Do kedy?

Keďže všetci robíte radi s predstihom, deadline je 31.12.2021 o 23:59.

Čo odovzdať?

Dotazy v separátnom textovom editore + dotaz a screenshot výsledku v pdf.

Hodnotenie

Úlohy spolu za 10 bodov rozdelené nasledovne:

- 1,2,3,4,5 a 6 sú za jeden bod,
- 7, 8 za 2 body.
- Bonus za 2 extra body.

Zadanie – Neo4j

V odpovediach nepoužívajte na zoraďovanie existujúce polia followers_count, retweet_count, etc. Tieto hodnoty sú prevzaté z Twitteru a nezodpovedajú našim vzťahom v datasete. Pri počítaní početnosti followerov, retweetov a podobne vždy počítajte zodpovedajúce vzťahy medzi uzlami.

1. Zoberte si nami vytvorený dataset z vašich tweetov: https://drive.google.com/file/d/1kufTAFP5DH8hiCT-BRLHqAicoulJ_0Qi/view?usp=sharing a importujte ho cez neo4j-admin

2. Vypíšte 5 Accountov s najvyšším množstvom followerov. Nezoraďujte Accounty podľa poľa followers_count. Zaujímajú nás followujúce Accounty v našom datasete cez vzťah FOLLOWS.

3. Nájdite najkratšie cesty medzi Katy Perry {screen_name: ‘katyperry‘} a Kim Kardashian {screen_name: ‘KimKardashian‘} cez vzťah FOLLOWS. Všetky cesty, kde Kim Kardashian followuje niekoho, kto followuje niekoho, kto..., kto followuje Katy Perry.

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

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.

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.

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.

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 follow uje nami vybraný Account rozumieme hĺbkou 2. Odporúčam pozrieť si procedúry v knižnici APOC pracujúce s collections, ale nie je to podmienkou na zvládnutie úlohy.

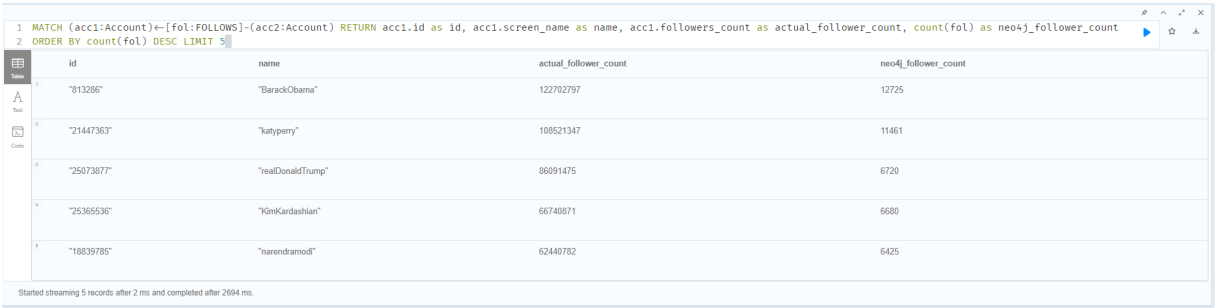
BONUS: Nájdite najkratšie cesty medzi Katy Perry (katyperry) a Donaldom Trumpom cez vzťah RETWEETS (a tým pádom aj POST). Všetky cesty, kde Katy Perry retweetla post Accountu, ktorý retweetol post Accountu, ktorý..., ktorý retweetol post Donalda Trumpa.

(Account) -> [POSTS] -> (Tweet) – [RETWEETS] -> (Tweet) <- [POSTS] – (Account) – [POSTS] -> (Tweet) – [RETWEETS] -> (Tweet) <- [POSTS] - (Account) -> …

