Dokumentácia k prvému odovzdávaniu projektu z UPA

Názov tímu: xryban00

Členovia tímu: Dávid Pukanec, Adam Rybanský, Michal Sandanus

Github repozitár: https://github.com/adaryb123/UPAproject

Zvolená téma: zdravotníctvo

Analýza datasetov:

K dispozícii máme dva datasety, 130142-21data043021.csv a export-sluzby-2021-10.

Súbor 130142-21data043021.csv obsahuje údaje o počtoch obyvateľov v územiach Českej republiky. Základnými údajmi sú:

* „hodnota“ = počet obyvateľov danej skupiny
* pohlavie danej skupiny
* vekové rozpätie skupiny
* rok merania
* územie, v ktorom sa meralo (okresy, kraje, Česká republika, Hlavné mesto)

Riadok tabuľky môžeme interpretovať ako počet obyvateľov konkrétnej skupiny (podľa veku a pohlavia), nameraný v konkrétnom roku pre konkrétny okres/ kraj / Českú republiku/ Hlavné mesto Praha.

Súbor export-sluzby-2021-10 obsahuje údaje o zdravotníckych zariadeniach v ČR. Základnými údajmi sú

* názov zariadenia
* druh zariadenia (asi 40 druhov)
* adresa zariadenia (ulica,číslo domu, PSČ, obec, okres, kraj)
* údaje o poskytovateľovi (názov, druh poskytovateľa, typ osoby (a pre právne osoby aj právnu formu), IČO, email, webstránku, fax, telefónne číslo a adresu poskytovateľa (rovnako rozdelenú na ulicu,číslo domu, PSČ, obec, okres, kraj)
* druh a formu zdravotnej starostlivosti/péče
* meno odborného zástupcu (ak zdravotné zariadenie nejakého má)
* GPS súradnice zariadenia

Záver:

Dátové sady môžeme zlúčiť podľa okresov, pretože to je údaj, ktorý majú obe sady spoločné. Dátové sady obsahujú veľké množstvo NaN hodnôt. Hodnoty v export-sluzby-2021-10 sú prevažne typu string, zatiaľ čo v 130142-21data043021.csv sú prevažne typu integer a float, a pre každý stĺpec s kategóriami je tu aj stĺpec s kódovanými hodnotami kategórii.

Riešenie projektu:

Na implementáciu sme zvolili datábazu MongoDB a jazyk Python. Databáza beží na cloudovom serveri <https://cloud.mongodb.com/> . Na komunikáciu s databázou používame knižnicu Mongoengine. Projekt je štrukturovaný nasledovne:

* Modul database\_driver.py sprostredkuje pripojenie na cloud
* Modul Region.py sprostredkuje uloženie .csv dát do json objektov definovaných ako Region, a následný import týchto objektov do databázy
* Program sa spúšťa súborom main.py, ktorý volá funkcie z modulov
* Jupyter notebook Data\_analysis.ipynb obsahuje postup, akým sme analyzovali dáta pomocou knižnice pandas