1. Úloha IOS (2022)

Popis úlohy

Cílem úlohy je vytvořit shellový skript pro analýzu záznamů osob s prokázanou nákazou koronavirem způsobujícím onemocnění COVID-19 na území České republiky. Skript bude filtrovat záznamy a poskytovat základní statistiky podle zadání uživatele.

Specifikace rozhraní skriptu

JMÉNO

- `corona` — analyzátor záznamů osob s prokázanou nákazou koronavirem způsobujícím onemocnění COVID-19

POUŽITÍ

- `corona [-h] [FILTERS] [COMMAND] [LOG [LOG2 [...]]`

VOLBY

- `COMMAND` může být jeden z:
  - `infected` — spočítá počet nakažených.
  - `merge` — sloučí několik souborů se záznamy do jednoho, zachovávající původní pořadí (hlavička bude ve výstupu jen jednou).
  - `gender` — vypíše počet nakažených pro jednotlivá pohlaví.
  - `age` — vypíše statistiku počtu nakažených osob dle věku (bližší popis je níže).
  - `daily` — vypíše statistiku nakažených osob pro jednotlivé dny.
  - `monthly` — vypíše statistiku nakažených osob pro jednotlivé měsíce.
  - `yearly` — vypíše statistiku nakažených osob pro jednotlivé roky.
  - `countries` — vypíše statistiku nakažených osob pro jednotlivé země nákazy (bez ČR, tj. kódu `CZ`).
  - `districts` — vypíše statistiku nakažených osob pro jednotlivé okresy.
  - `regions` — vypíše statistiku nakažených osob pro jednotlivé kraje.
- `FILTERS` může být kombinace následujících (každý maximálně jednou):
  - `-a DATETIME` — after: jsou uvažovány pouze záznamy PO tomto datu (včetně tohoto data). `DATETIME` je formátu `YYYY-MM-DD`.
  - `-b DATETIME` — before: jsou uvažovány pouze záznamy PŘED tímto datem (včetně tohoto data).
  - `-g GENDER` — jsou uvažovány pouze záznamy nakažených osob daného pohlaví. `GENDER` může být `M` (muži) nebo `Z` (ženy).
  - `-s [WIDTH]` u příkazů `gender`, `age`, `daily`, `monthly`, `yearly`, `countries`, `districts` a `regions` vypisuje data ne číselně, ale graficky v podobě histogramů. Nepovinný parametr `WIDTH` nastavuje šířku histogramů, tedy délku nejdelšího řádku, na `WIDTH`. Tedy, `WIDTH` musí být kladné celé číslo. Pokud není parametr `WIDTH` uveden, řídí se šířky řádků požadavky uvedenými níže.
  - (**nepovinný, viz níže**) `-d DISTRICT_FILE` — pro příkaz `districts` vypisuje místo LAU 1 kódu okresu jeho jméno. Mapování kódů na jména je v souboru `DISTRICT_FILE`
  - (**nepovinný, viz níže**) `-r REGIONS_FILE` — pro příkaz `regions` vypisuje místo NUTS 3 kódu kraje jeho jméno. Mapování kódů na jména je v souboru `REGIONS_FILE`
- `-h` — vypíše nápovědu s krátkým popisem každého příkazu a přepínače.

Popis

1. Skript filtruje záznamy osob s prokázanou nákazou koronavirem způsobujícím onemocnění COVID-19. Pokud je skriptu zadán také příkaz, nad filtrovanými záznamy daný příkaz provede.
2. Pokud skript nedostane ani filtr ani příkaz, opisuje záznamy na standardní výstup.
3. Skript umí zpracovat i záznamy komprimované pomocí nástrojů `gzip` a `bzip2` (v případě, že název souboru končí `.gz` resp. `.bz2`).

4. V případě, že skript na příkazové řádce nedostane soubory se záznamy (`LOG`, `LOG2`, ...), očekává záznamy na standardním vstupu.
5. Pokud má skript vypsat seznam, každá položka je vypsána na jeden řádek a pouze jednou. Není-li uvedeno jinak, je pořadí řádků dáno abecedně. Položky se nesmí opakovat.
6. Grafy jsou vykresleny pomocí ASCII a jsou otočené doprava. Hodnota řádku je vyobrazena posloupností znaku mřížky `#`.

