Baze, podaci i oko čega se truditi

Doni Pracner

9. Jun 2022.

Teme za danas

- Tipovi fajlova: binarni i tekstualni
- Kada su tekstualni bolji
- Baze podataka - prednosti i mane
- Izvoženje podataka u razne tekstualne formate
- Dalje obrada podataka
- Konkretni primeri konverzija
- … i opet baze

Tipovi fajlova

Svi fajlovi su u suštini niz bajtova

tekstualni fajlovi

- najčešće svaki bajt po jedno slovo
- Programski kodovi, HTML stranice, običan tekst, ...
- Slova se tumače na osnovu encoding-a
- UTF-8 koristi 1 ili 2 bajta

binarni fajlovi

- bajtovi se interpretiraju u zavisnosti od tipa podataka
- Izvršni fajlovi, MS Office fajlovi, slike, muzika, filmovi ...

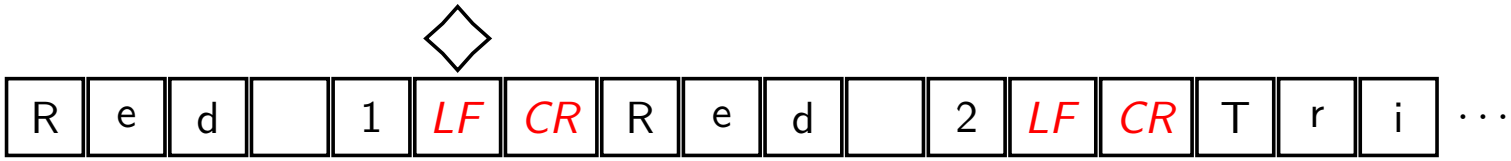
Tekstualni fajlovi

- Zgodni za brze promene, direktno čitanje bez dodatnih programa
- Mnogi sistemi preferiraju da su im ulazi i izlazi ovakvi fajlovi
- Klasično se otvaraju u notepad-u
  - Mada mogu i u drugima, daleko boljima
  - Notepad++, jEdit, Geany, Visual Studio Code, …

Znaci za prelom redova

Ne postoji fizički prelom reda - koriste se specijalni znaci
Na Windows-u:

- Kod 10: Line feed - red dole na štampaču
- Kod 13: Carriage return - povratak na početak reda



Primer u kome je učitan prvi red i kursor je na specijalnom znaku

Baze podataka

- Baze za tipičnog korisnika nisu ni jedno ni drugo
- Skladište se u binarnim fajlovima tipično
- Veoma brzi i kontrolisani pristupi
- Napredne veze između podataka
- Tipično centralizovane, pristupa se preko nekakvih programa/sajtova

Problemi sa bazama podataka

- Teško imati lokalne kopije
- Nije zgodno "igrati se" sa podacima - brljamo sve
- Tipično je obrada podataka ograničena programom
  - Npr ako nema dugme za "prosek" nećemo ga imati
- Baze su često "krute" sa podacima

Predstavljanje tabelarnih podataka

Predstavljanje velikog broja organizovanih podataka

Excel i razni srodni programi

- binarni fajlovi
- format nezgodan za lepa učitavanja za druge programe

CSV – comma separated values, tj vrednosti razdvojene zarezima

- čist tekstualan fajl
- lako se čita od strane ljudi, ali i raznih programa

CSV

- Separatori bi trebali biti zarezi

```
kol1,kol2,kol3
a,b,c
123,3345,125666534
```

- Ali nekad nažalost nisu

Izvoz iz Excel-a u CSV

- Excel je veoma moćan program
- Nekad je *suviše* pametan
- Budući da je integrisan sa Windows-om, pretpostavlja šta je korisniku najbolje

Lokalna podešavanja i Excel

- Engleski jezik preferira:
  - Decimalni separator `.`
  - Separator hiljada `,`
- Kad je sve na Srpskom (i mnogim drugim jezicima):
  - Decimalni separator je `,`
  - Separator hiljada `.`
- Separator u CSV je isto kao i separator hiljada, iz nekog razloga

```
kol1;kol2;kol3
a;b;c
123;3345;125666534
```

Excel i kontrolisani zarezi

- Nažalost teško - promene na nivou operativnog sistema tipično

Alternativa: izvoz iz Libreoffice

- Open/Libre Office
- Nudi da se bira i na uvozu i na izvozu
- Jake preporuke:
  - , za separator
  - UTF-8 za enkoding

Izvozi za Wombat

- Pedigre fajl
  - Id životinje
  - Id oca
  - Id majke

Redosled je bitan

- Wombat zahteva da su roditelji sa nižim Id-om nego deca
- Realno, često nisu

Renumeracija

- Program koji prima pedigre
- Nalazi nove brojeve i ispisuje novi fajl

Primeri

- Prikaz rada sa programom i fajlovima
- Samostalne vežbe

Fajlovi sa fiksiranim širinama kolona

- *Fixed width columns*
- Nema zareza, gledaju se tačno znaci u redu
- Womabat insistira na ovakvim ulazima

```
#      grlo       otac      majka
       3785       3784       3781
       4197       4196       4192
       4271       4270       4266
       4370       4368       4369
       4626       4625       4620
```

Izvoz u fiksirane širine

- Nažalost često problematičan
- Moderni programi ne vole ovo
- Excel/Libreoffice aproksimiraju širine
  - Moguće greške
  - Tzv *PRN* fajlovi
- Pri uvozu se tipično može uraditi kako treba sve

Program FiksneSirine

- Uzima ulazni CSV
- U okviru fajla `opcije.txt` treba da postoji specifikacija kolona
  - *Kolone=3,5,6,3,2*

Primeri

- Prikaz rada sa programom i fajlovima
- Samostalne vežbe

Baze podataka, opet

- U okviru BioITGenoSelect projekta:
  - Modeliranje podataka u bazi
  - Sajt za pristup, pretragu, promenu podataka
  - Razni izvozi, uvozi i izveštaji

Izvozi iz baza, prednosti i mane

- Wombat pedigre fajl
- Wombat dat fajl

Uvoženje u baze

- Problemi sa nekonzistentnim ili nedostajućim podacima
- Imena kolona i svašta slično mora biti konzsistentno
  - nekad može po brojevima kolona, nekad po imenima, pitanje koje bolje

Hvala na pažnji

Pitanja?