UPA - Projekt 2. část

Bc. Aleš Tetur (xtetur01),
Bc. Samuel Stuchlý (xstuch06),
Bc. Martin Litwora (xlitwo00)

Brno, 15. 12. 2021

1. Zadání

Našim cílem bylo zvolit si dva úkol ze skupiny A a jeden úkol ze skupiny B k našemu tématu. Poté jsme měli vymyslet dva vlastní nové úkoly. Pro všechny zvolené dotazy jsme dále měli implementovat nástroj pro extrakci potřebných dat z původního úložiště vytvořeného v první části projektu do souborů v CSV formátu. Následně jsme měli pro všechny zvolené dotazy navrhnout a implementovat řešení, které tyto dotazy zodpoví.

Dále jsme si měli zvolit jednu z dolovacích úloh ze skupiny C. Pro tuto úlohu připravit data tak, aby výsledná data mohla být použita dolovacím algoritmem. Připravit soubor ve formátu CSV, kde každý řádek bude odpovídat jednomu objektu, každý sloupec nějakému atributu. Dále pak ve vybraných datech detekovat odlehlé hodnoty a nahradit je jinou vhodnou hodnotou. Pro jeden zvolený sloupec provést normalizaci hodnot a pro jiný sloupec diskretizaci hodnot. Opět implementovat nástroj, který potřebná data extrahuje z úložiště vytvořeného v první části projektu, který je požadovaným způsobem upraví.

2. Výběr dotazů

Jako první dotaz ze skupiny A jsme si vybrali vytvořit sloupcové grafy zobrazující počty poskytovatelů určitého oboru pro Brno a zbytek Jihomoravského kraje. Vybrali jsme tyto obory (vybrali jsme 15 oborů s největším počtem poskytovatelů):

- Všeobecné lékařství
- zubní lékařství
- Fyzioterapeut
- gynekologie a porodnictví
- praktické lékařství pro děti a dorost
- praktické lékárenství
- vnitřní lékařství
- Zubní technik
- Chirurgie
- Všeobecná sestra
- ortopedie a traumatologie pohybového ústrojí
- psychiatrie
- neurologie
- veřejné lékárenství
- dermatovenerologie

Druhým dotaz A je Vytvořte spojnicového graf zobrazující historii počtu poskytovatelů v 5 zvolených oborech, kdy není nutné zobrazovat data pro každý měsíc, ale stačí čtvrtletní hodnoty. Vybrali jsme tyto obory:

- Fyzioterapeut
- Chirurgie
- patologie
- psychiatrie
- zubní lékařství

Jako dotaz ze skupiny B jsme si zvolili sestavit žebříček krajů dle počtu obyvatel na jednoho praktického lékaře (obor všeobecné praktické lékařství). Výsledky zobrazte graficky. Graf bude pro každý kraj zobrazovat počet praktických lékařů v kraji, celkový počet obyvatel a počet obyvatel na jednoho lékaře. Pro přesnější výsledky použijte počet obyvatel kraje nad 20 let.

Jako vlastní dotazy jsme zvolili 1) jak se v historii v Brně měnil poměr můžu a žen a zjistit 2) jaký je počet dětí na jednoho dětského lékaře v Jihomoravském a ve Středočeském kraji a jak se tento počet měnil poslední tři roky.