Podrobné požadavky

1. Skript analyzuje záznamy ze zadaných souborů v daném pořadí.

2. Formát souboru se záznamy je CSV, kde oddělovačem je znak čárky `,`. Celý soubor je v kódování ASCII. Formát je řádkový, každý *neprázdný* řádek (prázdné řádky jsou takové, které obsahují jen bílé znaky) odpovídá záznamu o jedné nákaze osoby ve tvaru (následující je hlavička CSV souboru)

   ```
   id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
   ```

   kde

   - `id` je identifikátor záznamu (řetězec neobsahující bílé znaky a znak čárky `,`),
   - `datum` je ve formátu `YYYY-MM-DD`,
   - `vek` je kladné celé číslo,
   - `pohlavi` je `M` (muž) nebo `Z` (žena),
   - `kraj_nuts_kod` je kód kraje, kde byla nákaza zjištěna,
   - `okres_lau_kod` je kód okresu, kde byla nákaza zjištěna,
   - `nakaza_v_zahranici` značí, zda zdroj nákazy byl v zahraničí (`1` značí, že zdroj nákazy byl v zahraničí, prázdné pole značí, že nebyl),
   - `nakaza_zeme_csu_kod` je kód země vzniku nákazy (pro nákazu vzniklou v zahraničí),
   - `reportovano_khs` značí, zda byla nákaza reportována krajské hygienické stanici.

   Příklad souboru se třemi záznamy:

   ```
   id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
   6f4125cb-fb41-4fb0-a478-07b69ba106a4,2020-03-01,21,Z,CZ010,CZ0100,1,IT,1
   f6b08ff5-203d-4a3e-aab0-a5d39ac9ab9e,2020-03-11,32,M,CZ080,CZ0804,,,1
   b03dcf40-04cd-4f7b-a13d-767fc43c3013,2020-03-14,38,M,,,,,
   ```

   - První záznam z 1. března 2020 značí nákazu 21leté ženy v kraji "Hlavní město Praha" (kód `CZ010`), v okrese "Hlavní město Praha" (kód `CZ0100`). Žena byla nakažena v Itálii (kód `IT`) a nákaza byla reportována krajské hygienické stanici.
   - Druhý záznam značí vnitrostátní nákazu 32letého muže v Moravskoslezském kraji (kód `CZ080`), v okrese Nový Jičín (kód `CZ0804`), zjištěnou 11. března 2020.
   - Třetí záznam značí nákazu 38letého muže zjištěnou 14. března 2020, u níž nejsou k dispozici další informace.
3. Skript žádný soubor nemodifikuje. Skript nepoužívá dočasné soubory.

4. Záznamy ve vstupních souborech nemusí být uvedeny chronologicky a je-li na vstupu více souborů, jejich pořadí také nemusí být chronologické.

5. Pokud není při použití přepínače `-s` uvedena šířka `WIDTH`, pak každý výskyt symbolu `#` v grafu odpovídá počtu nákaz (zaokrouhleno dolů) dle příkazu následujícím způsobem:

   - `gender` — 100 000
   - `age` — 10 000
   - `daily` — 500
   - `monthly` — 10 000
   - `yearly` — 100 000
   - `countries` — 100
   - `districts` — 1 000
   - `regions` — 10 000

6. Při použití přepínače `-s` s uvedenou šířkou `WIDTH` je počet nákaz na mřížku upraven podle největšího počtu výskytů v řádku grafu. Řádek s největší hodnotou bude mít počet mřížek `WIDTH` a ostatní řádky budou mít proporcionální počet mřížek vzhledem k největší hodnotě. Při dělení zaokrouhlujte dolů. Tedy např. při `-s 6` a řádku s největší hodnotou 1234 bude řádek s touto hodnotou vypadat takto: `######`.

7. Pořadí argumentů stačí uvažovat takové, že nejřív budou všechny přepínače, pak (volitelně) příkaz a nakonec seznam vstupních souborů (lze tedy použít `getopts`).

8. Předpokládejte, že vstupní soubory nemůžou mít jména odpovídající některému příkazu nebo přepínači.

9. V případě uvedení přepínače `-h` se vždy pouze vypíše nápověda a skript skončí (tedy, pokud by za přepínačem následoval nějaký příkaz nebo soubor, neprovede se).

10. V případě neexistující hodnoty atributu u příkazů `gender`, `age`, `daily`, `monthly`, `yearly`, `countries`, `districts`, `regions` agregujte záznamy s neexistující hodnotou pod hodnotu `None`, kterou ve výpisech uvádějte jako poslední.

11. Přepodkládejte, že je-li v záznamu uvedena hodnota pro nějaký atribut záznamu, pak je korektní (tj. není potřeba validovat vstup) s následujícími výjimkami:

    - ve sloupci `datum` je očekáváno korektní datum ve formátu `YYYY-MM-DD`, které odpovídá skutečnému dni (tedy, např., `yesterday` nebo `2020-02-30` jsou nevalidní hodnoty).
    - ve sloupci `vek` je očekáváno nezáporné celé číslo (tedy, např., `-42`, `18.5` nebo `1e10` jsou nevalidní hodnoty).

