PDT - Neo4j

https://github.com/MennoCoehoorn/PDT_6.git

Marek Štrba

December 2021

Obsah

9 BONUS: Nájdite najkratšie cesty medzi Katy Perry (katyperry) a Donaldom Trumpom cez vzťah RETWEETS (a tým pádom aj POST). Všetky cesty, kde Katy Perry retweetla post Accountu, ktorý retweetol post Accountu, ktorý..., ktorý retweetol post Donalda Trumpa. 11

1 Zoberte si nami vytvorený dataset z vašich tweetov a importujte ho cez neo4j-admin

S importom som bojoval asi 3,5 hodiny (som ten spolužiak, ktorý plakal - nie naozaj ale nervy som mal). Nakoniec som s pomocou Martina prišiel na to, že problém bol v tom, že dump bol verzia 4.2.1 a ja som to chcel naimportovať do DMBS verzie 4.4.0. Za trest mu nedonesiem pivo z New Yorku.

```
D:\FIIT\Apps\N4J\relate-data\dbmss\dbms-014a5628-ded8-42e8-9ded-959eccebbd60>D:\FIIT\A
pps\N4J\relate-data\dbmss\dbms-014a5628-ded8-42e8-9ded-959eccebbd60\bin\neo4j-admin lo
ad --from=D:\FIIT\ING\1_Semester\PDT\PDT_6\tweets2021.dump  --database=tweets --force
--verbose
```

Figure 1: Jeden z asi 11 importov, ktoré som robil. Finálny vyzeral rovnako, len nemal verbose a išiel priamo do neo4j databázy, nevytváral som dodatočne ďalšiu

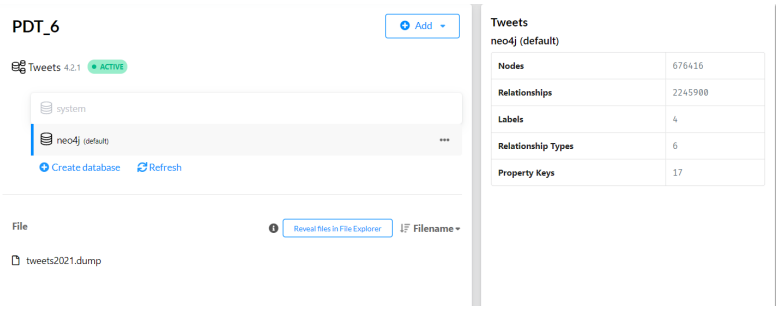

Figure 2: Ukážka úspešne naimportovaného dumpu

2 Vypíšte 5 Accountov s najvyšším množstvom followerov. Nezoraďujte Accounty podľa poľa followers_count, hodnota je prevzatá z Twitteru a nezodpovedá našim vzťahom v datasete. Zaujímajú nás followujúce Accounty v našom datasete cez vzťah FOLLOWS.

Ako prvé som si matchol všetky accounts, následne som ich cez with spolu s výpočtom počtu vzťahov hocijaký node follows account a preniesol k order by followerCnt desc a to som limitom orezal na top 5. Výsledok aj query sú v priloženom screenshote.

