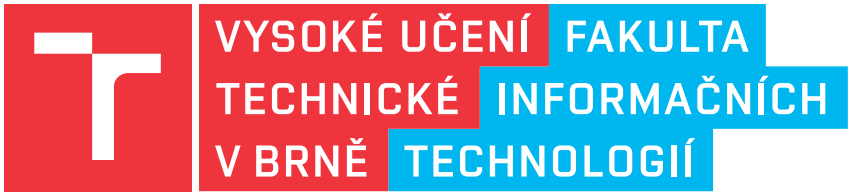

Projektová dokumentace

Varianta 3: COVID-19

Ukládání a příprava dat

Šimon Galba, Bc. (xgalba03)
Gladiš Damián, Bc. (xgladi00)
Jeřábek František, Bc. (xjerab25)

3. listopadu 2021

Obsah

1 Úvod

Cieľom projektu je stiahnuť dáta štatistík o rôznych skutočnostiach súvisiacich s pandémiou COVID-19 za účelom ich ďalšieho spracovania a zobrazovania do grafov. Jedná sa prevažne o dáta zobrazujúce počty nakazených, vyliečených, novo hospitalizovaných, zaočkovaných, alebo počte vykonaných testov.
Tieto dáta sú spracovávané jazykom Python verzie 3.8 a ukladané do NON-SQL databázy MongoDB.

2 Spustenie

Keďže sa jedná o programovací jazyk Python, program sa bez prekladu spustí príkazom **python3 download.py**

3 Využité knižnice

- datetime - zmena formátu dátumu pre uloženie do databázy
- urlopen - sťahovanie dát z web-u
- ijson - stream-like parsovanie dát, využité na rýchlejšie zmeny dátumu dát
- pymongo.collection - prístup do databáze MongoDB
- MongoClient - prístup do databáze MongoDB
- Pandas - spracovávanie dát v súbore typu csv

4 Využité dáta

Pre účely vytvorenia prehľadov a grafov v druhej časti tohto projektu sú stiahnuté a spracované nasledujúce datasety (kliknutie na kategóriu odkazuje na daný súbor), všetky dáta sú čerpané zo zdroja mzcr:

- Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů
- Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic
- Přehled vykázaných očkování podle krajů ČR
- Přehled vykázaných očkování podle profesí (očkovací místo, bydliště očkovaného)
- Prehľad celkovej populácie

5 Návrh a implementácia

V hlavnom programe sa postupne vykonávajú funkcie na základe jednotlivých otázok na dáta. V jednotlivých funkciách vždy najprv prebehne stiahnutie dát a následná úprava formátu dátumu pre konzistentnosť. Potom sú dáta v dávkach pridávané do databázy.

6 Parsovanie dát

Dáta na naplnenie databázy sú primárne sťahované v súboroch typu Json a .csv. Pre spracúvanie objebných dát bolo do implementácie pridané postupné spracúvanie, tz. streaming. V tomto prípade sa dáta sťahujú a spracúvajú v menších dávkach a samotné akcie na seba nečakajú. Urýchľuje sa tak proces plnenia databázy.