FIT5137 Individual Assignment - Sem 2/2019 (Weight = 15%)
Due date: Week 12, Wednesday 23-Oct-2019, 11:55pm

A. General Information and Submission

- This is an individual assignment.
- *Submission method*: Submission is online through Moodle.
- *Penalty for late submission*: 10% deduction for each day.
- *Assignment Coversheet*: You will need to sign the assignment coversheet.

B. Background – MonashBnB

MonashBnB is a new residential service that offers short time lodging to Monash students and staff around Melbourne. They have been keeping records of the accommodation listing, hosts, and reviews manually. Due to the increasing volume of people coming in and out of the accommodations listed in MonashBnB, the management team wondered to a way that could help them be in control of the environment without too much manual hassles.

They previously requested you to build a new database system through the incorporation of MongoDB and Cassandra Technologies. However due to some budget constraints, MonashBnB has decided to only maintain one database. They recently heard about graph databases and the benefit of having such database in maintaining data with complex relationships. Therefore, they have contacted you again to build another database system to store their residential service data by using Neo4j.

C. Tasks

The assignment is divided into **FOUR** main tasks:

C.1. Database Design.

You have been provided with the following three data files:

- **The Accommodation Host Report** (host.csv).
  This report provides some information about the hosts who are registered in MonashBnB. It consists of the host ID, URL, name, verification details, host registration date (host since), location, response rate, and information whether the host is a super-host.
- **The Reviews Report** (review.csv).

This report shows the reviews left by the guests after staying in a certain accommodation. The report contains the listing ID, review ID, review date, reviewer ID, reviewer name, scores rating, and comments.

- **The Listing Report** (listing.csv).
  The listing report stores information about the existing accommodation around Melbourne that are listed in MonashBnB. The report contains the information about the listing ID, listing name, summary, listing URL, picture URL, host ID, neighbourhood area, listing address, zipcode, longitude, latitude, accommodation room type, amenities, price per night, price for extra people, minimum booking length, number of reviews, last review date, reviews per month, calculated host listings count, and availability per 365 days.

Task Requirement:

In this task, you are required to do the following:

- Create a new project called ***FIT5137_Assign2***.
- Create a new graph called ***MonashBnBgraph***.
- Identify the potential nodes and edges for ***Listing***, ***Host,*** and ***Review***. You might need to illustrate your graph to show what nodes and edges are needed to build MonashBnB database.
- Import data from the CSV files and ensure that the imported data represent your identified nodes and edges.
- Use appropriate data types while inserting/importing the data.
- Use appropriate naming convention.
- Provide some explanations on the graph design (max 100 words).

The script for each task should be kept in a .cypher file called **TaskC1**.

C.2. Queries.

As you have built a new database system for MonashBnB, the management team is asking you to check if the newly built database can answer the following queries:

1. How many reviews does “Sunny 1950s Apartment, St Kilda East” have?
2. Show all reviews in Port Phillip.
3. Can you recommend accommodations that Jerome (reviewer 4162110) has never been but Sandy & Pete (reviewer 317848) have stayed and gave ratings above 90?
4. List all accommodation names and locations that do not provide Wi-Fi.
5. Count how many times a reviewer left reviews.
6. Display a list of pairs of accommodations having more than three amenities in common.
7. Which listings do not have any review?

8. Show all hosts who have multiple listings. Display both the host details and the listing name and price.
9. What is the average price for accommodations in Melbourne neighbourhood?
10. Where are the top 5 most expensive accommodations? Display the locations, host information, and names of those accommodation.
11. How many accommodations were reviewed in 2017?
12. What are the top 10 most popular neighbourhoods based on the total average review score ratings?
13. Find hosts whose location are different from their listings. Show the host name, host location, listing name, and listing location.
14. Assuming that each accommodation only accepts two guests, calculate the price of each accommodation for four people staying for five nights. Display only the accommodation name, location, price per night, extra people charge, and total price. Rank the accommodation from the cheapest price.
15. For each listing, rank other listings that are close to each other by their locations. You will need to use the longitude and latitude to calculate the distance between listings.

You are required to provide **five additional queries** that you consider to be useful to the MonashBnB management team's operation.

Task Requirement:

In this task, you are required to do the following:

- Provide the appropriate read operations for each query. For your answer you may follow where needed the format similar to that shown in the lectures and tutorials.
- When creating each query, you have to take into consideration the efficiency of the query operation.
- Create at least **two indices** including compound index for queries that are frequently used and justify why you have chosen the fields to be the index.
- Provide the necessary code for each query.

