COMPUTER SCIENCE TRIPOS Part IA 75%, Part IB 50%

Thursday 6 June 2019 1.30 to 4.30

COMPUTER SCIENCE Paper 3

*Answer **one** question from each of Sections A, B and C, and **two** questions from Section D.*

*Submit the answers in five **separate** bundles, each with its own cover sheet. On each cover sheet, write the numbers of **all** attempted questions, and circle the number of the question attached.*

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

STATIONERY REQUIREMENTS
Script paper
Blue cover sheets
Tags

SPECIAL REQUIREMENTS
Approved calculator permitted

SECTION A

1 Databases

Suppose that we have a relational database with the following tables.

Table	Primary Key
`Movies(mid, title, year)`	`mid`
`People(pid, name)`	`pid`
`Genres(gid, genre)`	`gid`
`ActsIn(pid, mid)`	`pid, mid`
`HasRole(pid, mid, role)`	`pid, mid, role`
`HasGenre(gid, mid)`	`gid, mid`

In tables `ActsIn` and `HasRole`, `pid` is a foreign key into `People` and `mid` is a foreign key into `Movies`. In table `HasGenre`, `mid` is a foreign key into `Movies` and `gid` is a foreign key into `Genres`.

Note that this database is similar to, but not the same as, the examples used in lectures and the database used for practicals.

(a) For the table `ActsIn`, carefully explain what is meant by saying that `pid` is a foreign key into `People`. [2 marks]

(b) Discuss potential problems this database might suffer due to data redundancy. [2 marks]

(c) Write an SQL query that produces triples of the form `genre1`, `genre2`, `total` that count the number of movies associated with a pair of distinct genres. Each pair of genres should only appear once in the result. That is, if the triple `genre1`, `genre2`, `total` appears in the result, then the triple `genre2`, `genre1`, `total` should not. [5 marks]

(d) Suppose that `kid` is the `pid` associated with Kevin Bacon. Write SQL that returns every `pid` for actors with a Bacon number of 2. This SQL should not include views. [5 marks]

(e) Simplify the SQL of Part (d) using views. [6 marks]

2 Databases

This question develops an Entity-Relationship (ER) model for a new database. The database will be called Meta-ER because it contains Entity-Relationship models! The entities of our ER model are

entity name	description
Model	each Model represents an ER model
Entity	each Entity represents an ER entity
Relationship	each Relationship represents an ER relationship
Attribute	each Attribute represents an attribute

Each of our entities will have an **id** attribute (the primary key) and a **name** attribute. In addition, the Attribute entity will have a **type** attribute indicating the data type of the Attribute:

entity name	attributes
Model	**id**, **name**
Entity	**id**, **name**
Relationship	**id**, **name**
Attribute	**id**, **name**, **type**

(*a*) We start with one many-to-many relationship ModelHasEntity between Model and Entity that indicates which entities belong to the Model. For example, we may have a model called "MoviesModel" related to the entities presented in lecture, or a model "Trucks-R-Us" for a transportation company. ModelHasEntity is many-to-many to allow different models to share entities. Your task now is to complete this ER model and consider implementing it in a relational database.

(*b*) Define a relationship between Entity and Attribute called EntityHasAttribute. What cardinality should this relationship have? Justify your answer. [2 marks]

(*c*) Define a relationship between Relationship and Attribute called RelationshipHasAttribute. What cardinality should this relationship have? Justify your answer. [2 marks]

(*d*) Define a relationship called RelationshipRelatesEntity between Relationship and Entity. What cardinality should this relationship have? Justify your answer. [2 marks]

(*e*) Should the relationship RelationshipRelatesEntity itself have attributes? Justify your answer. Let us assume that all of our relationships are binary. [2 marks]

(*f*) Describe a relational implementation of your ER model, including keys and foreign keys. [4 marks]

(*g*) Given your relational implementation, write an SQL query that takes a model name `mname` and returns all triples `ename1`, `rname`, `ename2` where `ename1` and `ename2` are names of entities in the model `mname`, and `ename1` is related to `ename2` via the relationship with name `rname`. [8 marks]

SECTION B

3 Introduction to Graphics

A homographic transformation of a point in a 3D space is expressed as:

$$\begin{bmatrix} x' \\ y' \\ z' \\ w' \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix}$$

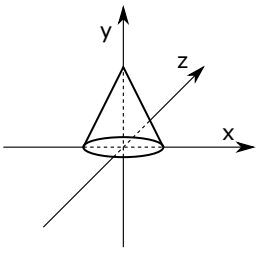
(*a*) For each of the following matrices, identify a basic transformation (translation, scaling, rotation, projection) or a sequence of them that a point will undergo when multiplied by that matrix. Name any axis of rotation and list the transformations in the correct order.