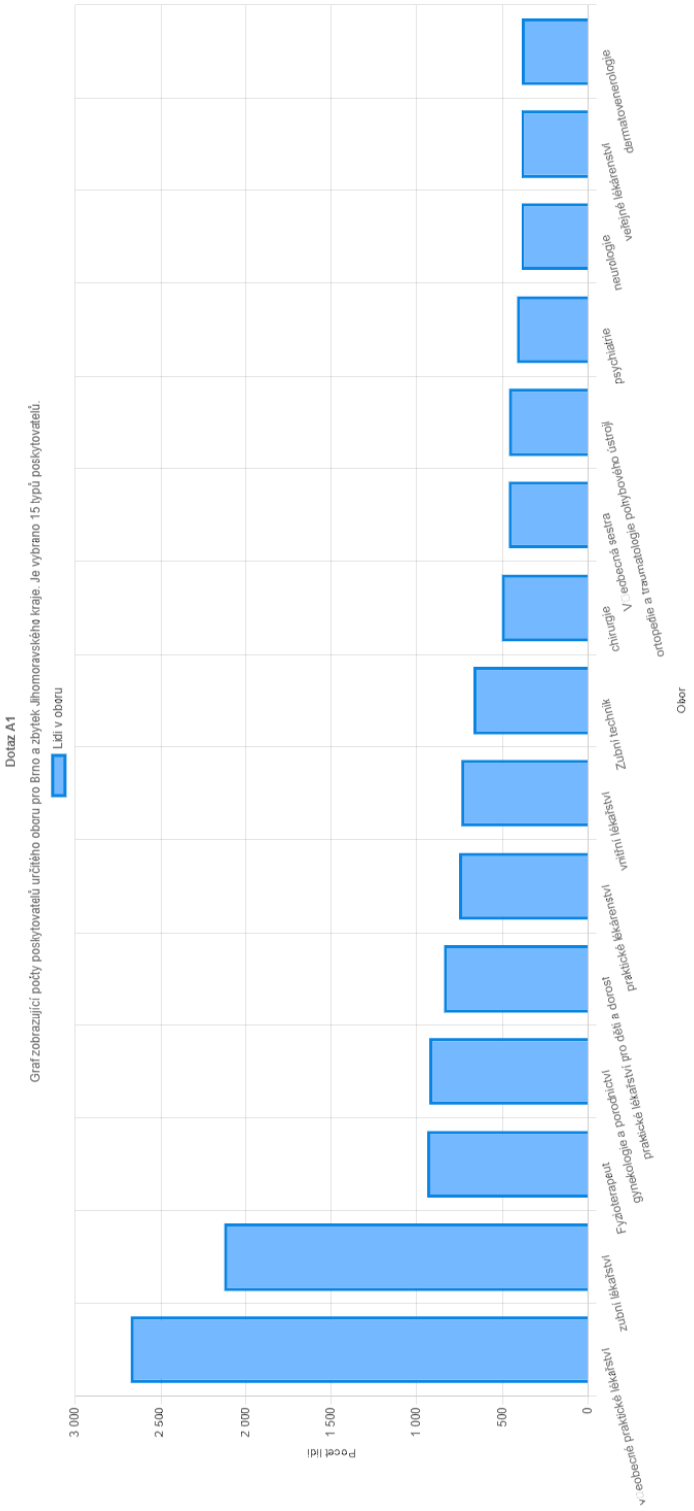
Dotaz ze skupiny C, pro dolovací úlohu, jsme si vybrali hledání skupin oborů péče s obdobným historickým vývojem. Atributy jsou: historické počty poskytovatelů péče daného oboru. Počty poskytovatelů jsou vyhodnoceny měsíčně. Dále jsme vybrali 20 oborů zdravotní péče:

- Všeobecné lékařství
- zubní lékařství
- Fyzioterapeut
- gynekologie a porodnictví
- praktické lékařství pro děti a dorost
- praktické lékárenství
- vnitřní lékařství
- Zubní technik
- Chirurgie
- Všeobecná sestra
- ortopedie a traumatologie pohybového ústrojí
- psychiatrie
- neurologie
- veřejné lékárenství
- dermatovenerologie

