Vysoké učení technické v Brně

Fakulta informačních technologií

Projekt 1. časť

Návrh spracovania a uloženia dát

Ak.rok: 2020/2021

Ročník: 1MIT

Predmet: Ukládání a příprava dat

Zodpovedný za projekt: RNDr. Marek Rychlý, Ph.D.

Zvolená téma: Covid-19

Riešitelia: Bettina Pinkeová (xpinke00), Michal Kabáč (xkabac00), Daniel Kavuliak (xkavul01)

Zvolené dotazy a formulácia vlastného dotazu:

- **Dotaz zo skupiny A:**
  - vytvořte popisné charakteristiky pro alespoň 4 údaje (např. věk, pohlaví, okres, zdroj nákazy) z datové sady COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (využijte krabicové grafy, histogramy, atd.)
- **Dotaz zo skupiny B:**
  - najděte skupiny nemocných s podobnými kritérii (např. podobný věk, místo, čas testů, atp.) a určete jejich vývoj v čase
- **Vlastný dotaz:**
  - porovnanie vývoja COVID-19 v Českej republike s vývojom v Európskej únii

Stručná charakteristika zvolenej dátovej sady:

Na vypracovanie dotazov zo skupiny A a skupiny B sme si zvolili dátové sady z nasledujúceho zdroja:

https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19

Na tomto odkaze sú dostupné viaceré dátové sady, my sme si vybrali konkrétne:

1. **COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (v2)**

Táto dátová sada obsahuje 184 123 záznamov a 7 stĺpcov.

- Prvý stĺpec predstavuje dátum, je typu object
- Druhý stĺpec predstavuje vek, je typu integer
- Tretí stĺpec predstavuje pohlavie, je typu object
- Štvrtý stĺpec predstavuje "kraj NUTS kód", ktorý je typu object
- Piaty stĺpec predstavuje "okres Lau kod", ktorý je typu object
- Šiesty stĺpec má názov "nákaza v zahraničí", ktorý je typu float
- Siedmy stĺpec má názov "nákaza zeme ČSU kód", ktorý je typu object

Každý riadok predstavuje osobu, pri ktorej bol preukázaný vírus COVID-19. Prvý atribút obsahuje dátum hlásenia preukázania vírusu u danej osoby. Ďalšie dva atribúty popisujú vek a pohlavie nakazenej osoby. Štvrtý a piaty atribút hovorí o tom, v ktorom okrese a v ktorom kraji bol diagnostikovaný COVID-19. Posledné dva atribúty popisujú, či sa osoba nakazila v zahraničí, a keď sa nakazila, tak v ktorom štáte.

datum	vek	pohlavi	kraj_nuts_kod	okres_lau_kod	nakaza_v_zahranici	nakaza_zeme_csu_kod
2020-04-03	33	M	CZ072	CZ0722	1.0	PA
2020-09-23	19	M	CZ031	CZ0317	1.0	HR
2020-07-18	50	Z	CZ010	CZ0100	1.0	HR
2020-08-04	23	Z	CZ010	CZ0100	1.0	HR
2020-09-10	16	Z	CZ031	CZ0311	1.0	SK

Obr. 1 - Ukážka dát z dátového súboru číslo 1.

2. COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu

Táto dátová sada obsahuje 1031 záznamov a 5 stĺpcov.

- Prvý stĺpec predstavuje dátum, je typu object
- Druhý stĺpec predstavuje "kraj NUTS kód" , ktorý je typu object
- Tretí stĺpec predstavuje "okres Lau kód" , ktorý je typu object
- Štvrtý stĺpec predstavuje kumulatívny počet nakazených, ktorý je typu integer
- Piaty stĺpec predstavuje kumulatívny počet vyliečených, ktorý je typu integer

- Šiesty stĺpec predstavuje kumulatívny počet úmrtí, ktorý je typu integer

datum	kraj_nuts_kod	okres_lau_kod	kumulativni_pocet_nakazenych	kumulativni_pocet_vylecenych	kumulativni_pocet_umrti
2020-03-01	CZ010	CZ0100	2	0	0
2020-03-01	CZ020	CZ020A	0	0	0
2020-03-01	CZ020	CZ020B	0	0	0
2020-03-01	CZ020	CZ020C	0	0	0
2020-03-01	CZ020	CZ0201	0	0	0

Obr. 2 - Ukážka dát z dátového súboru číslo 2.

