Projekt 1. část: návrh zpracování a uložení dat

Vysoké učení technické v Brně
Fakulta informačních technologií
UPA - Ukládání a příprava dat
25. října 2020

1 Zvolené téma

04: COVID-19 (dr. Rychlý)

2 Řešitelé

- Bc. Kapoun Petr – xkapou04,
- Bc. Nováček Pavel – xnovac16,
- Bc. Willaschek Tomáš – xwilla00.

3 Zvolené dotazy a formulace vlastního dotazu

- **Dotaz 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 B:** určete vliv počtu nemocných a jeho změny v čase na sousední okresy (aneb zjistěte jak se šíří nákaza přes hranice okresů).
- **Vlastní dotaz:** určete vliv věku na délku nemoci a úmrtnost.

4 Stručná charakteristika zvolené datové sady

Základním zdrojem dat pro všechny dotazy jsou **otevřené datové sady COVID-19 v ČR**[3] dostupné skrze veřejné API[1] ve formátech JSON a CSV.

Datová sada `COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (v2)` obsahuje následující charakteristiky o nakažených osobách:

- `datum` – datum, kdy byla osoba pozitivně testována,
- `vek` – věk nakažené osoby,
- `pohlavi` – pohlaví nakažené osoby,
- `kraj_nuts_kod` – identifikátor kraje podle klasifikace NUTS 3, ve kterém byla pozitivní nákaza hlášena krajskou hygienickou stanicí,
- `okres_lau_kod` – identifikátor okresu podle klasifikace LAU 1,
- `nakaza_v_zahranici` – příznak, zda došlo k nákaze mimo ČR,
- `nakaza_zeme_csu_kod` – identifikátor státu v zahraničí, kde došlo k nákaze (dvoumístný kód z číselníku zemí CZEM).

[1] `https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19`

Další dvě použité datové sady jsou `COVID-19: Přehled vyléčených dle hlášení krajských hygienických stanic` a `COVID-19: Přehled úmrtí dle hlášení krajských hygienických stanic`, které mají shodnou strukturu:

- `datum` – datum vyléčení nebo úmrtí osoby,
- `vek`,
- `pohlavi`,
- `kraj_nuts_kod`,
- `okres_lau_kod`.

Jelikož tyto datové sady používají identifikátory *NUTS 3*[2] pro kraje a *LAU 1*[3] pro okresy, využíváme dále číselníky od Českého statistického úřadu, které obsahují mapování těchto identifikátorů na odpovídající kraje a okresy a informace o nich.

Číselník okresů[4] a **číselník krajů**[5] jsou oba ve formátu CSV a obsahují:

- `KODJAZ` – kód jazykové verze textů,
- `AKRCIS` – akronym číselníku/klasifikace,
- `KODCIS` – kód číselníku/klasifikace,
- `CHODNOTA` – kód položky,
- `ZKRTEXT` – zkrácený název položky,
- `TEXT` – plný název položky,
- `ADMPLOD` – počátek administrativní platnosti,
- `ADMNEPO` – konec administrativní platnosti.

Zvoleným programovacím jazykem je `Python`[4]. K získání dat použijeme standardní moduly jazyka `Python`[4] `json`[6] a `csv`[7]. K nahrání dat použijeme modul `elasticsearch`[8].

5 Zvolený způsob uložení uložení surových dat

Pro uložení dat jsme zvolili NoSQL řešení `Elasticsearch`[2][9]. `Elasticsearch` je dokumentově orientovaný vyhledávací engine naprogramovaný v jazyce Java s použitím `Lucene`[1], díky kterému je velmi efektivní při komplexním vyhledávání na velkých objemech dat. Schéma struktur kolekcí odpovídá struktuře uložených dat ve vstupních datových sadách. Index je typu klíč-hodnota a pro každý vstupní soubor bude existovat právě jeden index. Klíč nabývá hodnoty `CHODNOTA` u číselníku okresů a krajů a ve zbývajících indexech se generuje automaticky, protože data neobsahují žádnou unikátní hodnotu.

[2]Nomenklatura územních statistických jednotek; úroveň 3 odpovídá krajům

[3]Local Administrative Units; úroveň 1 odpovídá okresům

[4]Číselník okresů: `http://apl.czso.cz/iSMS/cisexp.jsp?kodcis=109&typdat=0&cisvaz=80007_210&datpohl=20.10.2020&cisjaz=203&format=2&separator=,`

[5]Číselník krajů: `http://apl.czso.cz/iSMS/cisexp.jsp?kodcis=100&typdat=0&cisvaz=80007_885&datpohl=20.10.2020&cisjaz=203&format=2&separator=,`

[6]Dokumentace modulu `json` jazyka `Python`[4]: `https://docs.python.org/3/library/json.html`

[7]Dokumentace modulu `csv` jazyka `Python`[4]: `https://docs.python.org/3/library/csv.html`

[8]Modul `elasticsearch` pro `Python`: `https://pypi.org/project/elasticsearch/`

[9]Dokumentace `Elasticsearch`[2]: `https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html`

Reference

[1] Apache Software Foundation. *Apache Lucene* [online]. 1999 [cit. 2020-10-20]. Dostupné na: <https://lucene.apache.org/>.

[2] Elastic NV. *Elasticsearch* [online]. 2010 [cit. 2020-10-19]. Dostupné na: <https://www.elastic.co/>.

[3] Komenda M., Karolyi M., Bulhart V., Žofka J., Brauner T., Hak J., Jarkovský J., Mužík J., Blaha M., Kubát J., Klimeš D., Langhammer P., Daňková Š., Májek O., Bartůňková M., Dušek L.. COVID-19: Přehled aktuální situace v ČR. Onemocnění aktuálně. In [online]. Praha: Ministerstvo zdravotnictví ČR, 2020 [cit. 2020-10-20]. Vývoj: společné pracoviště ÚZIS ČR a IBA LF MU. Dostupné na: <https://onemocneni-aktualne.mzcr.cz/covid-19>. ISSN 2694-9423.

[4] Python Software Foundation. *Python* [online]. 1991 [cit. 2020-10-20]. Dostupné na: <https://www.python.org/>.