UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Monday 2nd May 2016

09:30 to 11:30

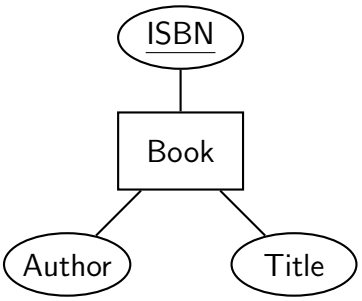
INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **CALCULATORS MAY BE USED IN THIS EXAMINATION.**

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. [*This question is worth a total of 40 marks.*]

(a) A book can be identified either by its author and title, or its ISBN (International Standard Book Number). The following diagram represents a *Book* entity for a planned relational database.



Here are four concepts from the field of relational databases. For each one, give a one-sentence description and an example using the *Book* entity above.

(i) Key
(ii) Superkey
(iii) Composite key
(iv) Primary key

[*8 marks*]

(b) The fictional online service *BitBarrow* provides a repository for shared software development. It uses a database to track various aspects of the service, including the following.

- Users, who are identified by their personal email address and each have a registered name and nickname.
- Projects, identified by a unique project title.
- Which users work on which projects. Each user can work on several different projects, and each project may have multiple contributors.
- For each project, exactly one user who is the project leader.
- Different kinds of project. Projects may optionally be declared as mobile, desktop, or server. Mobile projects need an identified platform, and desktop projects a named operating system.

Draw an entity-relationship diagram that represents this information. [*16 marks*]

QUESTION CONTINUES ON NEXT PAGE

(c) The BitBarrow database also includes information about *endorsements* where one user confirms that another user has expertise in a particular programming language. Users can also endorse themselves, to claim that they have such expertise. This is captured in the following three database tables.

```
create table User (
  email      varchar(254),
  name       varchar(200),
  nickname   varchar(200),
  primary key (email)
)
```

```
create table Endorsement (
  byUser   varchar(254),
  ofUser   varchar(254),
  forLang  varchar(100)
   ...
)
```

```
create table ProgLang (
  name varchar(100),
  primary key (name)
)
```

(i) The ProgLang table only has one column. Why might it be helpful to have a "relation" like this with just a single field?

(ii) The Endorsement table is missing primary and foreign key declarations. Write an appropriate set of declarations to complete the definition.

(iii) Write an expression in the tuple-relational calculus describing the set of all users who have been endorsed as having expertise in Haskell.

(iv) Write an expression in relational algebra to compute the nicknames of users who have endorsed someone as having expertise in C.

(v) Write an SQL command to list without duplication the names and email addresses of all users who have endorsed themselves.

[*16 marks*]

2. [*This question is worth a total of 30 marks.*]

MathML, the *Mathematical Markup Language*, is an XML dialect for describing mathematical and scientific notation. This covers both the *presentation* of mathematical expressions — how to render them on a web page or similar — and their *content* — what they mean.

Here is a small XML document using some of the content markup of MathML to describe the arithmetic expression $x + y^3$.

```
<?xml version="1.0" encoding="UTF-8"?>
<math>
  <apply>
    <plus/>
    <ci>x</ci>
    <apply>
      <power/>
      <ci>y</ci>
      <cn type="integer">3</cn>
    </apply>
  </apply>
</math>
```

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

Assume that this restricted MathML has the following conventions.

- The root math node can contain any expression.
- An expression can be either an application, an identifier, or a number; given by apply, ci and cn respectively.
- Each application is an operator followed by some arguments. The operators are plus, times and power. The arguments can be any expressions.
- Identifiers are any string; they have an optional attribute type of integer, real, or complex.
- Numbers are also strings, again with an optional type and also a base which defaults to 10 if not given.

(b) Construct a DTD that enforces these conventions. [*16 marks*]

Write XPath expressions to extract the following from a MathML expression that matches this DTD.

(c) All real-valued numbers used in the expression.

(d) Every application that involves a complex-valued identifier.

[*4 marks*]

3. [*This question is worth a total of 30 marks.*]

The non-existent start-up company *Find-a-Flick* has a website that makes film recommendations. The founders plan to monetise this some day by selling popcorn in flavours matched to individual movies through a deep-dive adaptive data-learning algorithm. For the moment, the site just suggests films based on keywords provided by a user.

For each of around 200,000 films, Find-a-Flick has a body of text built up from reviews, plot descriptions, and comments about the film. By counting occurrences of keywords in the text, the website makes recommendations of films to match user queries.

(a) The performance of an information retrieval system like Find-a-Flick can be evaluated in terms of its *recall* and *precision*. Informally, recall is the proportion of relevant results that are actually retrieved. Give a similar informal definition of precision.

(b) Which is more important for the Find-a-Flick service: recall or precision? Give a reason for your choice.

[*5 marks*]

The following table shows the keyword counts for text associated with three different films, computed to assist a search for "Exciting Scottish historical drama".

	Exciting	Scottish	Historical	Drama
Film A	30	15	0	10
Film B	2	2	4	1
Film C	0	0	4	3
Query	1	1	1	1

One way to identify which films are most relevant to the query is the *cosine similarity measure*, based on the *vector space model* of documents.

(c) Write out in full the formula for calculating the cosine of the angle α between the two four-dimensional vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

(d) Use this to rank these three films by relevance to the original query.

(e) The Find-a-Flick repository happens to have much more text about Film A than it does about either Film B or Film C — this is why the keyword counts are much higher for that film. Does this affect the ranking of Film A? If so, does it make it higher or lower? If not, why not?

[*15 marks*]

QUESTION CONTINUES ON NEXT PAGE

The Find-a-Flick founders are considering incorporating film ratings of zero to five stars into their recommender system. They want to investigate whether there is a correlation between the numerical ratings given to films on two other online services. They plan to use *Pearson's correlation coefficient* to do so. This has the following formula for the correlation coefficient $r_{x,y}$ between two samples $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ drawn from two separate larger populations — in this case, the ratings of films on two different sites.

$$r_{x,y} \;=\; \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{(n-1)s_x s_y}$$

(f) Say what the values m_x, m_y, s_x and s_y represent.

(g) Explain what it means to say that variables x and y are *positively* correlated, and what it means for them to be *negatively* correlated. How are these two types of correlation reflected in the value of Pearson's correlation coefficient?

[*10 marks*]