(*i*) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (*ii*) $\begin{bmatrix} \cos\theta & 0 & \sin\theta & 3 \\ 0 & 1 & 0 & 7 \\ -\sin\theta & 0 & \cos\theta & -2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

(*iii*) $\begin{bmatrix} 2\cos\theta & -\sin\theta & 0 & 0 \\ 2\sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (*iv*) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

[8 marks]

(*b*) You have a cone of height one unit; apex at $\begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$; and the base of diameter one unit centred at the origin. This cone is shown in the figure below.



You wish to transform this cone so that the apex is at $\begin{pmatrix} -1 & 2 & -3 \end{pmatrix}$, the base is centred at $\begin{pmatrix} 1 & 4 & -3 \end{pmatrix}$, and the base's diameter is two units. What transformations are required to achieve this and in what order should they be performed? Write down the product of individual transformation matrices. There is no need to multiply them together to compute the combined transformation. [12 marks]

4 Introduction to Graphics

(*a*) What are typical applications of RGB, HLS, and CIE L*a*b* colour spaces? Compare and contrast these spaces. [7 marks]

(*b*) Explain the purpose of two-step transformation from linear, scene-referred colour values to the display encoded values. [7 marks]

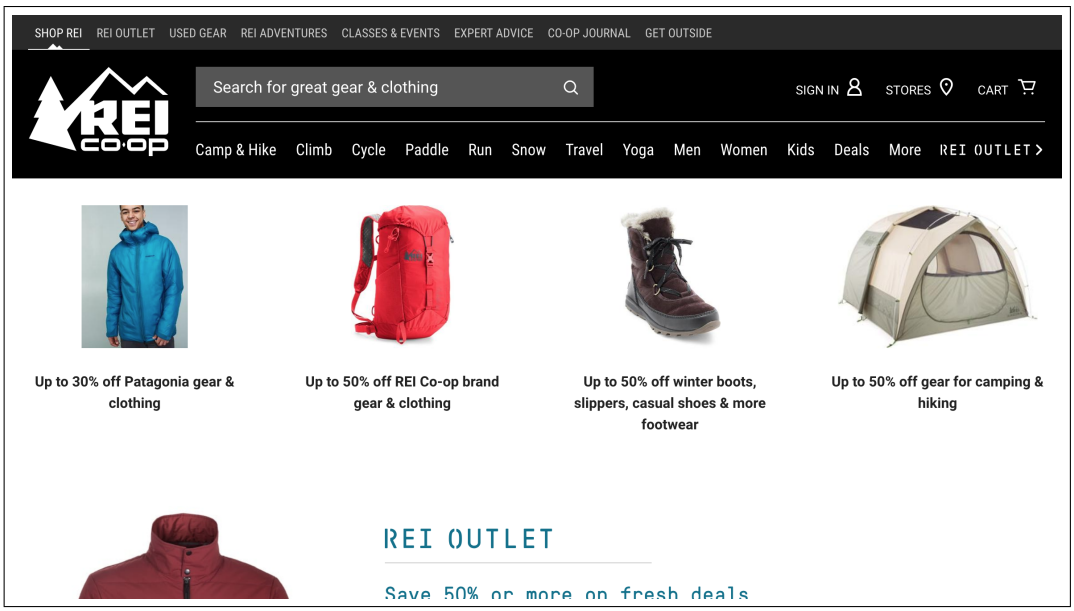
(*c*) An image is given in a *linear* ITU-R 2020 RGB colour space (display referred). Write down a sequence of equations to transform pixel values in that image into a *gamma-corrected* ITU-R 709 RGB colour space. Use the symbol $M_{2020|XYZ}$ to denote the 3×3 matrix for transforming from ITU-R 2020 to the CIE XYZ colour space; and the symbol $M_{709|XYZ}$ to denote the 3×3 matrix for transforming from ITU-R 709 to the CIE XYZ colour space. Use a standard gamma formula with $\gamma = 2.2$. [6 marks]

SECTION C

5 Interaction Design

(*a*) During your practical session you were asked to create a working app for a chosen primary stakeholder which works on a desktop or a laptop. This was done through iterative user-centred design and development. Provide a schematic description of iterative user-centred design and development. Explain which part or parts the practical sessions did not focus on and how this might have affected your working app. [6 marks]

(*b*) How would you apply Nielson's heuristics to evaluate a website for purchasing clothes which has a homepage similar to that in the figure below?



(*i*) How do the heuristics help you when looking at this homepage (or similar ones) compared to not using them?

(*ii*) Might fewer heuristics be better? Which might be combined and what are the trade-offs?

[6 marks]

[continued . . .]

(*c*) Describe what similarity analysis is in the context of card sorting and how it is conducted. Calculate the similarity rating using the data provided in the table below with four cards (1,2,3,4) and three users (A,B,C), and comment on the results.