Dátová sada obsahuje záznamy, ktoré popisujú denné kumulatívne počty nakazených, vyliečených a úmrtí osôb. Prvý atribút popisuje dátum hlásenia o kumulatívnych počtoch. Ďalej nasleduje atribúty popisujúce kód kraja v ČR a kód okresu v danom kraji. Ostatné atribúty obsahujú kumulatívny počet nakazených, vyliečených a úmrtí.

3. COVID-19: Přehled vyléčených dle hlášení krajských hygienických stanic

Táto dátová sada obsahuje 9898 záznamov a 5 stĺpcov.

- Prvý stĺpec predstavuje dátum, je typu object
- Druhý stĺpec predstavuje vek, je typu integer
- Tretí stĺpec predstavuje pohlavie, je typu object
- Štvrtý stĺpec predstavuje “kraj NUTS kód” , ktorý je typu object
- Piaty stĺpec predstavuje “okres Lau kód” , ktorý je typu object

datum	vek	pohlavi	kraj_nuts_kod	okres_lau_kod
2020-03-14	39	Z	CZ020	CZ020A
2020-03-15	49	M	CZ042	CZ0421
2020-03-15	18	Z	CZ042	CZ0421
2020-03-15	53	Z	CZ042	CZ0421
2020-03-15	20	Z	CZ010	CZ0100

Obr. 3 - Ukážka dát z dátového súboru číslo 3.

Táto dátová sada obsahuje záznamy, ktoré popisujú vyliečené osoby. Atribút dátumu obsahuje dátum, kedy bola daná osoba vyliečená. Nasledujú atribúty, ktoré popisujú vek a pohlavie osoby. Posledné atribúty obsahujú kód kraja v ČR a kód okresu v danom kraji.

4. **COVID-19: Přehled úmrtí dle hlášení krajských hygienických stanic**

Táto dátová sada obsahuje 17 479 záznamov a 6 stĺpcov.

- Prvý stĺpec predstavuje dátum, je typu object
- Druhý stĺpec predstavuje vek, je typu integer
- Tretí stĺpec predstavuje pohlavie, je typu object
- Štvrtý stĺpec predstavuje “kraj NUTS kód” , ktorý je typu object
- Piaty stĺpec predstavuje “okres Lau kód” , ktorý je typu object

datum	vek	pohlavi	kraj_nuts_kod	okres_lau_kod
2020-03-22	94	M	CZ010	CZ0100
2020-03-24	73	Z	CZ010	CZ0100
2020-03-24	44	M	CZ080	CZ0802
2020-03-25	91	Z	CZ010	CZ0100
2020-03-25	79	M	CZ053	CZ0532

Obr. 4 - Ukážka dát z dátového súboru číslo 4.

Tento dataset popisuje osoby, ktoré zomreli s vírusom COVID-19. Obsahom dátovej sady sú atribúty, ktoré popisujú dátum úmrtia, vek a pohlavie osoby, kód kraja v ČR a kód okresu v danom kraji.

V našom projekte sa chceme zamerať na zobrazenie COVID situácie v jednotlivých krajoch Českej republiky, z toho dôvodu plánujeme namapovať všetky vyššie uvedené záznamy na jednotlivé kraje. Pre každý kraj budeme mať príslušné štatistiky týkajúce sa počtu nakazených, mŕtvych, vyliečených a testovaných pre jednotlivé dni.

Dátové sady sú na stránke dostupné vo formátoch JSON a CSV. Preto sme sa rozhodli stiahnuť dátovú sadu vo formáte CSV a následne sa pozrieť na chýbajúce hodnoty.

```
datum                   0
vek                     0
pohlavi                 0
kraj_nuts_kod           0
okres_lau_kod           0
nakaza_v_zahranici    180184
nakaza_zeme_csu_kod   180184
dtype: int64
```

Obr. 5 - Chýbajúce hodnoty dátovej sady číslo 1.

```
datum                          0
kraj_nuts_kod                  0
okres_lau_kod                  0
kumulativni_pocet_nakazenych   0
kumulativni_pocet_vylecenych   0
kumulativni_pocet_umrti        0
dtype: int64
```

Obr. 6 - Chýbajúce hodnoty dátovej sady číslo 2.

