Vypracovanie úkolu na pozíciu dátovej špecialistky

Lucia Koštialová

24. augusta 2025

1 Automatické sťahovanie dát

Daná dátová sada obsahuje zoznam registrovaných knižníc v ČR. Na začiatok je potrebné zaistiť automatické sťahovanie a uloženie dát. Na túto časť práce s dátami som využila Jupyter Notebook v prostredí Google Colab. Implementácia bola realizovaná v jazyku Python.

Predpokladala som, že dáta sa dajú dohľadať na adrese: *https://mk.gov.cz/evidence-knihoven-adresar-knihoven-evidovanych-ministerstvem-kultury-a-souvisejici-informace-cs-341*. Zároveň predpokladám, že dátová sada sa môže priebežne aktualizovať (vrátane možnej zmeny názvu súboru). Preto som postup navrhla tak, aby skript najskôr automaticky identifikoval aktuálny odkaz na súbor vo formáte .XLSX. Následne sa tento súbor stiahne a uloží do zvoleného adresára.

Detaily sú dohľadateľné v priloženom kóde. Využila som Python baličky **requests**, **bs4** a **os**.

V prípade potreby je možné skript prispôsobiť:

- určením iného adresára pre ukladanie súborov
- rozšírenie pre iné formáty ako .XLSX, ako napríklad csv, json, tsv, ...
- pridanie timestampu do názvu, aby uživateľ vedel dátum sťahovania
- rozšírenie chybových hlášok

2 Predspracovanie a transformácia dát

Na ďalšiu prácu s dátami som použila Python balíček **pandas**. Na predspracovanie väčších dát je databázový nástroj PostgreSQL efektívnejší, ale pre menšie dátové sady je postačujúci aj Python a balíček pandas. Pôvodné dáta z .XLSX formátu som načítala do pandas dataframe. Následne prebehol prvý pohľad na dáta, čistenie a predspracovanie.

V prvom kroku boli odstránené stĺpce a riadky obsahujúce iba prázdne hodnoty (NaN). Ďalej som zmenila dátové typy stĺpcov z typu object na vhodnejší string alebo datetime, podľa charakteru dát. Táto úprava zjednoduší následnú prácu s dátami. Zároveň som odstránila diakritiku z názvov stĺpcov.

Ďalšie možné úkony v predspracovaní:

- Ďalšia úprava a skrátenie názvov stĺpcov pre jednoduchšiu manipuláciu (odstránenie whitespace a veľkých písmen, a pod), pričom pôvodné názvy by boli uchované spolu s dátovou sadou.
- Kontrola obsahu ďalších stĺpcov, identifikácia chybových či nezmyselných hodnôt (napríklad chyba v adrese, identifikátore IČ).
- Kontrola chýbajúcich hodnôt, prípadné doplnenie.
- Kontrola duplicitných hodnôt.
- Dátová sada obsahuje adresy sídla a samotnej knižnice. Preto by bolo vhodné pridať aj ČSÚ kódy okresov a krajov pre interoperabilitu s inými databázami.
- Využitie identifikátora B/H - IČ, pomocou ktorého je možné dohľadať doplňujúce informácie o provozovatelovi.

Ďalšie prípadné kroky by som vykonala v závislosti od ďalšej časti pipeliny a cieľov analýzy.

3 Uloženie dát

Po vyčistení dát by som ich znova uložila na vlastnom úložnom priestore, aby s nimi bolo možné ďalej pracovať aj v prostredí Python. Namiesto formátu .XLSX, ktorý je určený primárne pre MS Excel, by som zvolila všeobecnejší formát, napríklad .csv. Samozrejme, v závislosti od cieľov analýzy a špecificity dát vieme použiť iné formáty (HDF5, json).

Pre ich bezpečné uloženie by som určite využila relačnú databázu, ako je napríklad PostgreSQL. Relačná databáza sa hodí najmä v prípadoch, keď s dátami bude pracovať viac uživateľov, potrebujeme zdokladovať spracovanie alebo sa jedná o objemné dáta. Relačná databáza zabezpečuje bezpečné uloženie dát a chráni pred stratou. Zároveň ponúka šifrovanie dát počas prenosu. Dáta by som teda nahrala do relačnej databázy a sprístupnila používateľom na ďalšiu prípadnú analýzu.

Dáta by som teda nahrala do relačnej databázy a sprístupnila ich ďalším používateľom na ďalšie spracovanie a analýzu.

4 Popis dátovej sady pomocou CCMM

V poslednom kroku bolo potrebné zaistiť metadatový popis dátovej sady pomocou modelu Czech Core Metadata Model (*https://github.com/techlib/CCMM/wiki*).

Po preskúmaní dokumentácie CCMM som našla vzorovú XML reprezentácia metadát podľa modelu CCMM: *https://github.com/techlib/CCMM/blob/main/ccmm_sample.xml*.

Nasledovný postup by zahŕňal stiahnutie vzorových metadát a postupné manuálne dopĺňanie informácií do dokumentu. Prvých pár riadkov by mohlo vyzerať takto:

```
<?xml version="1.0" encoding="UTF-8"?>
<dataset xsi:schemaLocation="https://raw.githubusercontent.com/main/scheme/scheme.xsd"
    <iri>https://organization.cz/dataset_server/dataset_id</iri>
    <publication_year>2025</publication_year>
    <version>1.0.0</version>
    <title>Evidence knihoven 06.08.2025</title>
    <description>
        <description_text>Tato datova sada obsahuje seznam registrovanych knihoven v CR.
    <is_described_by>
        <date_created>2025-08-24</date_created>
        <qualified_relation>
            <relation>
                <person>
                    <name>Kostialova</name>
                    <given_name>Lucia</given_name>
                    <family_name>Kostialova</family_name>
                    <contact_point>
                        <email>kostial.lucia@email.com</email>
                        <phone>+0112345678</phone>
```

Na záver by som metadata sprístupnila ostatným uživateľom a nahrala do relačnej databázy.