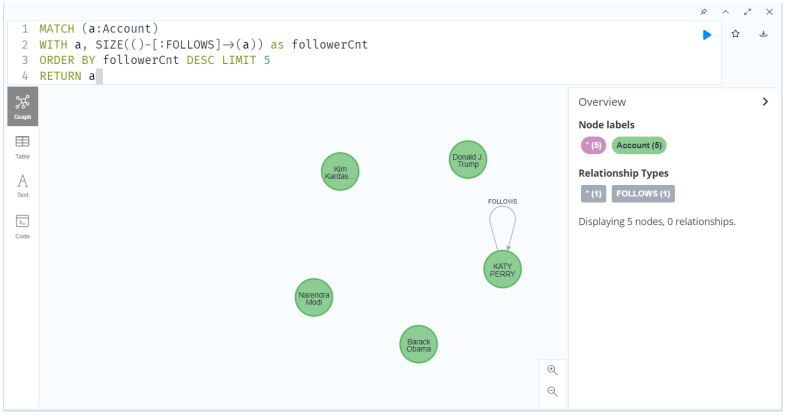

Figure 3: Ukážka query a výsledku

3 Nájdite najkratšie cesty medzi Katy Perry (screen_name: 'katyperry') a Kim Kardashian (screen_name: 'KimKardashian') cez vzťah FOLLOWS. Všetky cesty, kde Donald Trump followuje niekoho, kto followuje niekoho, kto..., kto followuje Katy Perry.

Dostať sa od Katy Perry ku Kim Kardashian bolo priamočiare, stačilo použiť funkiu shortestPath naviazanú na vzťah FOLLOWS, výsledok a query priložené v screenshote:

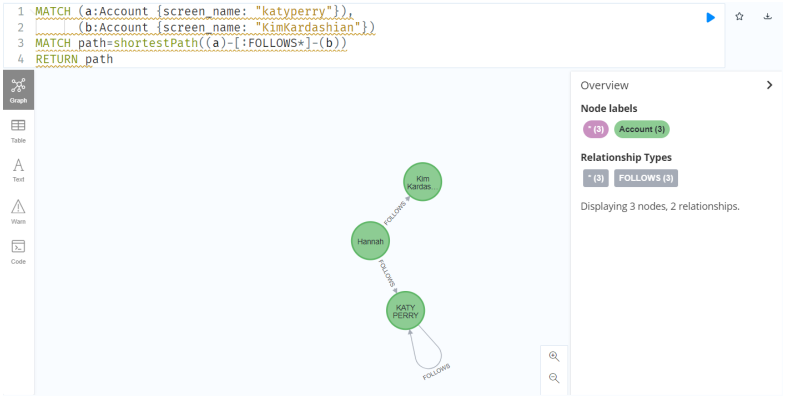

Figure 4: Katy -> Kim speedrun ukážka

Na to aby som zistil prepojenia hĺbky 4 medzi Donaldom a Katy som použil vzťah FOLLOWS s obmedzením na hĺbku 4 (dúfam, že som správne pochopil,

že to tak bolo myslené). Síce vo výsledkoch môžeme vidieť aj path Donald - BBC - Katy, ktorý má hĺbku 2 no ten je tam kvôli ceste Donald - Comport - Ashdax - BBC - Katy, čo je hĺbka 4.

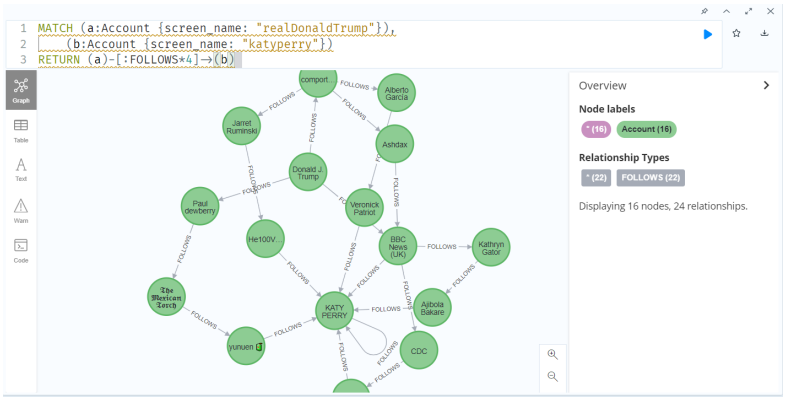

Figure 5: Ukážka Donaldových ciest ku Katy

4 Vyhľadajte neúspešné tweety influencerov. Vyhľadajte 10 najmenej retweetovanych tweetov od Accountov, ktoré sú na prvých 10 miestach v celkovom počte retweetov.