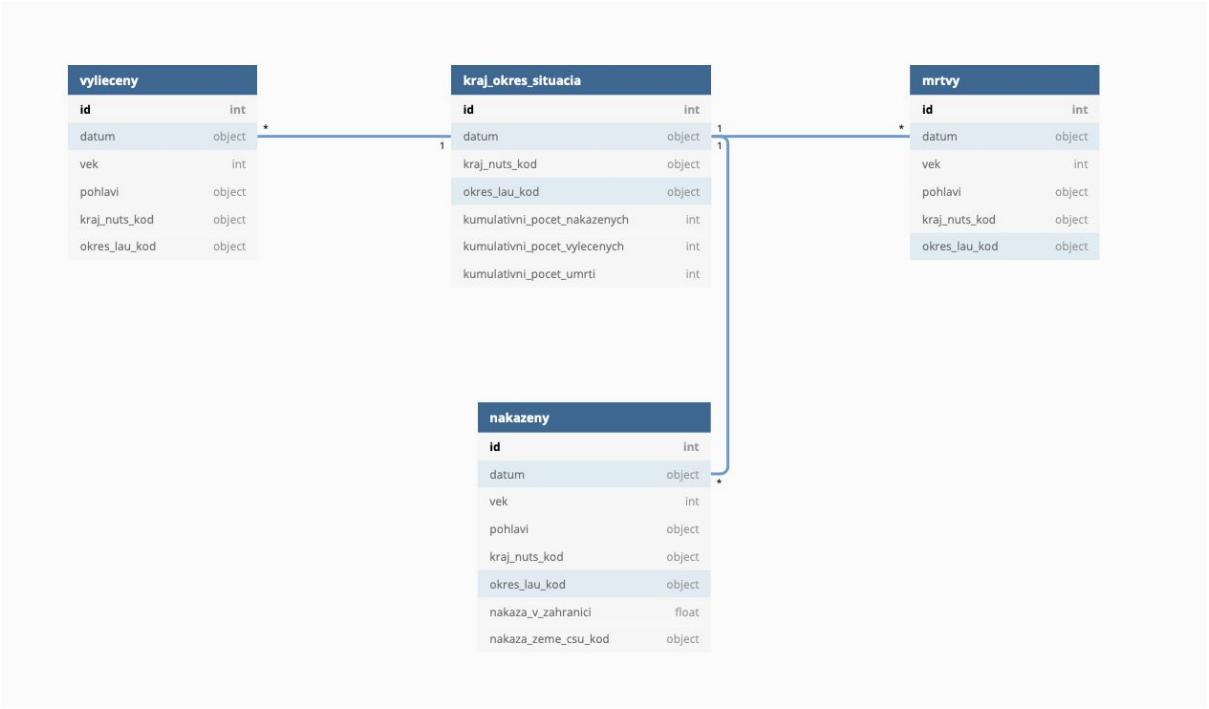
```
datum          0
vek            0
pohlavi        0
kraj_nuts_kod  0
okres_lau_kod  0
dtype: int64
```

Obr. 7 - Chýbajúce hodnoty dátovej sady číslo 3.

```
datum          0
vek            0
pohlavi        0
kraj_nuts_kod  0
okres_lau_kod  0
dtype: int64
```

Obr. 8 - Chýbajúce hodnoty dátovej sady číslo 4.

Po preskúmaní chýbajúcich hodnôt sme dospeli k záveru, že dáta chýbajú iba pri atribútoch (“nakaza_v_zahranici” a “nakaza_zeme_csu_kod” viď obr. 1), ktoré v našom prípade nie sú významné, nakoľko ich nepotrebujeme pre naplnenie zvolených dotazov.



Obr. 8 - Schéma zvolených dátových sád

Na obrázku vidíme schému zvolených dátových sád a ich prepojenia medzi sebou. Identifikovali sme nasledovné 4 entity:

1. “vylieceny”
2. “kraj_okres_situacia”
3. “nakazeny”
4. “mrtvy”

“Datum”, “kraj_nuts_kod” a “okres_lau_kod” sú spoločné pre všetky entity a na základe nich ich budeme spájať. Tieto identifikátory nie sú dostatočne jednoznačné, ale v našom prípade sú postačujúce, pretože sa zameriavame na zoskupenie všetkých mŕtvych, nakazených a vyliečených pod jeden kraj. Nakazený, mŕtvy alebo vyliečený patrí do práve jedného okresu, ale okres môže mať viacero vyššie uvedených.

Pre realizáciu nášho vlastného dotazu plánujeme použiť dátovú sadu zo zdroja: **https://documenter.getpostman.com/view/10877427/SzYW2f8n?version=latest#495ae7db-58bc-4dfd-9080-c234b080291b**

Dáta z tohto zdroja môžeme získať zavolaním get requestu na daný endpoint. Response zo servera je vo formáte JSON a obsahuje atribúty:

- Prvý stĺpec predstavuje “Country” (krajinu), je typu object.
- Druhý stĺpec predstavuje “last_update” (posledný update), je typu date
- Tretí stĺpec predstavuje “cases” (počet prípadov), je typu integer
- Štvrtý stĺpec predstavuje “deaths” (počet úmrtí) , ktorý je typu integer
- Piaty stĺpec predstavuje “recovered” (počet vyliečených) , ktorý je typu integer

Uvedené dáta plánujeme získať pomocou skriptu v jazyku Python. Následne ich plánujeme spárovať s dátami získanými zo zdroja vyššie na základe dátumu.