Github: https://github.com/jaruji/neo4j-PDT

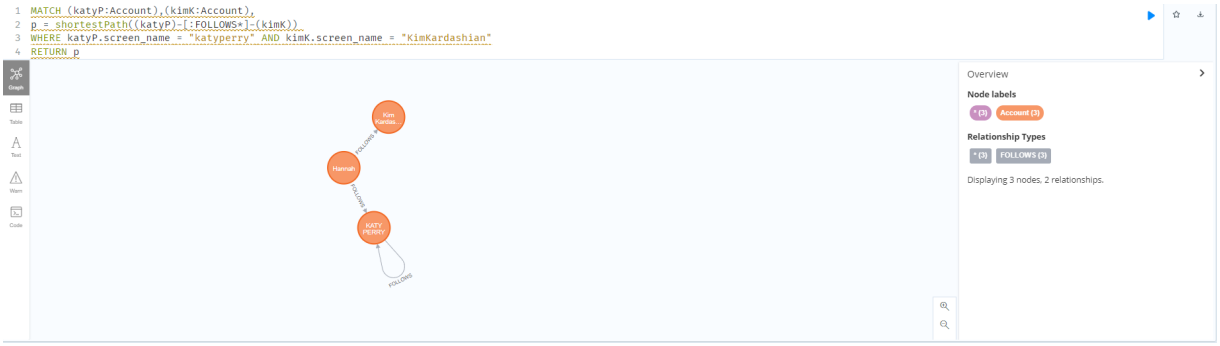
1. Neo4j sa mi podarilo rozchodiť v dockeri, dump som musel definovať cez docker-compose.

2. Query je zapísané aj v separátnom textovom súbore, výsledok vyšiel nasledovný:



Vypísal som si pre zaujímavosť aj pôvodný follower_count (zoraďoval som však podľa countu follow vzťahov z neo4j db).

3. Zistil som prostredníctvom query, že aká je najkratšia cesta medzi katyperry a KimKardashian -



Dĺžka cesty je 1, keďže existuje aspoň jeden účet ktorý followuje aj katyperry aj kimK. Keďže v zadaní máme nájsť všetky cesty, použijeme funkciu allShortestPaths namiesto shortestPath. Toto query vráti dosť výsledkov, graf mi niekam úplne zmizol.

Started streaming 304 records after 1 ms and completed after 452 ms.

Pridávam ukážku zopár prvých výsledkov:

```
"p"
[{"friends_count":224,"screen_name":"katyperry","statuses_count":10881,"followers_count":108521347,"name":"KATY PERRY","description":"Love. Light.","id":"21447363"},{},{"friends_count":288,"screen_name":"MeltyTeeth","statuses_count":4907,"followers_count":209,"name":"Reggie•Unus•Septicshit✿","description":"✨Hi it's me, Your Prince✨","id":"1278892576897105920"},{"friends_count":288,"screen_name":"MeltyTeeth","statuses_count":4907,"followers_count":209,"name":"Reggie•Unus•Septicshit✿","description":"✨Hi it's me, Your Prince✨","id":"1278892576897105920"},{},{"friends_count":127,"screen_name":"KimKardashian","statuses_count":33006,"followers_count":66740871,"name":"Kim Kardashian West","description":"Shop @kkwfragrance Diamonds II by Kourtney x Kim x Khloé now  & @SKIMS Sleep Naked Tuesday, 09.22 at 9AM PT","id":"25365536"}]
[{"friends_count":224,"screen_name":"katyperry","statuses_count":10881,"followers_count":108521347,"name":"KATY PERRY","description":"Love. Light.","id":"21447363"},{},{"friends_count":220,"screen_name":"SethWade","statuses_count":333,"followers_count":50,"name":"Seth Wade","description":"Looking for my music? Try facebook under Seth Wade. I use Twitter to shine light on injustice and support the voices of those I want heard.","id":"1266423607"},{"friends_count":220,"screen_name":"SethWade","statuses_count":333,"followers_count":50,"name":"Seth Wade","description":"Looking for my music? Try facebook under Seth Wade. I use Twitter to shine light on injustice and support the voices of those I want heard.","id":"1266423607"},{},{"friends_count":127,"screen_name":"KimKardashian","statuses_count":33006,"followers_count":66740871,"name":"Kim Kardashian West","description":"Shop @kkwfragrance Diamonds II by Kourtney x Kim x Khloé now  & @SKIMS Sleep Naked Tuesday, 09.22 at 9AM PT","id":"25365536"}]
[{"friends_count":224,"screen_name":"katyperry","statuses_count":10881,"followers_count":108521347,"name":"KATY PERRY","description":"Love. Light.","id":"21447363"},{},{"friends_count":179,"screen_name":"augustodias","statuses_count":37515,"followers_count":183,"name":"Augusto (Guga)Dias","description":"Augusto Dias, advogado especialista em propriedade intelectual, budista, é palestrante, escritor, poeta.","id":"24171057"},{"friends_count":179,"screen_name":"augustodias","statuses_count":37515,"followers_count":183,"name":"Augusto (Guga)Dias","description":"Augusto Dias, advogado especialista em propriedade intelectual, budista, é palestrante, escritor, poeta.","id":"24171057"},{},{"friends_count":127,"screen_name":"KimKardashian","statuses_count":33006,"followers_count":66740871,"name":"Kim Kardashian West","description":"Shop @kkwfragrance Diamonds II by Kourtney x Kim x Khloé now  & @SKIMS Sleep Naked Tuesday, 09.22 at 9AM PT","id":"25365536"}]
[{"friends_count":224,"screen_name":"katyperry","statuses_count":10881,"followers_count":108521347,"name":"KATY PERRY","description":"Love. Light.","id":"21447363"},{},{"friends_count":1036,"screen_name":"CUinNov","statuses_count":15516,"followers_count":503,"name":"trump thinks my brother is a LOSER and a SUCKER","description":"Tweeting is not leading!  To stop the trump crime family, those we elected have to STEP UP and DO SOMETHING.","id":"711083972969627648"},{"friends_count":1036,"screen_name":"CUinNov","statuses_count":15516,"followers_count":503,"name":"trump thinks my brother is a LOSER and
```

