Dokumentácia – 1. Časť projektu

Ukladanie a príprava dát

Názov tímu: xryban00

Členovia tímu: Dávid Pukanec, Adam Rybanský, Michal Sandanus

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

Zvolená téma: Zdravotníctvo

Zadanie

1. **Seznamte se s nabízenými tématy** pro řešení projektů, které popisují zdroje dat (odkazy na datové sady) a dotazy, které je potřeba na základě těchto dat zodpovědět.
2. Zvolte si jeden z nabízených zdrojů dat (téma projektu) a **analyzujte dílčí datové sady** z daného zdroje, jejich strukturu (schéma), typy datových položek, identifikátory, možnosti propojení datových sad (společné entity) či napojení na externí data (jiné zdroje, entity reálného světa, např. polohu v geografickém prostoru), změnu dat v čase (temporální data), a jiné. Po dohodě se cvičícími je zde v některých případech možné zvolit vlastní zdroj dat.
3. Navrhněte **vhodný způsob načtení datových sad z daného zdroje a jejich uložení ve zvolené NoSQL databázi** (s výběrem vhodné NoSQL databáze Vám pomohou cvičení z předmětu; jedná se často pro nestrukturovaná či velká data).
4. **Implementujte systém** pro získávání, předzpracování, uložení dat do/ve zvolené NoSQL databázi
5. **Výsledné řešení 1. části odevzdejte** do WISu. Očekává se, že odevzdáte nejen dokumentaci z analýzy datových sad a z návrhu řešení pro načtení a uložení v NoSQL, ale také zdrojové kódy řešení (skripty, aplikace, schémata, atp.) a stručnou dokumentaci jeho zprovoznění, případně také příklady zajímavých problémů, se kterými jste se v průběhu řešení setkali. Dokumentaci (mimo zdrojové kódy) odevzdávejte ve formátu PDF či jako prostý text v kódování ASCII/UTF-8 (např. jako Markdown).

Analýza dátových sád

K dispozícii máme dva datasety:

1. Zo stránky https://nrpzs.uzis.cz/index.php?pg=home--download&archiv=sluzby stiahneme súbor export-sluzby-2021-10.csv, ďalej spomínaný ako dátová sada „Zariadenia“
2. Zo stránky https://www.czso.cz/csu/czso/obyvatelstvo-podle-petiletych-vekovych-skupin-a-pohlavi-v-krajich-a-okresech stiahneme súbor 130142-21data043021.csv, ďalej uvádzaný ako dátová sada „Demografia“

Analýza dátových sád bola vykonaná a popísaná v prostredí Jupyter notebook (Data_analysis.ipynb) s použitím Python 3.8.

Dátová sada Zariadenia

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

- ID zariadenia
- názov zariadenia
- druh zariadenia (asi 40 druhov)
- adresa zariadenia (ulica, číslo, 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)
- obor, druh a formu zdravotnej starostlivosti/péče
- meno odborného zástupcu (ak zdravotné zariadenie nejakého má)
- GPS súradnice zariadenia

Dátová sada obsahuje duplikátne riadky, ktoré je potrebné odstrániť. Pre dôležité stĺpce, ako napríklad „OborPece“ sme sa rozhodli nedefinované hodnoty nahradiť slovom „neznamy“. Dáta sú prevažne textového typu.

Dataset bez duplikátov obsahuje 60 962 riadkov. Stĺpec „ZdravotnickeZarizeniId“ nie je jednoznačným identifikátorom každého riadku. Ten získame kombináciou stĺpcov „ZdravotnickeZarizeniId“, „OborPece“, „FormaPece“, „DruhPece“. Pre vkladanie do nami zvolenej databázy budeme teda používať nami vytvorený identifikátor riadku.

Ďalej analyzujeme stĺpce potrebné pre ďalšie pokračovanie projektu, a to najmä pre stĺpce „OborPece“, „FormaPece“, „DruhPece“ a „DatumZahajeniCinnosti“. Pre tieto stĺpce zisťujeme najmä počty NaN hodnôt a počty rozličných hodnôt. Výsledky sú uvedené v Jupyter notebooku.

Dátová sada Demografia

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)

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

Dátová sada je pomerne čistá, neobsahuje duplikátne riadky, ani NaN hodnoty, ktoré by boli nevyhnutné pre ďalšiu prácu s datasetom. Dataset obsahuje 63 063 riadkov. Dáta sú prevažne číselného typu.

Obyvatelia sú rozdelení vekovo do skupín po 5-ročných intervaloch.

Zlúčenie dátových sád

Dátové sady môžeme zlúčiť na základe názvu okresu, pretože to je údaj, ktorý majú obe sady spoločné. Kódy okresov sa v jednotlivých datasetoch nezhodujú, textové označenie okresu je však rovnaké, a teda datasety spájame na základe stĺpcov „Okres“ v datasete Zariadení a „vuzemi_txt“ v datasete Demografie.

