Homework #3 – Neo4j

Introduzione - L'esercizio richiede di creare un grafo di proprietà in Neo4j partendo da un dataset di contenuti erogati sulla piattaforma Netflix (film, serie TV). Il dataset (`netflix.csv`) contiene un film o serie TV su ogni riga, descritto da un identificatore (`show_id`), un tipo (Movie, TC Show), un titolo, un regista, un cast (nomi di attori separati da virgole), una nazione di produzione, la data di aggiunta nel dataset, l'anno di uscita, il rating, la durata (in minuti o stagioni per le serie TV), i generi (uno o più, separati da virgola), una descrizione testuale. Non tutte le informazioni sopra elencate sono sempre presenti.

Descrizione del compito - Scaricare il dataset e, utilizzando Python e Cypher, creare un nuovo database in Neo4j (suggerimento – utilizzare i DataFrame come primo step per caricare i film in Python):

- creare un nuovo nodo Movie per ogni film, con il proprio ID, il titolo, la data di aggiunta nel dataset, l'anno di rilascio, il rating, la durata (in minuti), la descrizione;
- creare un nuovo nodo SerieTV per ogni serie TV, con il proprio ID, il titolo, la data di aggiunta nel dataset, l'anno di rilascio, il rating, la durata (in stagioni), la descrizione;
- creare un nuovo nodo Person per ogni regista e ogni elemento del cast (eliminando i doppioni), ciascuno con il proprio nome;
- collegare i film e le serie TV alle persone del cast tramite la relazione :ACTED_IN;
- collegare i film e le serie TV alla persona che funge da regista tramite la relazione :DIRECTED;
- creare un nuovo nodo Categoria per ogni genere e collegare i film e le serie TV alla categoria tramite una relazione :IN_CATEGORY;
- creare un nuovo nodo Country per ogni nazione di produzione e collegare i film e le serie TV alla nazione di produzione tramite una relazione :WHERE.

Eseguire le seguenti interrogazioni sul database appena popolato:

1. Visualizzare lo "schema" del database (tipi di nodi e di relazioni tra i nodi) tramite Neo4j Desktop.
2. Estrarre i primi 5 attori come numero di film o serie TV a cui hanno partecipato.
3. Estrarre per ogni categoria il numero di film e serie TV di quella categoria, in ordine crescente, e visualizzare il risultato in un grafico a barre in Python.
4. Trovare il percorso più breve tra due persone (a scelta dello studente) e visualizzarlo come sottografo tramite Neo4j Desktop.
5. Calcolare la similarità tra tutte le possibili coppie di categorie, applicando la seguente formula (dove A e B rappresentano l'insieme di film associati alle due categorie, rispettivamente, e |.| rappresenta la cardinalità di un insieme):

$$Sim(Cat_A, Cat_B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

6. Riportare in un grafico l'evoluzione nel tempo del numero di film o serie TV per ogni nazione di produzione (il grafico presenterà gli anni sull'asse delle ascisse e N grafici ognuno riportante il numero di film e serie TV per ogni nazione al variare del tempo).

Dettagli sulla consegna – Predisporre un notebook Python (o un file .py) in cui è riportato lo svolgimento di tutti gli esercizi e un file PDF in cui sono riportati i commenti sui risultati o e l'output di Neo4j Desktop, laddove esplicitamente richiesto (per esempio, l'immagine di un sottografo o del percorso più breve tra due nodi). Includere tutti i file in uno zip e caricare l'archivio su Moodle.