Zadanie 1 – import tweetov do PostgreSQL

[Databazy/main.py at master · MartinJank/Databazy (github.com)]

Opis algoritmu:

Zvolil som jednoduchý algoritmus, kde som najskôr spracoval súbor *authors.jsonl* a následne som pokračoval na *conversations.jsonl*. Autorov a údaje o nich som zapísal do tabuľky authors. Konverzácie prechádzam 2 krát. Najskôr zapíšem dáta do tabuliek bez cudzích kľúčov a následne ostatné.

1. Otvorenie súboru *authors.jsonl*
2. Vytvorenie listu s údajmi pre tabuľku authors
3. Zapísanie údajov do tabuľky po 10000 riadkoch
4. Premazanie listu a pokračovanie na ďalších 10k
5. Zapísanie zvyšku zo súboru *authors.jsonl*
6. Otvorenie súboru *conversations.jsonl*
7. Vytvorenie listov s údajmi pre tabuľky conversation, hashtags, context_domains a context_entities
8. Zapísanie author_id do tabuľky authors ak sa tam nenachádzalo
9. Zapísanie ostatných listov do príslušných tabuliek po 10k
10. Zapísanie zvyšku zo súboru *conversations.jsonl*
11. Vytvorenie listov pre tabuľky annotations a links
12. Zapísanie listov do príslušných tabuliek po 10k
13. Pri linkoch sa kontroluje dĺžka
14. Vždy po 10k záznamoch sa vypíše čas

Použité technológie:

PostgreSQL:

- Bolo zadané
- Užívateľom definované typy
- Užívateľom definované operátory
- Podpora pre geografické objekty cez PostGIS
- Tabuľková dedičnosť

Python:

- Jednoduchosť
- Dobré prepojenie a komunikácia s databázami
- Je dynamicky typový jazyk
- Veľká komunita a veľa materiálov

Použité Queries:

```
args_str = ','.join( cursor.mogrify("(%s,%s,%s,%s,%s,%s,%s,%s)",
element).decode("utf-8") for element in list_to_insert)

cursor.execute('INSERT INTO authors VALUES {0} ON CONFLICT DO
NOTHING'.format(args_str), )
```

Každé query má rovnakú formu, kde sa najskôr vytvorí string z listu záznamov aby sa mohli záznamy vkladať vo väčšom množstve. Tento spôsob je efektívnejší ako keby sme použili *executemany.* Používam *ON CONFLICT DO NOTHING* aby som predišiel pádu tohto query. Ak sa rovnaké *id* nachádza v tabuľke, záznam sa ponechá nezmenený a program pokračuje ďalej. Iné queries ako insert nepoužívam.

Dĺžka trvania:

Po spustení programu sa v konzole vypíše čas od spustenia programu tak ako aj čas po každých 10k záznamoch.