    V případě detekování záznamu porušujícího některou z podmínek uvedených výše vypište chybu na chybový výstup a pokračujte ve zpracovávání dále (chybějící hodnota data/věku není porušením). Formát pro chybu je následující:

    ```
    Invalid date: 6f4125cb-fb41-4fb0-a478-07b69ba106a4,2020-04-31,21,Z,CZ010,CZ0100,1,IT,1
    Invalid age: 033fb060-2a10-42ce-80c1-72c2e39b1981,2020-03-05,dvacet,Z,CZ042,CZ0421,,,1
    ```

12. U příkazu `age` pracujte s následujícími intervaly a zarovnáním:

    ```
    0-5   :
    6-15  :
    16-25 :
    26-35 :
    36-45 :
    46-55 :
    56-65 :
    66-75 :
    76-85 :
    86-95 :
    96-105:
    >105  :
    ```

13. Implementace přepínačů `-d` a `-r` je nepovinná; korektní implementace může vynahradit jiné bodové ztráty.

14. U příkazů `gender`, `daily`, `monthly`, `yearly`, `countries`, `districts`, `regions` (bez přepínačů `-d` a `-r`) stačí vypisovat výstup ve formátu `hodnota: pocet` (bez mezery před dvojtečkou a s právě jednou mezerou za dvojtečkou), případně (pro přepínač `-s`) `hodnota: ###...#`. U příkazu `age` je zarovnání specifikováno výše.

    Pro nepovinné přepínače `-r` a `-d` je dvojtečka na pozici o jedna větší, než je počet symbolů nejdelší hodnoty, tj. např.

    ```
    hodnota      : 42
    delsi_hodnota: 1337
    ```

15. Příkaz `countries` nevypisuje počet nákaz v České republice (kód `CZ`).

16. Hodnoty ve sloupcích `nakaza_v_zahranici` a `reportovano_khs` ignorujte (tj. například u příkazu `countries` není třeba brát `nakaza_v_zahranici` do úvahy).

17. Záznamy nemusí nutně mít patřičný počet sloupců. V případě chybějícího sloupce postupujte stejně, jako kdyby v něm chyběla hodnota (pokud záznamu chybí `N` polí, znamená, to, že chybí hodnota `N` nejpravějších sloupců, tedy, např., pokud záznam obsahuje jen 7 polí, pak chybí hodnoty sloupců `nakaza_zeme_csu_kod` a `reportovano_khs`).

18. Nekontrolujte, zda obsahy sloupců `kraj_nuts_kod`, `okres_lau_kod` a `nakaza_zeme_csu_kod` odpovídají daným číselníkům. V případě implementace rozšíření `-d` a `-r` při použití hodnoty nedefinované v souboru s definicemi okresů/krajů vypisujte dané záznamy na chybový výstup v následujícím formátu:

```
Invalid value: 07958a56-6867-4245-b042-29c291c20359,2020-08-16,5,M,CZ099,CZ0999,,,1
```

Implementační detaily

1. Skript by měl mít v celém běhu nastaveno `POSIXLY_CORRECT=yes`.

2. Skript by měl běžet na všech běžných shellech (`dash`, `ksh`, `bash`). Pokud použijete vlastnost specifickou pro nějaký shell, uveďte to pomocí direktivy interpretu na prvním řádku souboru, např. `#!/bin/bash` nebo `#!/usr/bin/env bash` pro `bash`. Můžete použít GNU rozšíření pro `sed` či `awk`. Jazyky Perl, Python, Ruby, atd. povoleny nejsou.

   UPOZORNĚNÍ: některé servery, např. `merlin.fit.vutbr.cz`, mají symlink `/bin/sh -> bash`. Ověřte si proto, že skript skutečně testujete daným shellem. Doporučuji ověřit správnou funkčnost pomocí virtuálního stroje níže.

3. Skript musí běžet na běžně dostupných OS GNU/Linux, BSD a MacOS. Studentům je k dispozici virtuální stroj s obrazem ke stažení zde: http://www.fit.vutbr.cz/~lengal/public/trusty.ova (pro VirtualBox, login: `trusty` / heslo: `trusty`), na kterém lze ověřit správnou funkčnost projektu.

4. Skript nesmí používat dočasné soubory. Povoleny jsou však dočasné soubory nepřímo tvořené jinými příkazy (např. příkazem `sed -i`).