Riešenie projektu

Na implementáciu sme zvolili databázu MongoDB a jazyk Python 3.8. Databáza beží na cloudovom serveri https://cloud.mongodb.com/ . Na komunikáciu s databázou používame knižnicu Mongoengine. Projekt je štruktúrovaný nasledovne:

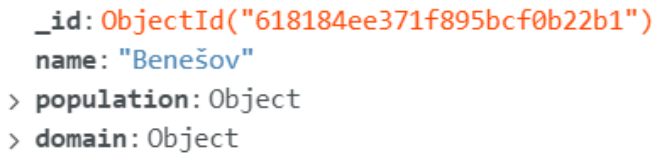
- Modul database_driver.py sprostredkuje pripojenie na cloud
- Modul Region.py predstavuje základnú štrukturálnu jednotku dát v databáze
- Program sa spúšťa súborom main.py, ktorý importuje csv súbory do databázy
- Jupyter notebook Data_analysis.ipynb obsahuje postup, akým sme analyzovali dáta pomocou knižnice pandas s využitím DataFramov

Štruktúra dát v databáze

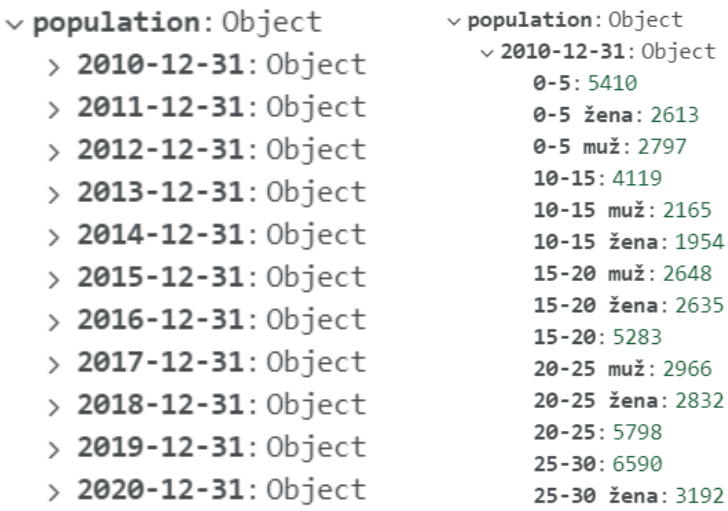
Datasety sú prepojené na základe regiónu, čo je hlavná položka v databáze. Región obsahuje atribút meno a taktiež informácie o populácii a zariadeniach roztriedených podľa oboru zdravotnej starostlivosti.

Populácia je ďalej rozvetvená na roky merania a každý rok merania obsahuje intervalové rozpätia po 5 rokov s hodnotami počtu obyvateľov v danom roku a danom vekovom intervale.

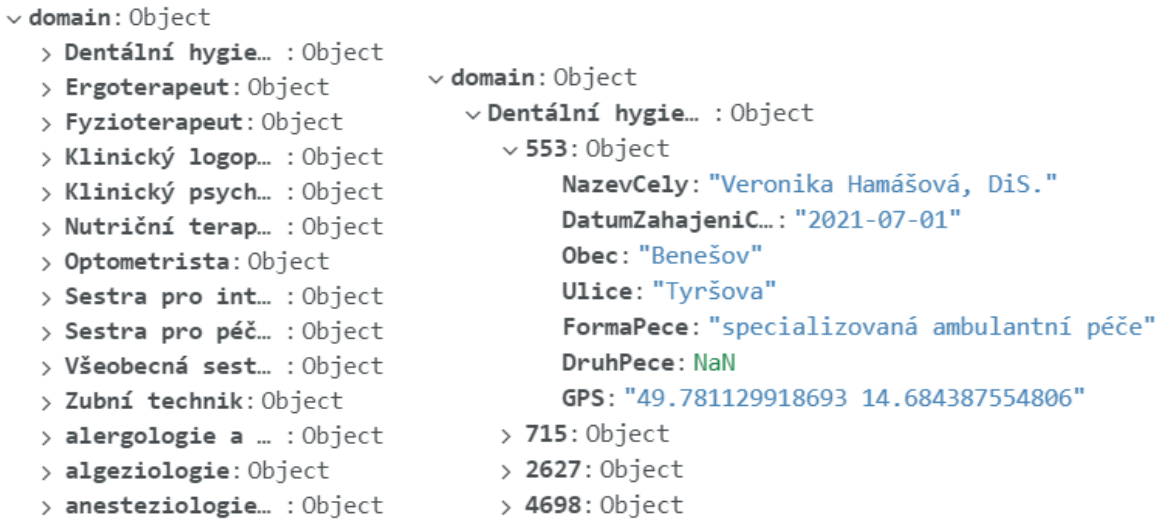
Pre každý región evidujeme iba obory, ktoré obsahujú aspoň jedno zariadenie.
O zariadeniach si ukladáme informácie obsiahnuté v stĺpcoch „NazevCely“, „DatumZahajeniCinnosti“, „Obec“, „Ulice“, „FormaPece“, „DruhPece“, „GPS“.



Základné členenie pre región



Členenie pre populáciu



Členenie pre obor zdravotnej starostlivosti

Spojazdnenie projektu na cudzom PC

Pre spustenie programu je potrebné urobiť nasledovné:

- Stiahnuť si Github repozitár
- Mať nainštalovaný Python, ideálne verziu 3.8 alebo vyššiu
- Importovať knižnice uvedené v requirements.txt (pandas, numpy, jupyter, mongoengine, certifi)
- spustiť súbor main.py

Pre spustenie programu s iným vstupným súborom dát, je potrebné zmeniť hodnotu premenných *csv_file_path_population* a *csv_file_path_facilities*, v main.py. Je v nich uložená cesta k .csv súborom s dátami.

Pripojiť sa na cloudovú databázu má povolené každé zariadenie, avšak ak je potrebné aj prezerať databázu a údaje v nej, musí mať používateľ vytvorený účet na www.cloud.mongodb.com, a musí byť pozvaný do zdieľaného projektu.