ELECTRONICS AND COMPUTER SCIENCE, UNIVERSITY OF SOUTHAMPTON

COMP1204 Coursework
Database Theory and Practice

Lecturer: Sarvapali D. Ramchurn

1 OVERVIEW

This coursework will cover Relational Algebra, ERDs, and SQL. Each part carries a percentage of the marks for this coursework (out of a 100) as detailed in Table 3.1. You should be able to complete various parts of this coursework as we go through the lectures. One or two lab sessions will be tabled to help with basic SQL statements that may be helpful to complete this coursework. The key points are:

- This Coursework counts for 15% of this module.
- The deadline[1] for submission of your report and scripts: 22^{nd} April 2016 by 5pm.
- Feedback will be given 4 weeks after the deadline.
- You are only allowed to use Unix, SQL, Lucidchart, and Latex (ShareLatex and other editors are acceptable). Use of other scripting languages or text editors will get you zero marks for the relevant sections.

For submission instructions, please Section 4 at the end of this document.

2 LEARNING OUTCOMES (LOS)

This coursework aims to achieve the following learning outcomes:

[1] Failure to submit by the deadline will incure a 10% penalty per working day. Submissions later by more than five working days will not be accepted.

- Knowledge of data storage approaches.
- Knowledge of empirical analysis techniques.
- Knowledge of Database Design including ERD diagramming and Normalisation.
- Knowledge of SQL.
- Knowledge of data management plans.

3 The Assignment

3.1 Assessment Criteria

Relational Model and ERD	30%
Relational Algebra	30%
SQL Commands	30%
Latex Report Writing	10%

Table 3.1: The weighting given to the different parts of this coursework.

3.2 Dataset

The dataset to be used for this coursework is the TripAdvisor dataset at:
`https://secure.ecs.soton.ac.uk/notes/comp1204/coursework/dataset/reviews_dataset.tar.gz`.
Download this file to a folder on your home drive (e.g., myworkspace). Extract the file using standard unix file decompression commands.

3.3 The Relational Model and Entity-Relationship Diagramming

In this part you will attempt to model the relationships you find in the TripAdvisor dataset.

EX1 Write down the relation you identify in the dataset. Ensure you include data types and the primary key. We will refer to this relation as **R1** in Section 3.5.

EX2 Write down the functional dependencies that exist in the dataset and potential candidate keys.

EX3 Normalise the relation so that the resulting relations are in BCNF.

EX4 Develop an ERD model of the relationship between Hotels and Reviews. Your ERD should include data types, associations, and primary/foreign keys. You may use LucidChart for this part.

Make sure you include all the above in your report. Expected completion time: 2 hours max.

3.4 Relational Algebra

In this part, you will use Relational Algebra to express various queries. These queries should be based on your NORMALISED version of the table. Write down these statements in your latex document in a new section.

EX5 Find all the reviews by the same user (i.e., given a user ID, return the list of all her reviews).

EX6 Find all the users with the number of reviews greater than 2 and return their name and number of hotels they reviewed for.

EX7 Find all the hotels with the number of reviews greater than 10.

EX8 Find all the hotels with overall rating greater than 3 and **average** cleanliness greater or equal to 5 (Note: use the Overall Rating attribute).

In your report, you need to ensure all the expressions above are written in the correct way and are readable. Expected completion time: 2 hours max.

3.5 SQL Queries

Your task is to transform the dataset into a set of INSERT statements that you can run to create and populate tables in an sqlite database. You will have to write and test some SQL queries that will help you extract some useful data from the database you will have created.

Pre-requisite
First, make sure you know how to activate sqlite from command line using 'sqlite3' at the prompt (see the lab instructions). Make sure you try this out and create some tables and run some queries (see w3schools.com for a few examples of how to do this).

Table Creation
You now need to complete the following steps:

EX9 In sqlite, create a table called HotelReviews that has all the attributes of relation **R1** as you've specified it in Section 3.3. Write down the statement you used to create this table in your report.

EX10 Write a Unix/Bash script that will transform your dataset into a set of INSERT statements into a file called "hotelreviews.sql". You should be able to re-use some of the scripts you used in the first coursework to do this: go through each file, extract the fields, and create INSERT statements for each review. Ensure you include the hotel ID as one of the attributes. Run your script using the command ".read hotelreviews.sql" at the sqlite prompt. This will load all the data into the DB. It is easy to test if all the data has been correctly loaded by running some SELECT statements. Provide a copy of your script in your report.

EX11 Create new tables as per the normalised version of your relations in the previous section. Write down these statements in your report.

EX12 Populate your newly created tables with the data from HotelReviews. Write down the SQL statements you use for this.

EX13 Create relevant indexes for all your tables to improve the queries that might be run on your dataset. Write down the indexes you choose and explain why you chose them in your report.

Data Retrieval and Analysis

EX14 Write down the SQL version of the queries you have written down in Relational Algebra in the previous section. Test your queries with the dataset. Write down only your queries in the report.

Expected completion time for this whole section: 10 hours max.

4 Submission Instructions

If you do not follow the submission instructions to the letter, you are very likely to lose marks.

4.1 Script Format

You are required to submit your unix script with the following properties:

1. Name it **generatesql.sh**.
2. It should only take as input the **reviews_folder** i.e., it should be called as follows: **./generatesql.sh reviews_folder**
3. The ONLY output generated by your script should be the hotelreviews.sql file.
4. Your script should NOT attempt to search the hard drive for the reviews folder. You can assume reviews_folder will be placed in the same directory as your script.
5. Your script should generate no other SQL statements than those required to create a table called HotelReviews.

You WILL lose marks if you do not follow the above instructions. If your code does not work as expected (call and output), you may lose all the marks for that part.

4.2 Report Format

You should write your report **in Latex**. The name of the generated file should be **report.pdf**. Your report should **not be more than 3 pages long** including the first page. Your report should contain:

1. A title, your name, and ID.
2. A section for each part of the coursework (i.e., ERD and Normalisation, Relational Algebra, and SQL). Make sure you create a **subsection** called EX[NUMBER] for each question you answer.
3. A "Conclusions" section explaining your approach, difficulties you faced.

The first page of your report can contain the submission info (title,name, id) as well as answers to the questions. Failure to follow the above guidelines will result in loss of marks.

4.3 Submission Format

You should upload your submission via the handin website. Your submission should be a **tar.gz** file (named comp1204.tar.gz) that contains:

1. Your 'report.pdf' file.
2. Your "generatesql.sh" file.

The files should ONLY be archived using **tar** and compressed using **gzip** to generate a tar.gz file. Using any other compression format will result in loss of marks. Uncompressing your archive should only generate the above two files. DO NOT INCLUDE your hotelreviews.sql file nor the reviews_folder. Failure to follow these conventions will result in loss of marks.