Odevzdání projektu

Odevzdávejte pouze skript `corona` (nebalte ho do žádného archivu). Odevzdejte do IS, termín Projekt 1.

Rady

- Dobrá dekompozice problému na podproblémy Vám může značně ulehčit práci a předejít chybám.
- Naučte se *dobře* používat **funkce** v shellu (uvědomte si, že spousta funkcionality, např. pro výpisy statistik, histogram, atd., je obdobná).

Návratová hodnota

- Skript vrací úspěch v případě úspěšné operace. Interní chyba skriptu nebo chybné argumenty budou doprovázeny chybovým hlášením na stderr a neúspěšným návratovým kódem.

Příklady použití

- Ukázky záznamů o nakažených jsou dostupné na oficiálních stránkách MZČR: https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/osoby.csv (pozor, má cca 250 MiB). Na této stránce jsou dispozici další datové sady včetně popisů jejich schémat.
- Vzorové záznamy, na kterých jsou ukázány příklady použití níže, jsou k dispozici na této stránce. Jsou to konkrétně následující:
- Kopie souboru `osoby.csv` z 21. února 2022 zde (cca 250 MiB).
- Zkrácená verze verze `osoby-short.csv` zde (cca 150 KiB).
- Podmnožina záznamů za leden 2022 rozdělených dle jednotlivých dnů je k dispozici zde.
- Komprimované verze souborů `osoby.csv` a `osoby-short.csv` (`osoby.csv.gz`, `osoby.csv.bz2`, `osoby-short.csv.gz` a `osoby-short.csv.bz2`).
- Soubor `osoby2.csv` s ukázkami těžších záznamů, které je potřeba umět korektně zpracovat, je zde.

Příklady:

```
$ cat osoby.csv | head -n 5 | ./corona
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
6f4125cb-fb41-4fb0-a478-07b69ba106a4,2020-03-01,21,Z,CZ010,CZ0100,1,IT,1
5841443b-7df4-4af9-acab-75ca47010ec3,2020-03-01,43,M,CZ042,CZ0421,1,IT,1
5cdb7ece-97a2-4336-9715-59dc70a48a2c,2020-03-01,67,M,CZ010,CZ0100,1,IT,1
d345e0e2-9056-4d3f-b790-485b12831180,2020-03-03,21,Z,CZ010,CZ0100,,,
```

```
$ ./corona infected osoby.csv
3510360
```

```
$ ./corona infected infected-jan22/infected-22-01-*.csv
560894
```

```
$ ./corona merge infected-jan22/infected-22-01-*.csv
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
741d72a4-2b6e-4703-872d-928748ca0ade,2022-01-01,3,Z,CZ020,CZ0203,,,1
f39754b8-5e7f-44fd-8b65-e4e7e3b89521,2022-01-01,52,Z,CZ052,CZ0522,,,1
...
1a27f58f-8950-40c5-89fa-3795f4a906f4,2022-01-31,19,Z,CZ063,CZ0635,,,1
9aebc069-89d5-4ba0-96c5-aefa1f2c6746,2022-01-31,19,M,CZ064,CZ0642,,,1
```

```
$ cat osoby.csv | ./corona gender
M: 1703679
Z: 1806681
```

```
$ curl -s 'https://pajda.fit.vutbr.cz/ios/ios-22-1-inputs/-/raw/main/data/osoby.csv' | ./corona -a 2021-07-19
infected
1835517
```

```
$ cat osoby.csv | ./corona daily
2020-03-01: 3
2020-03-02: 0
2020-03-03: 2
2020-03-04: 1
...
2022-02-19: 8218
2022-02-20: 4267
```

```
$ cat osoby.csv | ./corona monthly
2020-03: 3316
2020-04: 4385
2020-05: 1615
...
2022-01: 560894
2022-02: 465810
```

```
$ cat osoby.csv | ./corona yearly
2020: 732808
2021: 1750848
2022: 1026704
```

```
$ ./corona countries osoby.csv
99: 1
AD: 1
AE: 444
AF: 13
...
ZA: 36
ZM: 2
ZW: 1
```