4. Úlohu som spravil cez subquery prostredníctvom CALL funkcie. Najprv si vyberieme 10 najviac retweetovaných účtov, následne nad touto množinou hladáme 10 najmenej retweetovaných tweetov. Selectoval som zaujímavé hodnoty ako názov účtu, tweet_id, tweet content, počet retweetov účtu a počet retweetov samotného tweetu.

```
CALL
{
    MATCH (acc:Account)-[p:POSTS]→(t:Tweet)←[ret:RETWEETS]-()
    RETURN acc, count(ret) as acc_retweet_count ORDER BY acc_retweet_count DESC LIMIT 10
}
MATCH (acc)-[p:POSTS]→(t:Tweet)←[ret:RETWEETS]-()
RETURN acc.screen_name as name, acc_retweet_count, t.id as tweet_id, t.content as content, count(ret) as tweet_retweet_count ORDER BY tweet_retweet_count ASC
LIMIT 10
```

name	acc_retweet_count	tweet_id	content
"ewarren"	1883	"1287140698131750913"	"Last night, the federal evictions moratorium expired, and rent is due next week—the same week coronavirus unemployment benefits are set to end. This is a completely preventable crisis. Congress must act immediately to extend these critical protections. https://t.co/9gALMcbR3k"
"CarlosLoret"	1486	"1289243214428303360"	"Este estudio en el @washingtonpost dice que para reabrir las universidades habría que hacer una prueba de Covid cada dos días a los alumnos https://t.co/69T4wlguXG"
"replouiegohmert"	1613	"1288554276231548929"	"Lots of #FakeNews going around about this https://t.co/OQhHYZJZvh"
"ewarren"	1883	"1289255470293757955"	"We know what we need to do to contain the virus and save lives and our economy—but Republicans refuse to invest enough in widespread testing and contact tracing. Trump and his Republican buddies don't have what it takes to get us out of this crisis. https://t.co/1CDbH26bmu"
"CarlosLoret"	1486	"1288962040401399815"	"Para morirse de envidia: cómo Francia está viviendo su nueva normalidad. Este articulista sale a restaurantes, va a conciertos y a centros comerciales. El truco: pruebas y rastreo de contactos. https://t.co/XcVG6T7E0P"

Started streaming 10 records after 70 ms and completed after 935 ms.

