode znakov. Končni rezultat smo shranili v `unique_characters_codes.xlsx`.