UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1: DATA & ANALYSIS
COURSEWORK ASSIGNMENT 2012

Deadline: 2pm Friday 23rd March 2012

Hand in at ITO

This is the Data & Analysis resit exam from August 2011. It is being released on Friday 9th March 2012 as the 2012 Data & Analysis coursework assignment. You have **two weeks** to complete this assignment. (The original exam time was two hours.) Solutions must be handed in, on paper, at the ITO (AT Room 4.02) by **2pm Friday 23rd March 2012**. Ensure that all sheets you submit are stapled together, and that each sheet is clearly marked with your name. Your scripts will be marked, commented and returned to you in the Week 12 Data & Analysis tutorials (3–4 April).

INSTRUCTIONS TO CANDIDATES

1. ALL QUESTIONS ARE COMPULSORY.

2. DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.

1. The organisers of the *EXAM 2011* international multi-conference need to keep track of a large collection of workshops associated with the event. Initial requirements analysis brings out the following information about what needs to be recorded.

   - Each workshop has a name, and happens on a particular date — or dates, as some workshops last more than one day.
   - There are several participants, each of which may sign up to one or more workshops.
   - For each participant, it is important to record their name, email address, and the workshops which they wish to attend.
   - There are a number of meeting rooms at the conference venue, each of a fixed capacity. Meetings rooms are identified by a floor and room number.
   - Every workshop needs an allocated meeting room; where a workshop lasts for two days, it will use the same room on both days.

   (a) Draw an entity-relationship diagram suitable for representing this information, in particular the connections between participants, workshops, rooms, and dates. [*20 marks*]

   (b) For each of the following concepts give a brief description of what it means, and give an example from your ER diagram for the previous part.

   (i) Key
   (ii) Composite key
   (iii) Total participation
   (iv) Key constraint

   How is total participation shown in an ER diagram? How is a key constraint shown? [*9 marks*]

   (c) Further analysis reveals additional requirements. However, not all of these can be captured easily in an ER diagram.

   - Each workshop must have an identified organiser among the conference participants.
   - No participant may register for two workshops on the same day.
   - Every participant must register for at least one workshop.

   (i) Identify two of these which can be captured in an ER diagram.
   (ii) For those two, show the additions required to your diagram. [*6 marks*]

2. The following XML document represents some of the taught degree programmes available at the University of Edinburgh.

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE university SYSTEM "university.dtd" >
<university>
  <school>
    <name>
      The School of Informatics
    </name>
    <website>
      http://www.inf.ed.ac.uk
    </website>
    <degree code="UTAINTL" type="undergraduate">
      <name>
        BSc Artificial Intelligence
      </name>
    </degree>
    <degree code="PTMSCCMPSI1F" type="postgraduate">
      <name>
        MSc Computer Science
      </name>
    </degree>
  </school>
</university>
```

(a) Draw the XPath data model tree for this XML document. [*10 marks*]

(b) Write a DTD which specifies the format of this XML document. Assume that a complete document would include several degrees from several different schools; and that all degrees are either undergraduate or postgraduate. [*10 marks*]

(c) Write XPath expressions to return the following information from this document.

(i) The names of all schools.

(ii) The names of all undergraduate degree programmes.

(iii) The web addresses of all schools offering postgraduate degrees. [*6 marks*]

(d) It is suggested that this data would be better represented in a relational database with two tables, Schools and Degrees. Write a schema in the SQL Data Declaration Language to do this. [*8 marks*]

(e) Based on your schema, write SQL queries to answer the three queries from part (c). [*6 marks*]

3. (a) What is the *information retrieval task*? Give an example of such a task, indicating how it matches your description. [*4 marks*]

(b) The performance of an information retrieval system can be evaluated in terms of its *precision*, P, and *recall*, R. Give an English-language definition of these two terms. [*2 marks*]

(c) Precision and recall are computed as follows:

$$P = \frac{TP}{TP + FP} \qquad\qquad R = \frac{TP}{TP + FN}$$

Name and define the three values TP, FP, FN appearing here. [*6 marks*]

(d) Two retrieval systems, X and Y, are being compared. Both are given the same query, applied to a collection of 1500 documents. System X returns 400 documents, of which 40 are relevant to the query. System Y returns 30 documents, of which 15 are relevant to the query. Within the whole collection there are in fact 50 documents relevant to the query.

Tabulate the results for each system, and compute the precision and recall for both X and Y. Show your working. (You should not need a calculator for this.) [*8 marks*]

(e) Both precision and recall need to be taken into account when evaluating retrieval systems. The *F-score* is a measure which combines them using a *weighting factor* α, where high α means that precision is more important. Give the formula defining the F-score. [*2 marks*]

(f) For the example task you gave in part (a), suggest an appropriate weighting factor α. Justify your choice. [*3 marks*]