Výsledky:

"name"	"acc_retweet_count"	"tweet_id"	"content"	"tweet_retweet_count"
"ewarren"	1883	"1287140698131750913"	"Last night, the federal evictions moratorium expired, and rent is due next week—the same week coronavirus unemployment benefits are set to end.\n This is a completely preventable crisis. Congress must act immediately to extend these critical protections. https://t.co/9gALMcbR3k"	1
"CarlosLoret"	1486	"1289243214428303360"	"Este estudio en el @washingtonpost dice que para reabrir las universidades habría que hacer una prueba de Covid cada dos días a los alumnos https://t.co/69T4wlguXG"	2
"replouiegohmert"	1613	"1288554276231548929"	"Lots of #FakeNews going around about this https://t.co/OQhHYZJZvh"	3
"ewarren"	1883	"1289255470293757955"	"We know what we need to do to contain the virus and save lives and our economy—but Republicans refuse to invest enough in widespread testing and contact tracing.\n Trump and his Republican buddies don't have what it takes to get us out of this crisis. https://t.co/1CDbH26bmu"	3
"CarlosLoret"	1486	"1288962040401399815"	"Para morirse de envidia: cómo Francia está viviendo su nueva normalidad. Este articulista sale a restaurantes, va a conciertos y a centros comerciales. El truco: pruebas y rastreo de contactos. https://t.co/XcVG6T7E0P"	6

"ewarren"	1883	"1289287273880547328"	"We need to make sure schools have all the resources they need to determine whether and how to safely reopen. Anything less is recklessly endangering lives for political gain. \nhttps://t.co/XVwGQ7yRF4"	10
"CarlosLoret"	1486	"1289354415372025858"	"688 fallecimientos documentados en 24 horas, ya son 46 mil 686 decesos por #Covid en #México. https://t.co/gh6x8hYCt2"	12
"CarlosLoret"	1486	"1289417499654541312"	"Murió Paco Valverde, un gran luchador por la naturaleza, valiente defensor de la Vaquita Marina. Hubo una enorme solidaridad para tratar de salvarlo. Gracias a todos los que estuvieron pendientes. Descanse en Paz el buen pescador. Abrazo entrañable para Alan y toda su familia. https://t.co/4nzvIbqto"	50
"maddieevelasco"	3143	"1288978547898322945"	"2. You put employees at risk for getting sick. Yes we wear a mask, but we are there to serve you and have families and friends we are afraid to be around now because we don't know how long ago we came in contact with someone or if we are infected until it's too late."	54
"maddieevelasco"	3143	"1288978061501722625"	"1. You put yourself at unnecessary risk of contracting covid-19. We sanitize as often as we can, but you still take your mask off to eat around strangers and you don't know where they have been or who they have been in contact with."	66

5. V tejto úlohe bolo potrebné najprv nájsť najviac retweetovaný tweet účtu Donalda Trumpa prostredníctvom subquery (CALL rovnako ako predošlá úloha). Následne pomocou CREATE môžeme vytvoriť nový účet, ktorému cez vzťah FOLLOWS pridáme follow na Trumpov účet a cez vzťah POSTS vytvoríme nový tweet s ľubovoľným obsahom.

```
CALL{
    MATCH (acc1:Account)-[p:POSTS]→(t:Tweet)←[ret:RETWEETS]-()
    WHERE acc1.screen_name = 'realDonaldTrump'
    RETURN acc1, t, count(ret) as neo4j_retweet_count
    ORDER BY count(ret) DESC
    LIMIT 1
}
CREATE(acc2:Account {screen_name: 'jaruji', name: 'Juraj'})-[f:FOLLOWS]→(acc1), (acc2)-[p:POSTS]→(myT:Tweet {content: "Hello TRUMP"})←[r:RETWEETS]-
()
RETURN acc1.screen_name, neo4j_retweet_count, t.id, t.content, acc2.screen_name, myT.content
```

acc1.screen_name	neo4j_retweet_count	t.id	t.content	acc2.screen_name	myT.content
"realDonaldTrump"	487	"1289260338135699456"	"Great job by Jim Jordan, and also some very good statements by Tony Fauci. Big progress being made! https://t.co/8Oeca9H3yq"	"jaruji"	"Hello TRUMP"

Added 2 labels, created 3 nodes, set 3 properties, created 3 relationships, started streaming 1 records after 60 ms and completed after 235 ms.