User A	**User B**	**User C**
1,4	1,2,3	1,2,4
2,4	3,4	1,3
1	4	3,4

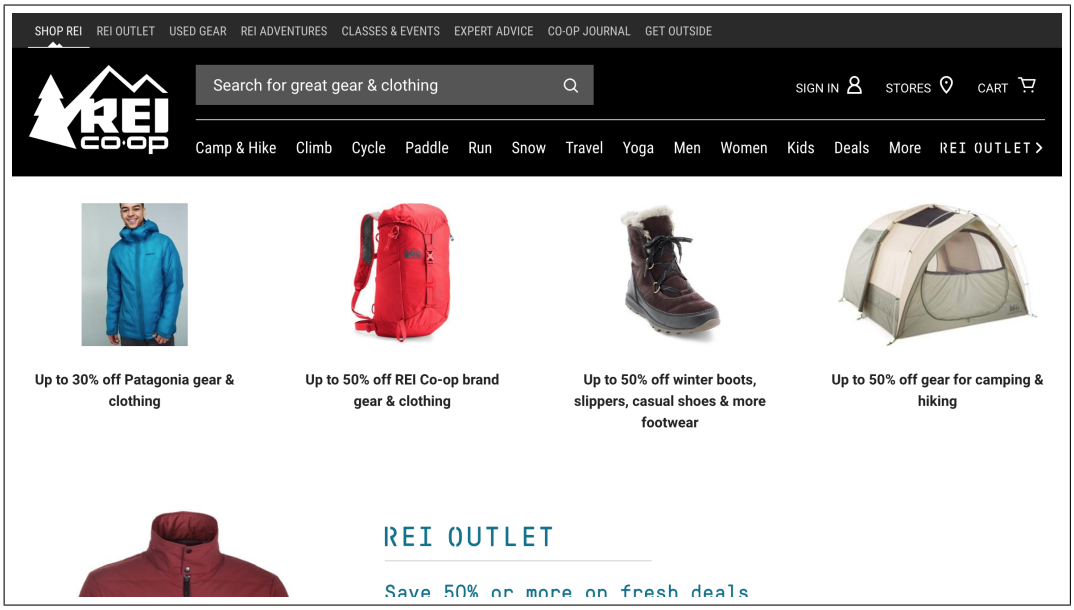
[8 marks]

6 Interaction Design

(*a*) During your practical session you were asked to create a working app for a chosen primary stakeholder which works on both a desktop and a laptop.

Describe the primary stakeholder the app was developed for, and describe three data gathering techniques your group used for the app to identify the user requirements. Explain the reasons behind this choice. [5 marks]

(*b*) Consider a website for purchasing clothing similar to that in the figure below. Would it be more appropriate to use Cognitive Walkthrough or Heuristic Evaluation to evaluate this website? Give three criteria on which to base your decision. [6 marks]



[**continued . . .**]

(*c*) What does Gestalt theory describe and what is its implication for interaction design? Describe which principle(s) are being applied for each item in the figure below, and how, and what it tells us about the interface and the interaction.

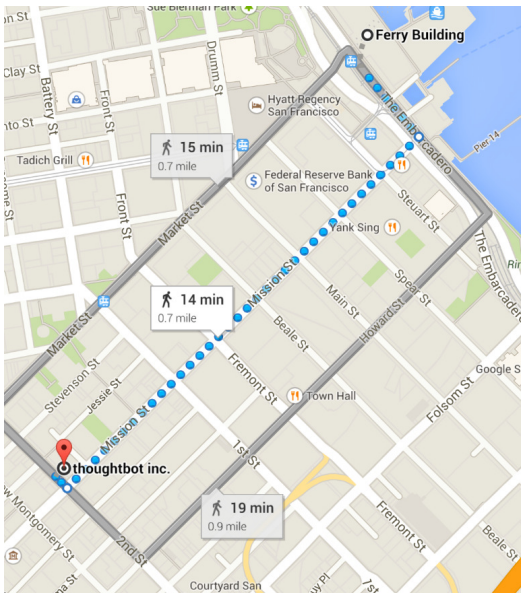
[9 marks]

(a) An interface to create a blog post.

(b) The notifications icon in Twitter's interface.

(c) Layout of Twitter's profile information.

(d) A screenshot of Google maps walking directions.



SECTION D

7 Machine Learning and Real-world Data

You want to compare the performance of two classification systems and perform a significance test on their results. You use six items as detailed in the table below, where the correct answer ("Gold Standard") and the answers of Systems 1 and 2 are listed.