The script for this task should be kept in a .cypher file called **TaskC2**.

Note:

- The marking criteria for this task will be based on your query variation and complexity in doing the read operation. Just attempting five additional queries does not mean that you could get full mark for this section.

C.3. Database Modifications.

The management team of MonashBnB wishes to make the following changes to the implemented MonashBnB database systems:

1. Go to AirBnB website and **add three new listings**, including the hosts details and some related reviews of the listings you chose. The IDs in this case can be assigned manually by yourself.
2. Update the host verification for those who registered in 2009 and add Facebook to the list of existing verifications.
3. Update hosts who respond "within an hour" to a superhost. For this update you may only use the "host response time" and "host is a super host" information.
4. Update hosts who do not receive any reviews for their accommodation since 2017 and add a new property called active. This new property accepts Boolean value.
5. Delete all listings with zero availability and have no reviews.

Task Requirement:
In this task, you are required to do the following:

- Create the necessary modifications given in Task C.3.
- Make sure data which are not required to be modified are not modified in any way.

The script for each task for should be kept in a .cypher file called **TaskC3**.

C.4. Advanced Topic.

There are two options in this task that you can choose. Option 1 requires you to build an accommodation recommendation system, while Option 2 requires you to do some research on Graph Database in real life. You only need to **choose one task** out of the two options specified in Task C.4.

Option 1:
In this task, you are required to build an accommodation recommendation system for MonashBnB. You have created a basic graph database for MonashBnB, however the management team wants you to include a more advanced feature to the database. The idea is to suggest accommodations that have never been booked by users (you can use reviewers for this case) based on the other user's ratings. For example, user A has never booked accommodation 1, while in the system, accommodation 2 and accommodation 3 have ratings suitable for user A. Thus in this case, we can recommend accommodation 2 and 3 to user A since he/she has never booked these two accommodations.

Task Requirement:

In this task, you are required to do the following:

- Create the accommodation recommendation system and ensure the system is running well.
- You would need to install the Graph Algorithm Plugin to work on this task.
- You may need to implement k-nearest neighbours and cosine similarity in order to create your accommodation recommendation system.
    - K-Nearest Neighbour (kNN) is one of the widely used machine learning techniques that can find the *k* number of objects which have the nearest distance from a certain query. For instance, assume that we are looking for 2 nearest stations (2-NN) from Monash University Caulfield Campus. The kNN here is 2-NN as we are only looking for 2 stations, while the query is invoked from Monash University Caulfield Campus. The answer to this query will then be Caulfield Station and Carnegie Station.
- This page provides an example of movie recommendations using the kNN and cosine similarity. You may use the page as a guide and reference for you to work on this task.

The script for this task should be kept in a .cypher file called **TaskC4**.

Option 2:

A lot of online travel accommodation booking systems, such as AirBnB and Expedia, are using Neo4j to store their accommodation data. They choose Neo4j for the easiness of maintaining connectivity between data, such as the connections between host and travellers. All the complex connections can easily be visualised through Neo4j compared to the other database types.

As you are currently developing MonashBnB's graph database, the management team has asked you to do some research on how those companies store their graph data using Neo4j. You are required to produce a report to MonashBnB management team and give some suggestions on what can be improved to the current graph database you recently built for MonashBnB.

Task Requirement:

In this task, you are required to do the following:

- Provide some reports on how Neo4j has been used to maintain travel and hospitality data in real life.
- Provide some suggestions to improve the current graph database of MonashBnB.
- This report should only contain a maximum of 750 words.

D. Submission Checklist

1. One **combined pdf file** containing all tasks mentioned above:
   - □ Cover page
   - □ A signed cover sheet
   - □ A **report** that combines all the tasks, including some screenshots of the output for each task.

2. **txt files** from the following tasks:
   - □ Task C.1 Database Design (**TaskC1**)
   - □ Task C.2 Queries (**TaskC2**)
   - □ Task C.3 Database Modifications (**TaskC3**)
   - □ Task C.4 Advanced Queries (**TaskC4**) – if you choose Option 1

   All of the above txt files must contain scripts that are run-able using Neo4j.

3. Zip all the files above (pdf from #1 above, and txt files from #2 above), and upload this zip file to Moodle.

Reference

http://insideairbnb.com/get-the-data.html

THE END