Infrormacija in kodi

Poročilo druge laboratorijske vaje

Avtor: Bor Starčič

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

Datum: 21. april 2024

Kazalo

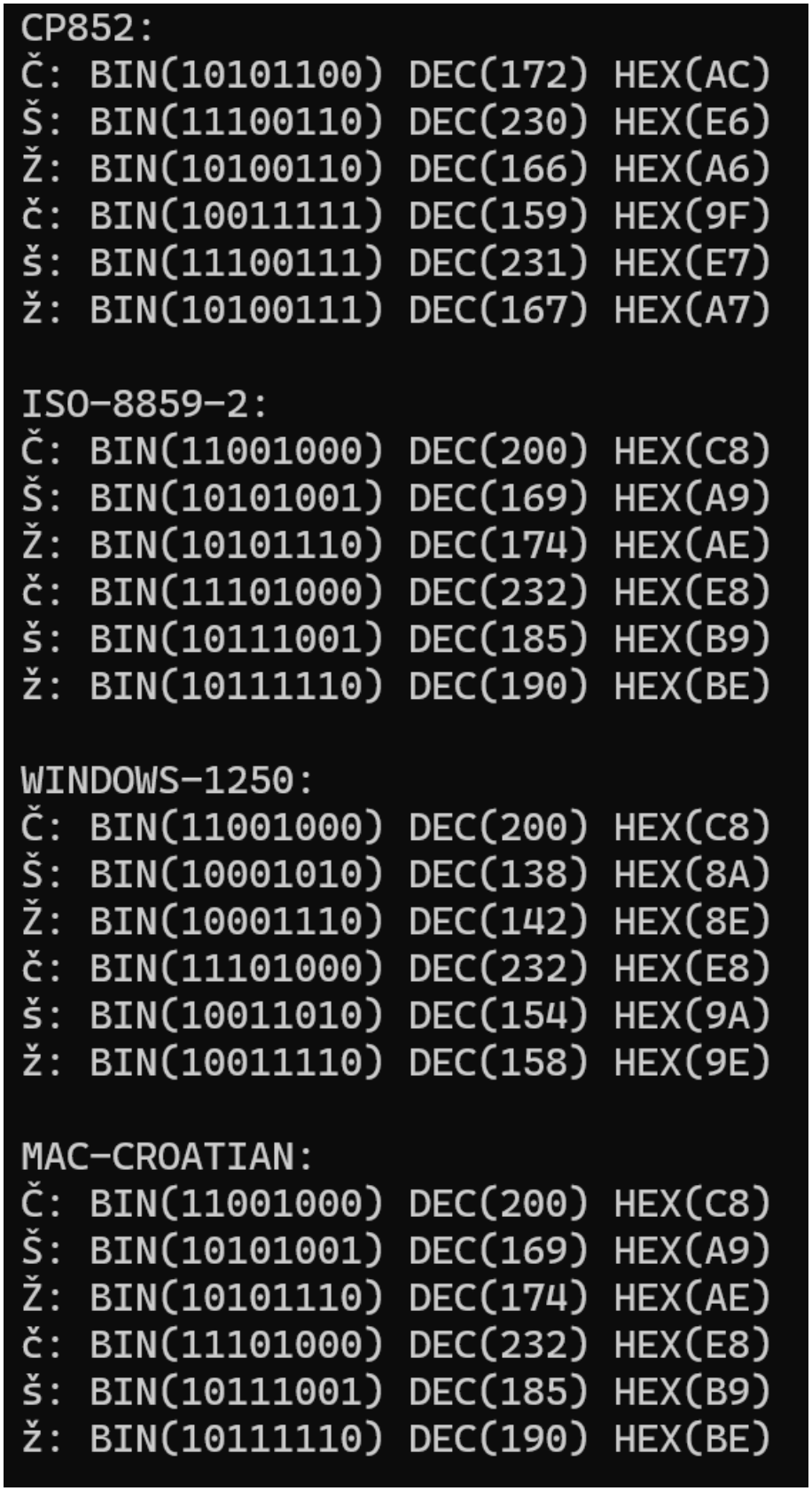
1 Prva Naloga

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.

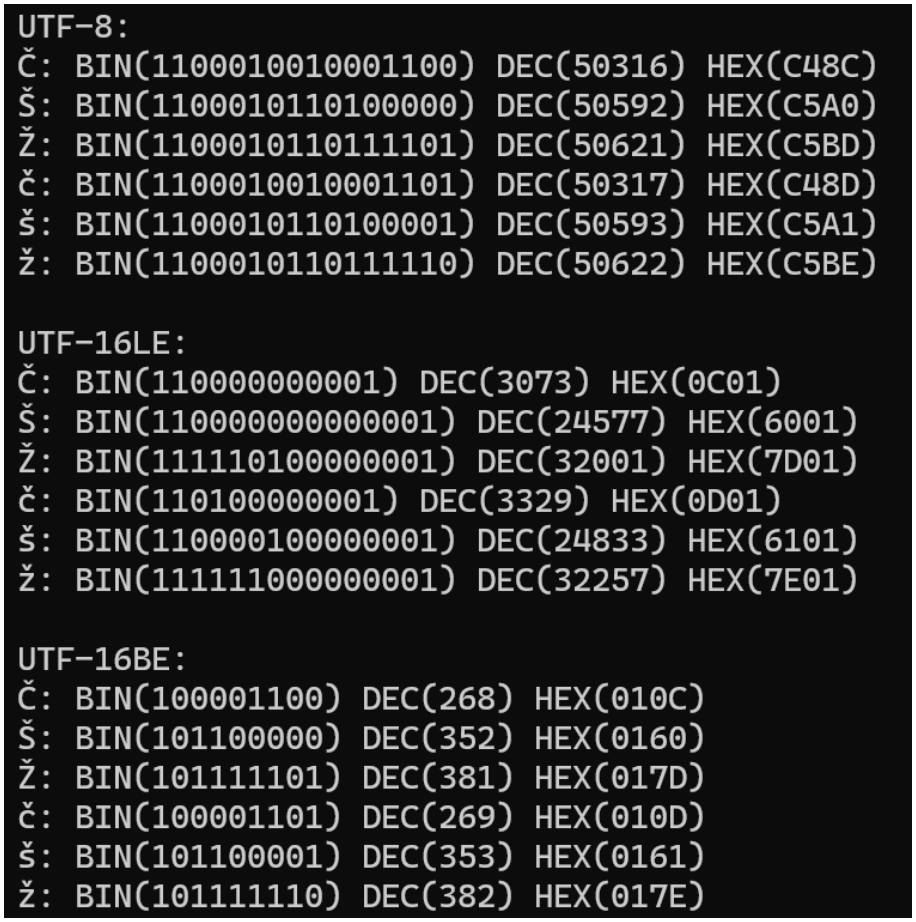
Za vsako od teh kodnih tabel smo izvedli kodiranje šumnikov in rezultate prikazali v binarni, decimalni in hexadecimalni obliki. Tako smo lahko primerjali kako kodne tabele obravnavajo posamezne znake. Uporabili smo funkcijo encode, ki omogoča pretvorbo stringov v določeno kodiranje.

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.



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



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`.