$Id: duomenų-analizė-next-year.txt 5243 2018-12-17 17:10:53Z saulius $

Duomenų tyrimas Unix įrankiais

Šią užduotį reikia atlikti programa, kurią parašėte 3-jai užduočiai. Jei tokios programos neparašėte, arba ji neveikia tinkamai, tuomet kreipkitės į dėstytoją, ir Jums bus parūpinta kita programa ir/arba kita užduotis; tačiau tokiu atveju Jūsų 4-os užduoties balai bus dauginami iš koeficiento 0.75 (t.y. bus užskaityti tik 3/4 Jūsų surinktų taškų).

Atlikite duomenų tyrimą su didesne (300–1000 failų) duomenų imtimi iš Jums skirtos duomenų bazės. Nukelkite failus; nukėlimą dokumentuokite (set -x; ..) metodu, shell skriptu („scenarijumi“) arba Makefile'u. Pradinius *nemodifikuotus* duomenis sudėkite į direktoriją inputs/; šių duomenų *nekeiskite*!

Jei pradiniai duomenys didesni, negu 100MB, jų kelti į repozitoriją nereikia, bet tokiu atveju *būtina* įkelti identifikatorių sąrašą. Kaip identifikatorius galima naudoti *stabilius* URI arba duomenų bazės identifikatorius (PDB ID, COD ID, Uniprot ID, DOI, etc.).

Parašykite Makefile'ą, kuris iš pradinių duomenų inputs/ direktorijoje sugeneruotų tarpinius ir galutinius rezultatų failus direktorijoje outputs/. Parašykite taisyklę tikslui ‘validate’, kuris patikrintų Jūsų programos sukurtų failų validumą pagal formato aprašymą (regexp arba schemą). Ar reikia tikslą ‘validate’ paskelbti kaip .PHONY, .PRECIOUS ar .INTERMEDIATE?

Įgyvendinkite savo Makefile tikslus ‘clean’ ir ‘distclean’. Tikslas ‘clean’ turi ištrinti tarpinius rezultatus (ypač tuos, kurie paskelbti .INTERMEDIATE); tikslas ‘distclean’ turi atlikti tas pačias komandas, kaip ir ‘clean’, ir, be to, ištrinti galutinius suskaičiuotus rezultatus.

Visi tarpiniai rezultatų failai turi turėti laiko žymes, t.y. juose turi būti eilutė (tai gali būti pirmoji eilutė), automatiškai įterpiama ‘make’ pagalba, su skaičiavimo momento data ir laiku; pvz.:

```
saulius@varanas PDB/ $ head outputs/downloads/pdb/1h/1h2x.biblst
# 2018-02-16 04:53:24 EET
10.1074/JBC.M208043200  10.1074/JBC.M007003200
...
```

Make taisyklės pavyzdys:

```
%.biblst: %.cif.gz
    @mkdir -p $(dir $@)
    date +"# %F %T %Z" > $@
    pdbx-bibliography $< >> $@
```

Galutinių rezultatų faile Unix komandų pagalba suskaičiuokite lentelę (Keyword-space-value (KSV) formatu) su Jums paskirto parametro reikšmėmis kiekvienam tyrinėtam failui ir DB identifikatoriui; jei reikia, taip pat lenteles su Jums paskirtu parametru ir kitais parametrais, nuo kurių priklausomybę turite ištirti (tokiu pat KSV formatu). Pvz.:

```
# 2018-12-17 18:57:39 EET
#FILE: pdb-sequence-length-resolution.tab
#PDBID chain seqlen resolution
1KNV A 290 2.17
1XYZ A 320 1.40
```

Naudodami R, Gnuplot ar kitą Jums žinomą grafinę programą, valdomą komandomis, sugeneruokite tyrinėjamo savo parametro reikšmių histogramą ir priklausomybės nuo kitų parametrų grafiką (scatterplot). Suskaičiuokite savo nagrinėjamo parametro vidutinę reikšmę ir vidutinį kvadratinį nuokrypį. Komandos, generuojančios paveiksliukus ir skaičiuoja parametrus, turi būti skriptuose, kurie kviečiami Make sistemos.

Parašykite 1-2 A4 formato psl. ataskaitą (9-11 pt šriftu; naudokite HTML, TXT, ODT arba LaTeX formatus); joje pateikite savo parametro histogramą ir priklausomybės grafiką; aptarkite, kokias išvadas galite padaryti iš šių grafikų – ar stebite koreliacijas tarp parametrų; kokia gali būti koreliacijų (ne)buvimo priežastis; pabandykite neformaliai įvertinti, ar koreliacijos ir Jūsų nustatyti parametrai yra statistiškai reikšmingi; ar jie atsikartos, jei pakartosite savo tyrimą dar kartą su kita duomenų imtimi. Ataskaitą paverskite PDF formato failu; įkelkite šiuos rezultatus į repozitoriją. Vertinimui pateikite Moodle sistemoje savo sutvarkytos Subversijos darbinės kopijos .zip arba .tgz archyvą.

Jūsų parašyta ataskaita bus Jūsų pranešimo kurso pabaigoje pagrindas.