Prvým krokom bolo zistiť, ktoré účty majú najviac retweetov. Preto som si matchol accounts cez post na tweets, to som si s with preniesol spolu so súčtom retweet relationships na dané tweety a zoradil som to podľa tohoto počtu s limitom 10, takto som dostal top infulencerov. Nasledne som si znova matchol tweety na týchto influencerov cez POSTS relationship. Zrátal im znova RETWEETS relationships a zoradil teraz opačne, od najmenšieho a limitoval znovu na 10. Tieto tweety som následne vrátil ako najneúspešnejšie tweety influencerov.

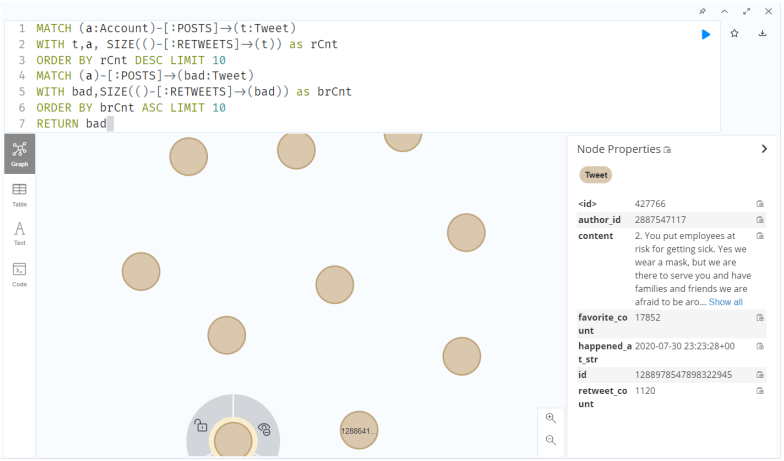

Figure 6: Query a výsledok najslabších tweetov

5 Vytvorte volaním iba jednej query nový Account s Vašim menom, ktorý bude followovať Donalda Trumpa (screen_name:"realDonaldTrump") a v tom istom volaní vytvorte tweet, ktorý bude retweetom Donaldovho najretweetovanejšieho tweetu.

Prvým krokom bolo získať najretweetovanejší Donaldov tweet. Na to som si matchol všetky Donaldové tweety (vzťah POSTS), zrátal im RETWEETS relationship, zoradil, limitoval na 1 a tak som dostal jeho najretweetovanejší tweet. Tento tweet som si pomocou with preniesol do zvyšku query. Tam som si matchol Donalda, vytvoril som si svoj účet a svoj tweet. Potom som pomocou premenných vytvoril relationships, kde ja followujem Donalda, môj tweet retweetuje jeho tweet a ja postujem svoj tweet.

Figure 7: Query vytvarajúca môj účet, tweet a potrebné relationships

Následne som vykonal kontrolné dopyty, na to či followujem správny účet a či som retweetol správny tweet:

6 Odporučte používateľovi (screen_name:"777stl") followovanie ďalších Accountov, na základe followovania rovnakých Accountov: Vyhľadajte 10 Accountov, ktoré followujú najviac rovnakých Accountov ako náš používateľ, ale náš používateľ ich ešte nefollowuje.

Ako prvé som si matchol accounty, ktoré followuje stl, nazval som ich followers (mohlo byť aj lepšie meno ale tak čo už). Následne som matchol fofollowers (ludí, ktorí followujú followerov, ale ako podmienku som pridal, že nie sú stl a ani jeden z followerov). Fofollowerov spolu s vypočítaným počtom vzťahov fofollower FOLLOWS follower som si pomocou with preniesol ďalej, tam som ich zoradil podľa countu zostupne a limitoval na 10. Výsledných fofollowerov som vrátil ako účty, ktoré by stl mal followovať.