Zvolený spôsob uloženia surových dát:

Na uloženie dát plánujeme použiť databázu MongoDB, pomocou ktorej budeme mať dáta uložené vo formáte JSON. Objekty v databáze budú jednotlivé kraje, tieto objekty obsahujú atribúty, ktoré sú uvedené v entite “kraj_okres_situacia” v schéme (viď Obr. 8). Okrem týchto atribútov bude obsahovať aj polia objektov vyliečených, mŕtvych a nakazených osôb z daného dňa. Príklad reprezentácie môžeme vidieť z kód 1. Po predspracovaní dáta vo filnálnej forme plánujeme presunúť z databázy MongoDB do relačnej databázy PostregSQL.

```
{
     "datum": 2020-06-06,
     "kraj": "neviem",
     "okres": "neviem2",
     "kumulativni_pocet_nakazenych": 2,
     "kumulativni_pocet_mrtvych": 0,
     "kumulativni_pocet_vylecenych": 0,
     "nakazeny": [
          {
               "datum": 2020-06-06,
               "vek": 90,
               "pohlavie": "M",
               "kraj": "neviem",
               "okres": "neviem2"
          },
          {
               "datum": 2020-06-06,
               "vek": 84,
               "pohlavie": "F",
               "kraj": "neviem",
               "okres": "neviem2"
          }
     ],
     "mrtvy": [],
     "vyleceny": []
}
```

Kód 1 - predstava ukážky uloženia objektu v databáze MongoDB

Návrh spracovania dotazov

Našim cieľom je vytvoriť desktopovú aplikáciu v programovacom jazyku Python s využitím frameworku Flask, ktorej účelom bude zobrazenie štatistických informácií o priebehu ochorenia COVID-19 pre každý kraj v rámci ČR. Hlavnou súčasťou bude vizualizácia Českej republiky prostredníctvom geografickej mapy rozdelenej podľa krajov. Táto vizualizácia bude interaktívna, mapa obsahuje rozšírenie v podobe štatistík jednotlivých krajov ČR.

Dotaz skupiny A

Tento dotaz plánujeme realizovať prostredníctvom spomínanej vizualizácie prostredníctvom mapy ČR pre jednotlivé kraje. Pre vykonanie tohto dotazu použijeme prvé 4 dátové sady spomenuté vyššie. Dátové sady budú spracované do deskriptívnych štatistík, grafov - boxplotov, histogramov, scatterplotov… V štatistikách budú zahrnuté denné prírastky nakazených, vyliečených a mŕtvych. Vizualizáciu týchto štatistík chceme spracovať pre jednotlivé kraje. Tieto údaje by mali byť zobrazené na hlavnej stránke po vstupe do systému. Detailnejšie zobrazenie bude možné zobraziť po kliknutí na príslušné tlačidlo.

Dotaz skupiny B

Tento dotaz plánujeme takisto realizovať prostredníctvom spomínanej vizualizácie prostredníctvom mapy ČR pre jednotlivé kraje. Pre tento dotaz použijeme štyri dátové sady, ktoré sú spomenuté vyššie. Spracovanie tohto dotazu bude súčasťou aplikačnej funkcie, kde je implementovaný dotaz A. Po zobrazení deskriptívnych štatistík, používateľ klikne na tlačidlo a po jeho kliknutí sa zobrazí okno s lineplotmi. Tieto grafy popisujú vývoj počtov nakazených, vyliečených a mŕtvych v čase. V príslušnom textovom poli si môže používateľ zadať dátum od ktorého chce určiť vývoj v čase.

Vlastný dotaz

Tento dotaz neplánujeme realizovať prostredníctvom spomínanej vizualizácie mapou, ale bude implementovaný v samostatnej funkcionalite. Pre naplnenie vlastného dotazu plánujeme použiť všetky uvedené dátové sady. V tomto dotaze plánujeme porovnať šírenie COVID-19 v Českej republike so šírením COVID-19 v Európskej únii. Porovnanie chceme zrealizovať

ako štatistické - počet nakazených, mŕtvych a vyliečených. Následne toto porovnanie plánujeme vizualizovať v príslušných grafoch.