6. Cez CALL som si vybral účet používateľa 777stl a účty ktoré followuje. Následne som hľadal iné účty ktoré followujú tieto účty a počítal prekryv, resp. početnosť zhodných účtov ktoré followujú s prvotne vybraným účtom. Bolo potrebné ošetriť aj situáciu, aby sme nenašli rovnaký účet (cez acc1 <> acc2).

```
CALL{
    MATCH (acc1:Account)-[fol1:FOLLOWS]→(accounts:Account)
    WHERE acc1.screen_name = '777stl'
    RETURN acc1, accounts
}
MATCH (acc2:Account)-[fol2:FOLLOWS]→(accounts)
WHERE acc1 <> acc2 AND NOT (acc1)-[:FOLLOWS]→(acc2)
RETURN acc2.screen_name as screen_name, count(fol2) as identical_follow_count
ORDER BY count(fol2) DESC
LIMIT 10
```

"screen_name"	"identical_follow_count"
"Aa43677901"	2
"TuiteraActiva1"	1
"Steelersgirl690"	1
"acacciatore94"	1
"jassar_gurjant"	1
"moesnaps78"	1
"ACrowdedPlanet"	1
"heruyaheru"	1
"han0110if"	1
"Anonymous5426"	1

MAX COLUMN WIDTH:

7. Pomocou CALL som si vybral všetky tweety ktoré retweetuje používateľ DaynerWilson, následne som hľadal zhodu prostredníctvom tejto predom vybranej množiny tweetov. Výsledkov mi našlo iba 7.

```
CALL{
    MATCH (acc1: Account)-[p:POSTS]→(:Tweet)-[ret1:RETWEETS*]→(t:Tweet)
    WHERE acc1.screen_name = 'DaynerWilson'
    RETURN acc1, t
}
MATCH (acc2: Account)-[p:POSTS]→(:Tweet)-[ret2:RETWEETS*]→(t)
WHERE acc1 <> acc2
RETURN acc2.screen_name, count(ret2) as identical_retweets_count
ORDER BY count(ret2) DESC
LIMIT 10
```

"acc2.screen_name"	"identical_retweets_count"
"Aquilesreyes9"	3
"GusRodr05589737"	3
"GoretiLiza"	3
"BetyRod50219672"	2
"CCARSOLIOO"	1
"elizagomalcala"	1
"viralvideovlogs"	1

8. V prvom rade bolo potrebné spojazdniť APOC, keďže treba mat v plugins .jar file inak tie funkcie nezbehnú. Nainštaloval som to cez docker CLI príkazom wget na najnovšiu APOC verziu a následne reštartom neo4j. Pri úlohe som postupoval nasledovne: najprv som si vybral tweet, všetky účty ktoré followujú účet ktorý tweetol tweet do hĺbky 5 (neviem či mám dávať 0-4 alebo 1-5 hĺbku podľa toho čo je napísane v zadaní - dával som 1-5 ale skúšal som aj 0-4 a výsledky boli trochu odlišné, 1-5 našlo o jednu 6 slovnú zhodu viac) a všetky tweety týchto účtov, ktoré nie sú retweetom nášho pôvodného tweetu. Nasledovne som cez APOC funkciu intesection vyhľadával počet zhodných slov medzi dvoma listami (získanými cez split s parametrom “ “) a podľa počtu zhôd som následne našiel najpodobnejšie tweety. Najlepší match bol 6 slov.

```
MATCH (acc1:Account)-[:POSTS]→(t1:Tweet),
(acc2:Account)-[:FOLLOWS*1..5]→(acc1),
(acc2)-[:POSTS]→(t2:Tweet)
WHERE t1.id = '1289380305728503808' AND NOT (t2)-[:RETWEETS*]→(t1)
RETURN t1.content as tweet1, t2.content as tweet2, apoc.coll.intersection(split(t1.content, " "), split(t2.content, " ")) as matching_words,
size(apoc.coll.intersection(split(t1.content, " "), split(t2.content, " "))) as matching_words_count
ORDER BY matching_words_count DESC
LIMIT 5
```