(kód země `99` na prvním řádku je chyba v datové sadě; neřešte ji)

```
$ ./corona -g M osoby.csv | head -n 6
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
5841443b-7df4-4af9-acab-75ca47010ec3,2020-03-01,43,M,CZ042,CZ0421,1,IT,1
5cdb7ece-97a2-4336-9715-59dc70a48a2c,2020-03-01,67,M,CZ010,CZ0100,1,IT,1
496a049f-656e-4274-a51f-72aa92d01f33,2020-03-05,49,M,CZ042,CZ0421,1,IT,1
815a2219-2735-46ae-8b14-658459481b2f,2020-03-06,47,M,CZ010,CZ0100,1,IT,1
9f78dd0d-2e71-4d37-89a2-665b44b2a607,2020-03-06,44,M,CZ010,CZ0100,1,IT,1
```

```
$ cat /dev/null | ./corona
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
```

```
$ ./corona -s daily osoby.csv
2020-03-01:
2020-03-02:
2020-03-03:
2020-03-04:
...
2022-02-19: ################################################################################
2022-02-20: ##########################################
```

```
$ ./corona -s monthly osoby.csv
2020-03:
2020-04:
2020-05:
...
2022-01: ############################################################
2022-02: ##############################################
```

```
$ ./corona -s 20 yearly osoby.csv
2020: ########
2021: ####################
2022: ###########
```

```
$ cat osoby.csv.gz | ./corona | head -n 5
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
6f4125cb-fb41-4fb0-a478-07b69ba106a4,2020-03-01,21,Z,CZ010,CZ0100,1,IT,1
5841443b-7df4-4af9-acab-75ca47010ec3,2020-03-01,43,M,CZ042,CZ0421,1,IT,1
5cdb7ece-97a2-4336-9715-59dc70a48a2c,2020-03-01,67,M,CZ010,CZ0100,1,IT,1
d345e0e2-9056-4d3f-b790-485b12831180,2020-03-03,21,Z,CZ010,CZ0100,,,
```

```
$ cat osoby.csv.bz2 | ./corona | head -n 5
id,datum,vek,pohlavi,kraj_nuts_kod,okres_lau_kod,nakaza_v_zahranici,nakaza_zeme_csu_kod,reportovano_khs
6f4125cb-fb41-4fb0-a478-07b69ba106a4,2020-03-01,21,Z,CZ010,CZ0100,1,IT,1
5841443b-7df4-4af9-acab-75ca47010ec3,2020-03-01,43,M,CZ042,CZ0421,1,IT,1
5cdb7ece-97a2-4336-9715-59dc70a48a2c,2020-03-01,67,M,CZ010,CZ0100,1,IT,1
d345e0e2-9056-4d3f-b790-485b12831180,2020-03-03,21,Z,CZ010,CZ0100,,,
```

```
$ ./corona districts osoby.csv
CZ0100: 448252
CZ0201: 34423
CZ0202: 33545
CZ0203: 54368
CZ0204: 36166
...
CZ0806: 103556
None: 2959
```

```
$ ./corona regions osoby.csv
CZ010: 448252
CZ020: 482138
...
CZ080: 387509
None: 2926
```

```
$ ./corona age osoby.csv
0-5   : 118107
6-15  : 511868
16-25 : 410980
26-35 : 511672
36-45 : 649751
46-55 : 570064
56-65 : 359275
66-75 : 225485
76-85 : 110360
86-95 : 39405
96-105: 2651
>105  : 302
None  : 440
```

```
$ ./corona infected osoby2.csv
9
Invalid date: 0dc57759-d153-45c2-8d14-fb92fc028060,2020-15-03,62,Z,CZ010,CZ0100,,,1
Invalid age: 5b0a9692-a72a-4f34-a014-83ae08a79f20,2020-03-10,3.1415,Z,CZ071,CZ0712,1,IT,1
```

```
$ ./corona daily osoby2.csv
2020-03-01: 3
2020-03-03: 2
2020-03-04: 1
2020-03-05: 3
Invalid date: 0dc57759-d153-45c2-8d14-fb92fc028060,2020-15-03,62,Z,CZ010,CZ0100,,,1
Invalid age: 5b0a9692-a72a-4f34-a014-83ae08a79f20,2020-03-10,3.1415,Z,CZ071,CZ0712,1,IT,1
```

Rozšíření

```
$ ./corona -d okresy.csv districts osoby.csv
Benesov          : 34423
Beroun           : 33545
Blansko          : 34374
Brno-mesto       : 123692
...
Zdar nad Sazavou: 37928
None             : 2959
```

```
$ ./corona -r kraje.csv regions osoby.csv
Hlavni mesto Praha  : 448252
Jihocesky kraj      : 206288
Jihomoravsky kraj   : 374972
Karlovarsky kraj    : 77709
Kraj Vysocina       : 158169
Kralovehradecky kraj: 190181
Liberecky kraj      : 148988
Moravskoslezsky kraj: 387509
Olomoucky kraj      : 208593
Pardubicky kraj     : 183208
Plzensky kraj       : 193696
Stredocesky kraj    : 482138
Ustecky kraj        : 248821
Zlinsky kraj        : 198910
None                : 2926
```