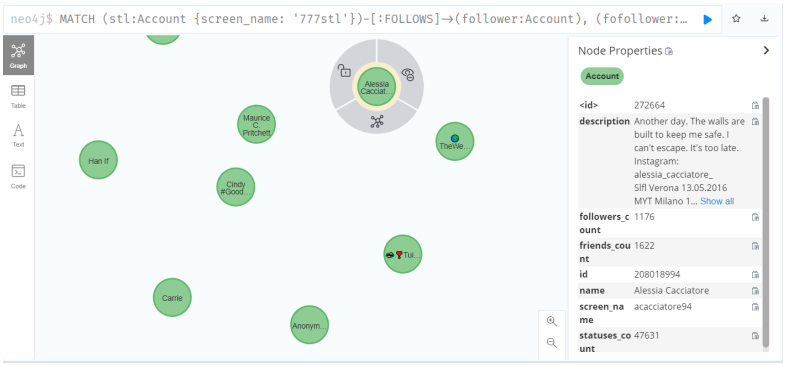

Figure 8: Query vracajúca recommendation pre používateľa 777stl

7 Odporučte používateľovi (screen_name:"DaynerWilson") followovanie ďalších Accountov na základe zhody v retweetovaní rovnakých tweetov: Vyhľadajte 10 accountov, ktoré retweetli najviac tých istých tweetov, ako náš používateľ. Ak tweet ktorý retweetujeme, je už tiež retweetom, rátajte za zhodu aj retweetovanie jeho parent tweetu – retweetovanie teda zohľadňujte rekurzívne.

Keďže asi Dayner nemá záujem followovať sám seba a hľadá nových ludí na followovanie, tak som do podmienky dal, že account nemôže byť ani Dayner a ani followovany Daynerom. Následne som cez rekurzívny vzťah RETWEETS našiel všetky tweety, ktoré retweetujú rovnaké tweety ako Dayner. Cez vzťah POSTS som na dané tweety naviazal accounty. Potom som cez size a sum zistil, ktoré accounty majú najviac tweetov, ktoré retweetujú rovnaké tweety ako Dayner. Podľa tohoto súčtu som potom zoradil výsledok a limitoval ho na 10, čo ale nebolo potrebné, keďže mi vyšlo iba 7 accountov. Query a výsledok v priloženom screenshote:

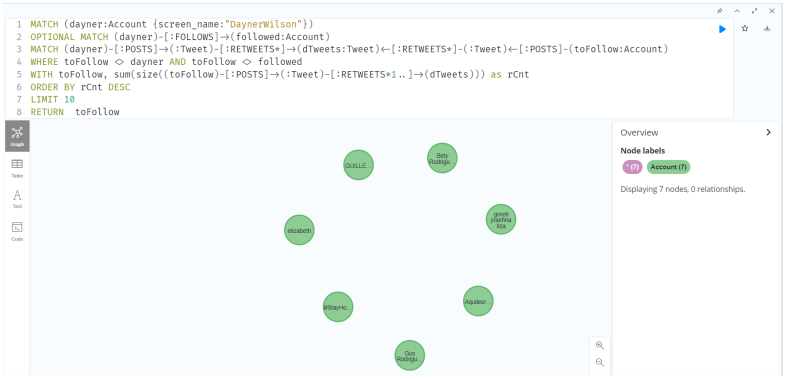

Figure 9: Noví kamaráti pre Daynera

8 Vyhľadajte 5 tweetov ostatných Accountov, ktoré do hĺbky 5 followujú account, ktorý napísal tweet (id: ”1289380305728503808”), ktoré síce nie sú retweetom vybraného tweetu, ale napriek tomu majú čo najviac rovnakých slov v poli content zhodných s vybraným tweetom (stačí rozdeliť content na slová cez split(tweet.content, ” ”). Account, ktorý followuje Account, ktorý followuje nami vybraný Account rozumieme hĺbkou 2. Odporúčam pozrieť si procedúry v knižnici APOC pracujúce s collections, ale nie je to podmienkou na zvládnutie úlohy.