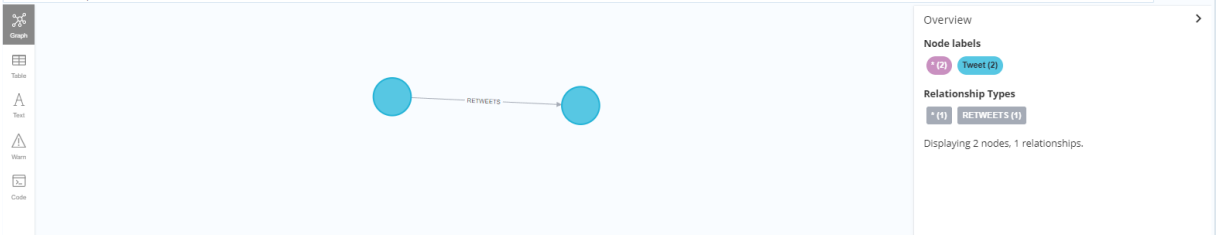
"tweet1"	"tweet2"	"matching_words"	"matching_words_count"
"@JadeRhinos @eturleye @dougducey @SenMcSallyAZ Every decision they have made regarding COVID has been economically driven, which has allowed the virus to ravage our communities, and yet, when it came down to actually helping our economy, they adandonded our state's unemployed."	"@PolitiBunny From what I've read, the teachers will be much more likely to get COVID from other staff than they would from a pupil. \n\n22 | | countries in Europe have schools open w/ no big problems. Day care facilities have been open with no big problems."	["been","COVID","have","they","the","to"]	6
"@JadeRhinos @eturleye @dougducey @SenMcSallyAZ Every decision they have made regarding COVID has been economically driven, which has allowed the virus to ravage our communities, and yet, when it came down to actually helping our economy, they adandonded our state's unemployed."	"@uaccn @SafeReturnUA @UofAlabama This law only applies to people actually infected by COVID, and not to the audience of your heinous email, which referenced "alternative arrangements" for day care and the like. So no. That wasn't your intent, or your email would have explicitly said "if you get sick, etc""	["and","have","which","the","actually","to"]	6
"@JadeRhinos @eturleye @dougducey @SenMcSallyAZ Every decision they have made regarding COVID has been economically driven, which has allowed the virus to ravage our communities, and yet, when it came down to actually helping our economy, they adandonded our state's unemployed."	"RT @natasharoyy: my mom's friend in India called to tell her that her husband insulted the dinner she made & called it bland and tastel ess,…"	["and","made","it","the","to"]	5

"@JadeRhinos @eturleye @dougducey @SenMcSallyAZ Every decision they have made regarding COVID has been economically driven, which has allowed the virus to ravage our communities, and yet, when it came down to actually helping our economy, they adandonded our state's unemployed."	"RT @natasharoyy: my mom's friend in India called to tell her that her husband insulted the dinner she made & called it bland and tastel ess,…"	["and","made","it","the","to"]	5
"@JadeRhinos @eturleye @dougducey @SenMcSallyAZ Every decision they have made regarding COVID has been economically driven, which has allowed the virus to ravage our communities, and yet, when it came down to actually helping our economy, they adandonded our state's unemployed."	"RT @dougmar_: the United States is the only country that still have a corona virus problem and Trump wants to focus on tik tok?"	["virus","and","have","the","to"]	5

BONUS:

Podarilo sa mi nájsť najkratšiu cestu, no nevedel som ako použiť shortestPath ak chceme mať komplexnú sekvenciu relationships, preto som vyhľadával medzi množinou tweetov katyP a donaldaT najkraťsiu cestu cez retweets vzťah. DĺžKa cesty bola 1 (katyP retweetla účet ktorý retweetol donalda s obsahom tweetu “Cool Retweet donaldovho Tweetu). Najkratšia cesta bola iba jedna.

```
CALL{
    MATCH (donaldT:Account)-[:POSTS]→(dtt:Tweet)
    WHERE donaldT.screen_name = 'realDonaldTrump'
    return donaldT, dtt
}
CALL{
    MATCH (katyP:Account)-[:POSTS]→(:Tweet)-[:RETWEETS*]→(t:Tweet)←[:POSTS]-(:Account)-[:POSTS]→(kpt:Tweet)
    WHERE katyP.screen_name = 'katyperry'
    return katyP, kpt
}
MATCH p = AllShortestPaths((dtt)-[:RETWEETS*]-(kpt))
RETURN p
```




```
"p"
[{"happened_at_str":"2020-03-28 02:29:57+00","favorite_count":122342,"
id":"1243726993537073152","author_id":"25073877","content":"I love Mic
higan, one of the reasons we are doing such a GREAT job for them durin
g this horrible Pandemic. Yet your Governor, Gretchen “Half” Whitmer i
s way in over her head, she doesn’t have a clue. Likes blaming everyon
e for her own ineptitude! #MAGA","retweet_count":23858},{},{"content":
"Cool Retweet donaldovho tweetu"}]
```