System 1	System 2	Gold Standard
N	0	N
N	P	P
P	0	0
P	N	P
P	0	P
0	N	0

(*a*) Name the standard evaluation metric for classification, give its formula, and calculate its value for the two systems' results. [2 marks]

(*b*) Apply the sign test at significance level $\alpha = 0.05$ to test whether System 1 is significantly better than System 2. [5 marks]

(*c*) If instead we are testing whether Systems 1 and 2 are statistically different, how does that change your calculations in Part (*b*)? [1 mark]

(*d*) A saboteur appears in your laboratory, and creates fake versions of the results table above. She does this by swapping the values of System 1 and 2 in the same row. She decides randomly for each version how many different rows she subjects to this treatment.

(*i*) How many different fake versions of the table can be generated? [2 marks]

(*ii*) Somebody suggests that the saboteur's actions can be used as the foundation of a new statistical test, based on the idea that if the Null Hypothesis were true, this would imply that results can be randomly swapped without overall changes in the result. Explain how you can use this idea for significance testing. Illustrate how you apply the new test using the table above. [6 marks]

(*e*) The table does not contain any ties, but a high number of ties are often a reality in experiments.

(*i*) How does the presence of many ties affect the sign test? [2 marks]

(*ii*) How does it affect your newly developed test from Part (*d*)(*ii*) above? [2 marks]

8 Machine Learning and Real-world Data

You want to determine which exact crops were grown on a particular field in mediaeval times in each year. You have records of the overall yield of the field (classified into good, average and poor), but the records don't say which crop was grown. You know empirically that certain crops tend to yield more than others:

Rye (R)	good: 50%, average: 40%, poor: 10%
Beans (B)	good: 20%, average: 30%, poor: 50%
Clover (C)	good: 10%, average: 60%, poor: 30%

An historical document provides a sample of consecutive years' crops, which shows that the villagers did not keep to a strict crop rotation:

R C C R B C B B R C C C B B R C B R C B B C C C R

You want to apply a Hidden Markov Model (HMM) to the task of predicting the crop sequences for years outside of your sample.

(*a*) Define the components of an appropriate First-Order Hidden Markov Model. Estimate the missing parameters from the information given. You may assume that each of the crops is equally likely to start a sequence. Apply smoothing. Ignore the end state (treat the sequence as if it ran forever). [6 marks]

(*b*) The Viterbi algorithm can be applied to infer a sequence of crops given an observation sequence.

(*i*) State the purpose of the variable $\delta_j(t)$ in the Viterbi algorithm and give its defining equation.

(*ii*) Consider the partial observation sequence `good, good, average, ...`, with the HMM trained as above. At t=2, the following δ have been calculated: $\delta_R(1) = \frac{1}{6}, \delta_B(1) = \frac{1}{15}, \delta_C(1) = \frac{1}{30}, \delta_R(2) = \frac{2}{11 \cdot 15}, \delta_B(2) = \frac{1}{8 \cdot 15}, \delta_C(2) = \frac{1}{8 \cdot 12}$. Simulate the Viterbi algorithm at t=3, i.e., the point when `average` is encountered, showing intermediate results. [6 marks]

(*c*) Which assumptions does an HMM make? To which degree are these assumptions justified in the situation described above? [4 marks]

(*d*) In order to mitigate the effects of the potentially violated assumptions mentioned in Part (*c*), somebody suggests an increase in the order of the HMM. Do you think this would have the desired effect, and why (or why not)? [2 marks]

(*e*) You were instructed to smooth the HMM in Part (*b*) above. There is also an argument for not applying smoothing. What would happen if the estimates above were not smoothed, and what is potentially desirable about this? [2 marks]

9 Machine Learning and Real-world Data

This question concerns a sample of *English* language texts written by an author.

(*a*) When investigating a text we are concerned with its types and tokens.

(*i*) What are the types and tokens of a text? [1 mark]

(*ii*) Provide a sentence with exactly 4 types and 5 tokens. Explain any assumptions you make about the nature of tokens. [2 marks]

(*b*) Describe the expected frequency distribution of the English language types in the author's texts. Include any relevant formulas. [4 marks]

(*c*) We are interested in the written vocabulary size of the author. Could we estimate this from our sample texts? Include any relevant formulas that express the expected relationship between the size of a text and its vocabulary. [3 marks]

(*d*) Now we have another sample of English language texts written by a second author. We are interested to see if we can use a Naive Bayes classifier to automatically classify texts from the two authors.

(*i*) Define a Naive Bayes classifier for this task and describe how we use Maximum Likelihood Estimations to train the classifier. Provide equations. [4 marks]

(*ii*) Describe how the frequency distribution of types and the type/token ratio in the samples might affect the classifier. [2 marks]

(*iii*) A piece of writing has been discovered which both of our authors claim to be theirs. Could the classifier be used to settle this authorship dispute? Explain your answer. [4 marks]

END OF PAPER