Prvým krokom bolo nájsť accounty, ktoré cez vzťah FOLLOWS followujú autora nášho tweetu. Následne som zo všetkých tweetov, ktoré sú na dané accounty naviazané cez vzťah POSTS odfiltroval všetky, ktoré sú vo vzťahu RETWEETS s pôvodným tweetom. Potom som cez funkciu split rozdelil aj pôvodný tweet aj vyfiltrované tweety na listy slov, (splitter ” ”). Potom som pomocou apoc.col.intersection a size() vyrátal splitScore - intersection vracia pole, ktoré je prienikom dvoch polí a size vracia dĺžku poľa, čiže splitScore bol počet rovnakých slov. Následne som už len zoradil tweety zostupne podľa splitScore a limitoval výsledok na 5. Query a výsledok priložené v screenshote nižšie:

Figure 10: Tweety s najlepším splitScore

9 BONUS: Nájdite najkratšie cesty medzi Katy Perry (katyperry) a Donaldom Trumpom cez vzťah RETWEETS (a tým pádom aj POST). Všetky cesty, kde Katy Perry retweetla post Accountu, ktorý retweetol post Accountu, ktorý..., ktorý retweetol post Donalda Trumpa.

Cesty som našiel pomocou apoc.path.expandConfig. Do relationshipFiltera som dal POSTS a RETWEETS (museli ostať obojstranné). Ako terminatingNode som dal Donalda, lebo k nemu sme sa chceli dostať. Postupne som nastavoval maxLevel na číslo, ktoré dalo aspoň 2 výsledky rôznej dĺžky aby som si vedel overiť funkčnosť zvyšku query. V produkcii by toto bolo byť veľké číslo alebo by sa ten atribút omitol a tým pádom by tam nebol limit. Toto mi vrátilo zoznam všetkých ciest do danej dĺžky. Cez min som zistil dĺžku najkratšej cesty a následne som znovu použil apoc.path.expandConfig s tým, že maxLevel bolo predtým získané minimum.

```
MATCH (katy:Account {screen_name: 'katyperry'}), (donald:Account {screen_name: 'realDonaldTrump'})
CALL apoc.path.expandConfig(katy, {
    relationshipFilter: "POSTS|RETWEETS",
    minLevel:1,
    maxLevel:15,
    terminatorNodes: [donald]
})
YIELD path
WITH katy,donald,path,length(path) as hops
WITH katy,donald,min(hops) as minHops
CALL apoc.path.expandConfig(katy, {
    relationshipFilter: "POSTS|RETWEETS",
    minLevel:1,
    maxLevel:minHops,
    terminatorNodes: [donald]
})
YIELD path
RETURN path
```

Figure 11: Celá volaná query

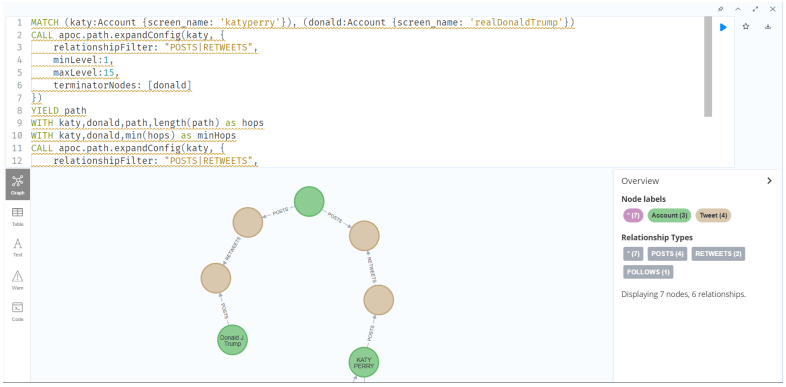

Figure 12: Výsledok volania query, najkratšia bola len 1 cesta a mala 6 hopov (relationshipov)