Dokumentácia - iDnes scraper do grafu

4IZ553 Graph Databases and Graph Data Visualization

Michal Dupkala

# **Účel programu**

Program umožňuje transformáciu dát zo spravodajského portálu iDnes do grafovej podoby v databázi neo4j.

Transformácií podliehajú:

* nadpis člańku
* autor
* súvisiace články
* témata

Výsledkom sú naloadované dáta, ktoré obsahujú články (node Article), autorov (node Author) a témata (node Topic) so vzťahmi:

* Article -> relates\_to -> Article
* Article -> written\_by -> Author
* Article -> is\_about -> Topic

# Spustenie a používanie

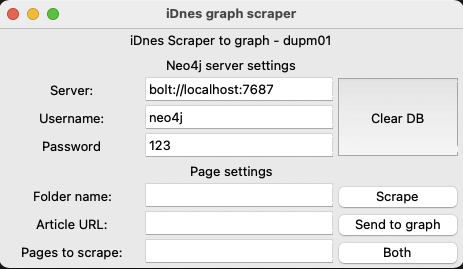
## Požiadavky

1. Nainštalovaný Python 3.9 a vyšší.
2. Nainštalované potrebné balíčky pomocou : *pip install -r /path/to/requirements.txt*
3. Spustená neo4j databáza na ktorú sa môže program napojiť (vyvíjané na verzii 4.3.5)

## Spustenie

Program sa spúšťa skriptom bin/main.py . Napríklad pomocou príkazu z root zložky :

*python3 bin/main.py*



Obrázok 1 - Obrazovka programu [zdroj: Autor]

Pre správnu funkcionalitu tlačidla Clear DB, ktoré vymaže všetky uzly a vzťahy z danej databáze je potrebné mať vyplnené všetky položky v časti „Neo4j server settings“.

Tlačidla „Scrape“ , „Send to graph“ a „Both“ fungujú správne pri vyplnení všetkých polí.

## Fungovanie programu a popis kódu

Program je rozčlenený na zložku /conf, ktorá obsahuje konfiguračné súbory o serveri a scrapovaní a na zložku /bin.

V zložke bin sa nachádza main.py, ktorý tvorí grafické rozhranie a funkcionalitu si volá z modulov v zložke /bin/lib – *yaml\_editor.py*, *scraper.py* a*graphtransfer.py* .

### main.py

Grafické rozhranie programu pomocou knižnice tkinter. Rozhranie je rozdelené na 2 stavebné prvky – header a center. V sekcii header nájdeme Label pre nadpis a center je následne rozdelený na top\_row a bottom\_row.

V priestore top\_row sú prvky súvisiace s definíciou a prácou čisto neo4j serverom a v priestore bottom\_row sú prvky pre hlavnú funkcionalitu programu.

Tlačítka sa pomocou lambda funkcií odkazujú a skripty v zložke lib.

|  |
| --- |
| tk.Button(top\_row, text = "Clear DB", command = lambda: [yaml\_editor.edit\_conf\_graph(server\_entry.get(), username\_entry.get(), password\_entry.get()), graphtransfer.clear\_db()], width= 10).grid(row = 1, column = 2, rowspan = 3, sticky = tk.N + tk.S) |

### yaml\_editor.py

Tento modul obsahuje funkcionalitu pre úpravu súborov v /conf. Spúšťa sa po kliknutí tlačidla pred ostatnými súbormi a uloží vstup z textových polí do súborov.

Scraper a graphtransfer k týmto súborom následne pristupujú a berú ich ako vstup.

* conf\_scrape.yaml

url: url na článok

pages\_to\_scrape: počet strániek na zoscrapovanie (do grafu sa dostanú spolu

dir\_name: názov zložky do ktorej sa články uložia v podobe json súborov

* conf\_graph.yaml

server: neo4j server, napíklad: bolt://localhost:7687

username: neo4j

password: password

dir\_name: názov zložky s json súbormi pre vloženie do grafovej databáze

### scraper.py

### graphtransfer.py