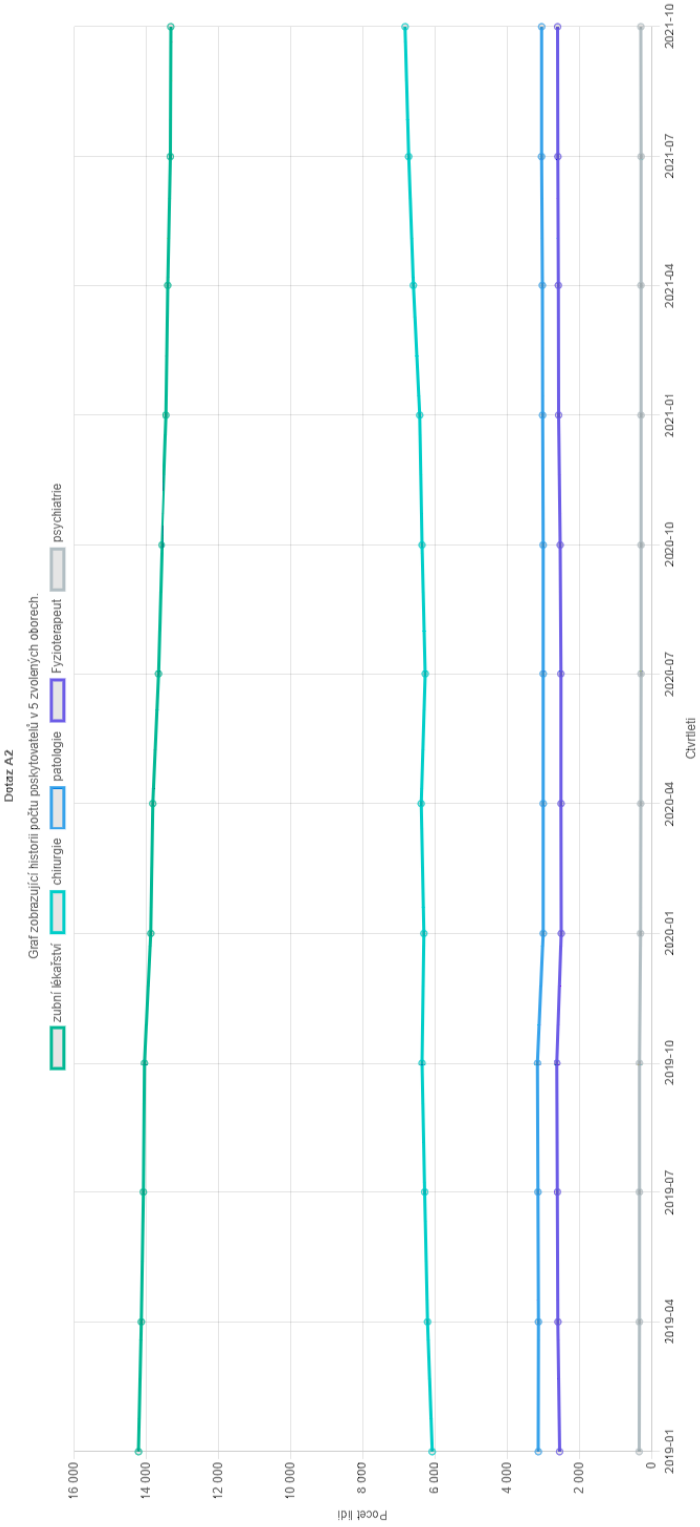
3. Výsledky získané dotazy

Výsledky, které jsme získali pomocí dotazů jsou přístupné on-line na adrese: https://www.stud.fit.vutbr.cz/~xtetur01/UPA/. Dále jsou obrázky výsledků.

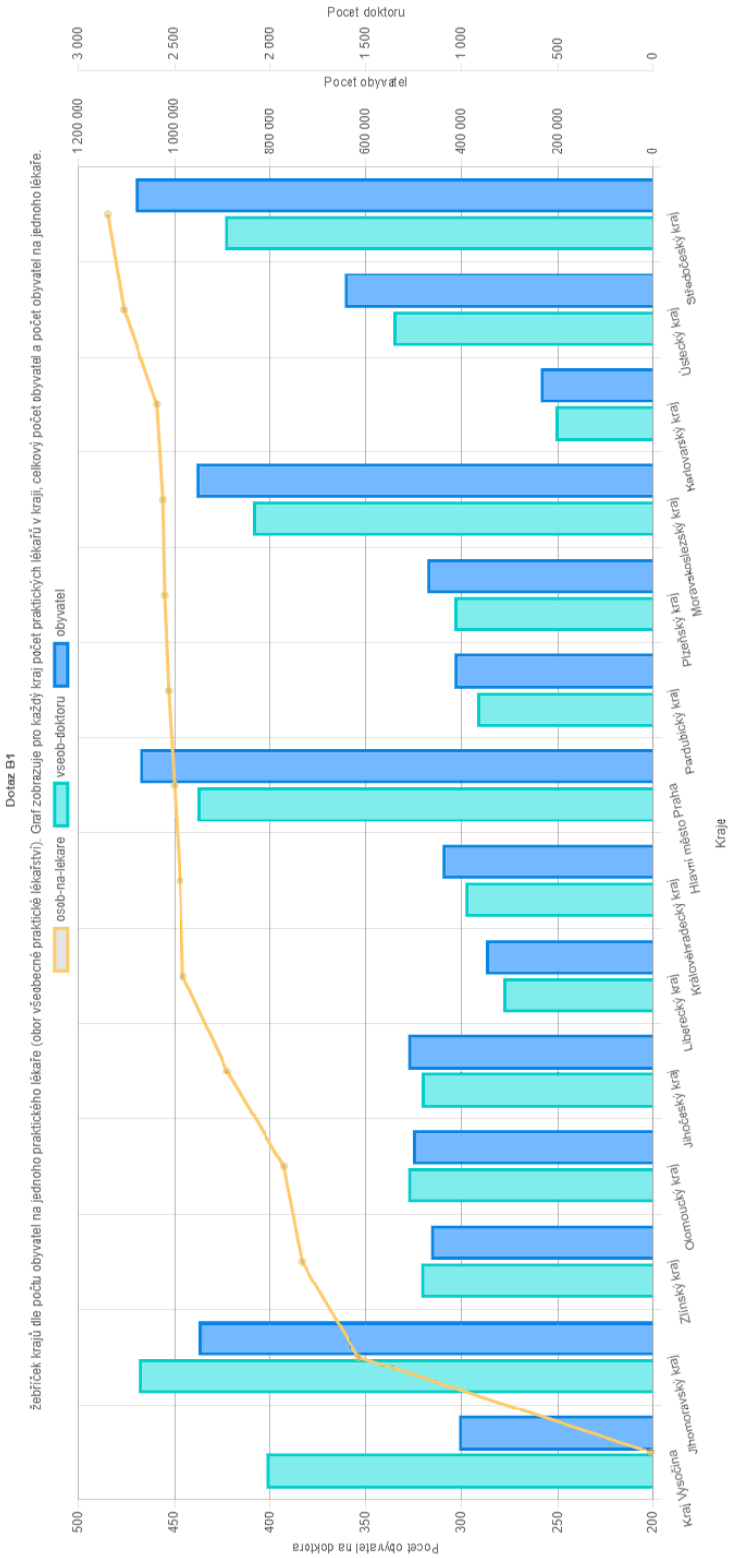
3.1. První dotaz A



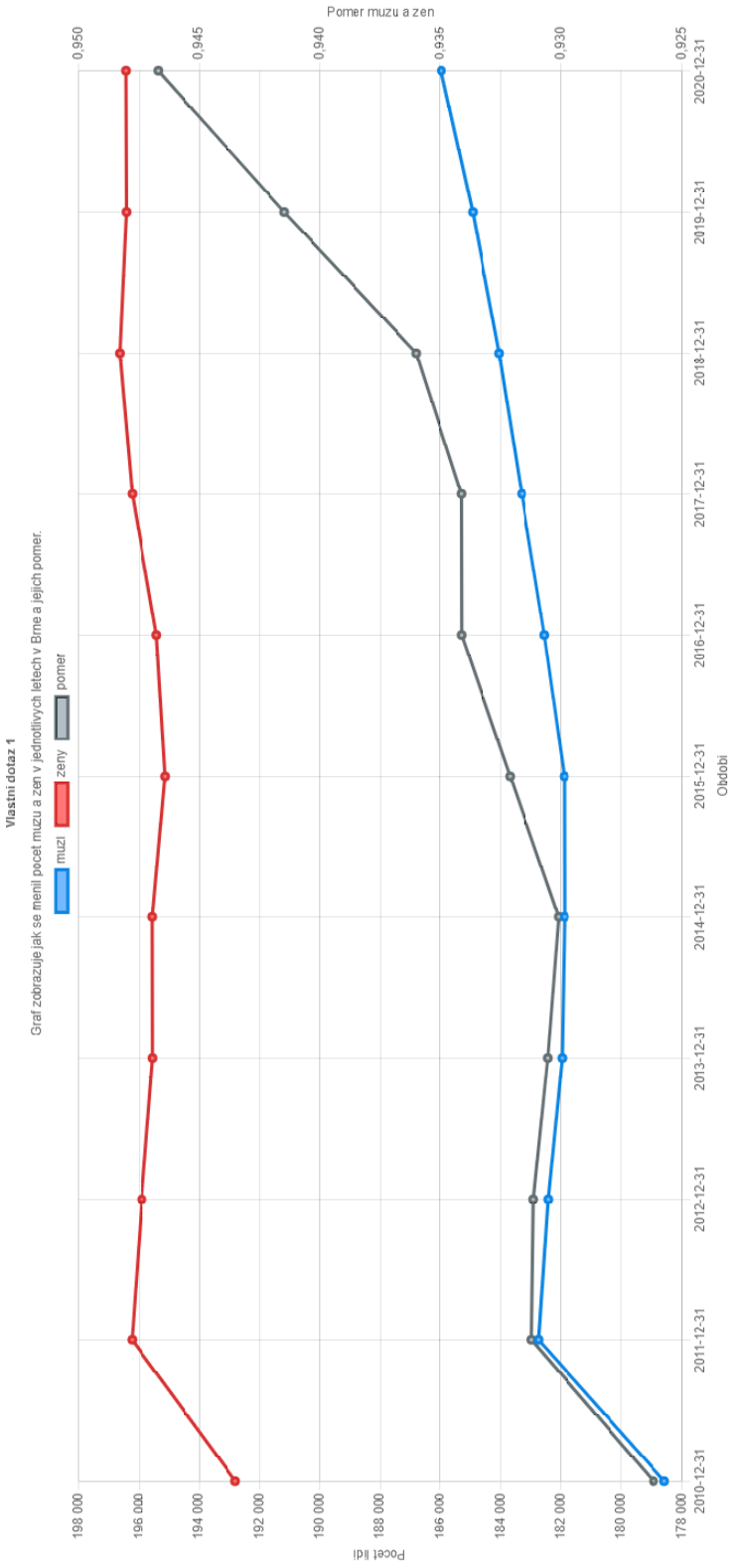
3.2. Druhý dotaz A



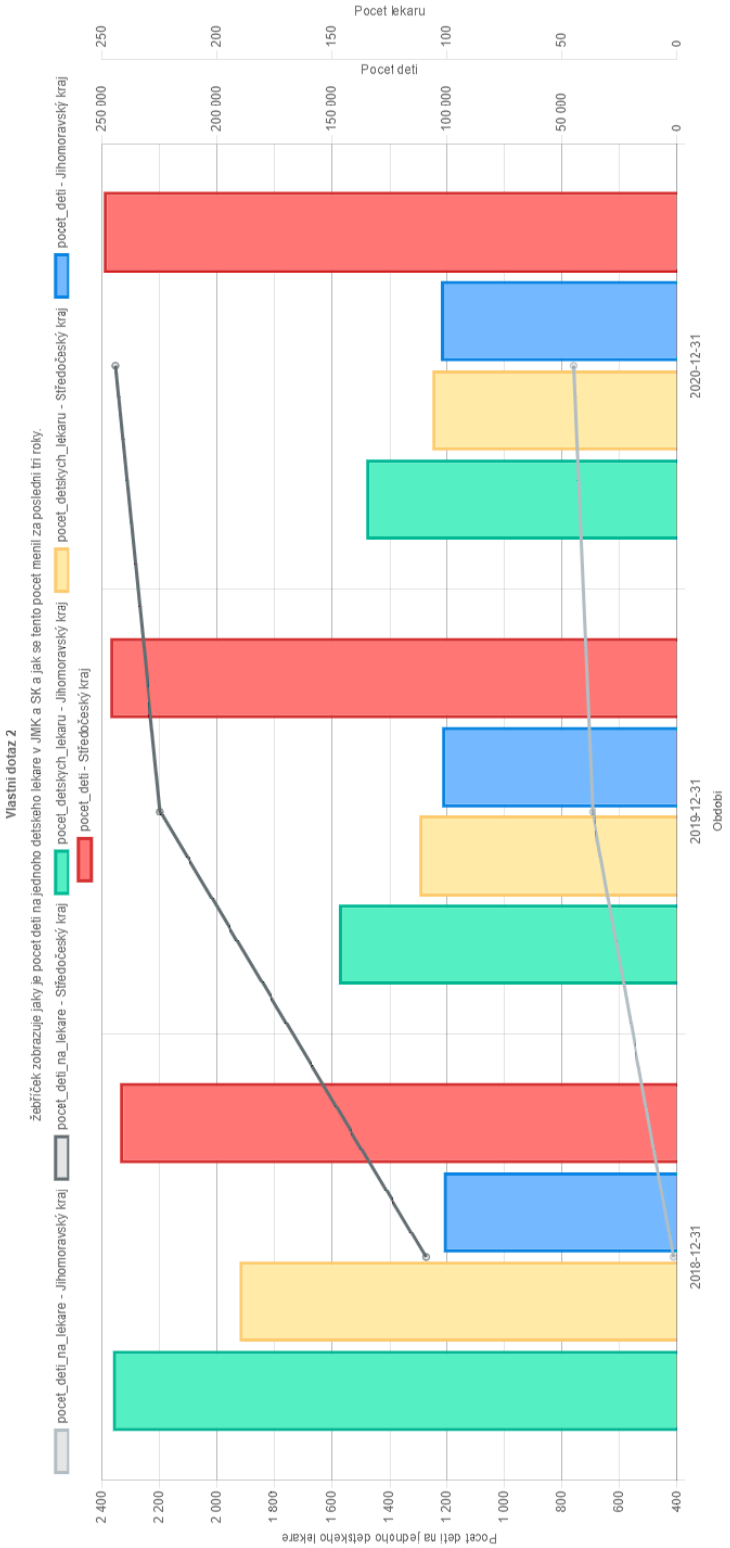
3.3. První dotaz B



3.4. První vlastní dotaz



3.5. Druhý vlastní dotaz



4. Dotaz pro dolovací úlohu C

Formát souboru ve formátu csv je následující. Jednotlivé sloupce jsou atributy obsahující informace o počtu poskytovatelů péče daného oboru v daném období. Řádky jsou pak jednotlivé entity, tj. jednotlivé obory. Při detekci odlehlých hodnot jsme žádnou nezjistili, jelikož u počtu poskytovatelů není žádný extrémní výkyv. Jediné čeho si lze povšimnout je, že po prvních třech kvartálech (2019-01 až 2019-07) se asi o třetinu snížily počty poskytovatelů. Tento jev se ovšem objevuje u všech poskytovatelů, proto jsme se rozhodli tyto hodnoty nijak neupravovat a nechali jsme je jak jsou.

Dále pro zařízení “praktické lékárenství” chybí data v databázi pro celý rok 2019. Tento fakt jsme zohlednili tak, že jsme do csv souboru napevno uložili hodnotu “nan” místo chybějících hodnot.

Normalizaci jsme provedli pro atribut “2021-07”. Zvolili jsme metodu min-max normalizace. Pro diskretizace jsme zvolili metodu plnění (“binning”), tedy kvantilové rozdělení na intervaly s (přibližně) stejným počtem prvků v každém intervalu. Na počet košů/bins jsme zvolili čtyři, protože to nejlépe vystihuje obsah daného sloupce. Na diskretizaci jsme použili funkci *qcut()* z knihovny Pandas. Diskretizace byla provedena pro atribut “2021-10”. Výsledek před úpravou je uložen v souboru “query-c.csv” a po úpravě v “query-c-after.csv”

5. Implementace

Další částí řešení je vytvoření programu pro přípravu dat do souborů CSV. Pro implementaci tohoto programu jsme zvolili programovací jazyk Python, protože se v něm velmi jednoduše pracuje s CSV a JSON.

Samotný program postupně pomocí dotazů na databázi MongoDB a nad jednotlivými kolekcemi získává potřebné informace a hodnoty. Poté potřebné informace upraví a uloží je do výsledných souborů.

Vizualizace výsledků dotazů je implementována pomocí HTML stránky, které zobrazuje jednotlivé grafy. Pro ně jsme využili knihovnu Chart.js, která je určena pro vytváření grafů. Výsledky jsou dostupné online na:
https://www.stud.fit.vutbr.cz/~xtetur01/UPA/

GitHub repozitář projektu:
https://github.com/SamuelStuchly/UPA2021

6. Zprovoznění

Pro fungování je nutné mít nainstalován *Python3* a *Docker*.
Před spuštěním je nutné spustit

make clean

dále vytvořit virtuální prostředí pomocí příkazu

make venv

source venv/bin/activate

spustit skript příkazem

make

Grafy lze zobrazit i lokálně, když je spuštěn www server, např:

python3 -m “http.server”,

Pote stačí přejít v prohlížeči na:

http://0.0.0.0:8000/

Podrobnější postup je v souboru README.md (dostupný v repozitáři projektu)