UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFORMATICS 1 — DATA & ANALYSIS

Deadline: 4pm Thursday 19 March 2015

Submit to box outside ITO office on Appleton Tower Level 4

This paper contains Data & Analysis exam questions from 2013 and 2014. It is being released on Thursday 5 March 2015 as a written coursework assignment. You have **two weeks** to complete this assignment. It will not necessarily take that long to complete, but the time is there to help you schedule against other assignment loads from your different courses. The original exam time was two hours.

Submit your solutions on paper to the labelled box outside the ITO office on Level 4 of Appleton Tower by **4pm Thursday 19 March 2015**. Please ensure that all sheets you submit are firmly stapled together, and on the first page write your name, matriculation number, tutor name, tutorial group number, and the course code INF1-DA. If these are not clearly stated then your work will not reach your tutor and may not be marked.

Your tutor will mark your work and return it to you in your Week 11 tutorial, with written and verbal feedback. However, these marks will not affect your final grade for Inf1-DA — this *formative* assessment is entirely for your feedback and learning. Because of this you can freely share help on the questions, ask on *Piazza* for advice, and discuss your work with other students. Please do.

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.**

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

A phone company wants to set up their own App Store for mobile devices. Requirements analysis for the controlling database highlights the following information about what must be recorded.

- Every app in the store needs a unique name, a publisher, and a rating.
- There are two subclasses of app: a *premium* app has to be paid for before installation; a *freemium* app is free to download, but has in-app purchases which cost money.
- The database should record the price of each premium app.
- Each user of the store is identified by their email address.
- A user may have several subaccounts, each identified by a nickname.
- The database needs to record which users have installed which apps.
- Users can use subaccounts to restrict access to freemium apps: the database needs to record which nicknames are allowed to run which ones.

(a) Draw an entity-relationship diagram to represent this information. [*20 marks*]

The app store groups apps into *themes* such as "Games", "News + Magazines", or "Health + Fitness". An app can be in multiple themes, and each theme can have a current "Top App". This is captured by the following SQL data declarations.

```
create table App (
  name      varchar(30),
  publisher varchar(25),
  rating    integer,
  primary key (name)
)
```

```
create table Theme (
  title    varchar(20),
  topApp   varchar(30),
  primary key (title),
  foreign key (topApp) references App(name)
)
```

```
create table InTheme (
  name  varchar(30),
  title varchar(20),
  primary key (name,title),
  foreign key (name) references App,
  foreign key (title) references Theme
)
```

(b) What do the terms "arity" and "cardinality" mean when describing database tables? [*2 marks*]

(c) Write relational algebra expressions to compute the following.

(i) The name of the top app in the "Games" theme.

(ii) For every app in the "Games" theme, its name and rating. [*6 marks*]

(d) Write expressions in the tuple-relational calculus that express the following queries.

(i) The names of all apps in the "Office" theme.

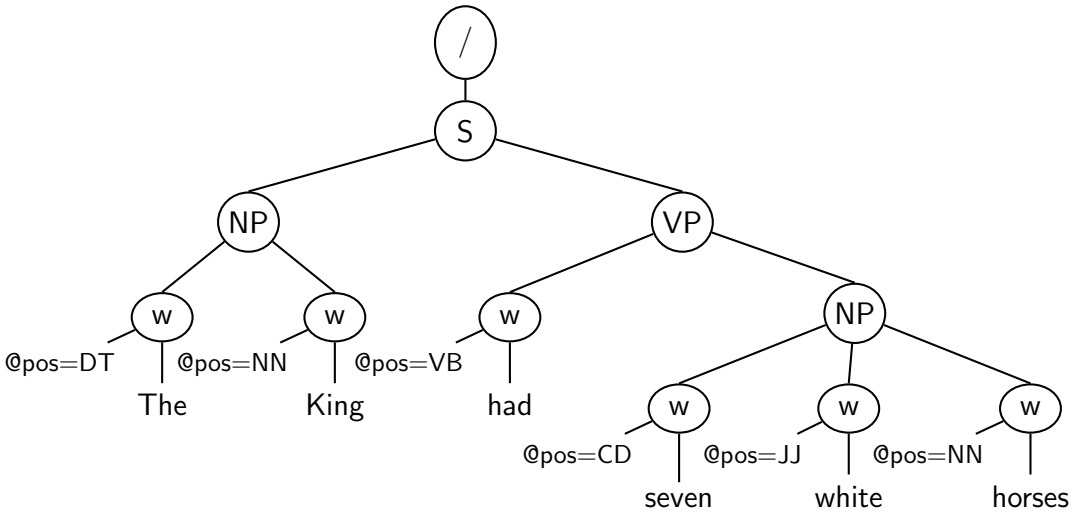
(ii) The publishers of all top apps. [*6 marks*]

(e) Write SQL queries to answer the following questions.

(i) How many apps are there in the database?

(ii) What is the highest and lowest rating given to apps in the "Utilities" theme? [*6 marks*]

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



(a) The tree above is the XPath data model for an annotated parse of the sentence "The King had seven white horses", using the following abbreviations for syntactic components and part-of-speech tags:

S	Sentence	DT	Determiner (e.g. the, a, each)
NP	Noun phrase	NN	Noun (e.g. King, horse)
VP	Verb phrase	VB	Verb (e.g. has, lives)
w	Word	CD	Cardinal number (e.g. seven)
pos	Part of speech	JJ	Adjective (e.g. large, white)
CC	Conjunction	IN	Preposition (e.g. for, or, in)

Write out the XML text form for this document. [*12 marks*]

(b) The following passage has been marked up with parts of speech, as in the Corpus Workbench used with the Corpus Query Processor (CQP).

> The/DT proud/JJ King/NN of/IN distant/JJ Spain/NN had/VB seven/CD beautiful/JJ white/JJ horses/NN . An/DT evil/JJ wizard/NN cursed/VB the/DT King/NN and/CC stole/VB the/DT horses/NN .

A *noun phrase* is a phrase of one or more words that taken together play the grammatical role of a noun. For example, this passage includes the following noun phrases:

- The proud King of distant Spain
- The proud King
- distant Spain
- seven beautiful white horses
- An evil wizard
- the King
- the horses

In the CQP syntax an expression like **[pos="NN"]** matches a single noun. Write CQP regular expressions to match the following:

(i) A sequence of two or more adjectives in a row.

(ii) A sequence of words, each of which is either a noun or a verb.

(iii) All of the noun phrases given above.

[*12 marks*]

(c) A large research corpus such as the *British National Corpus* or the *Corpus of Contemporary American English* contains hundreds of millions of words from many different sources. Building such a corpus requires balancing and sampling to ensure it is representative.

Explain briefly the meaning of *balancing*, *sampling* and *representative* as used here. [*6 marks*]

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

(a) An information retrieval system is searching a European Parliament archive for documents on the topic of "offshore fishing boundary disputes". The following document matrix indicate three possible matches.

	offshore	fishing	boundary	disputes
Document A	4	2	7	0
Document B	3	3	3	3
Document C	12	6	0	0
Query	1	1	1	1

One way to rank these documents for potential relevance to the topic is the *cosine similarity measure*.

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

Use this to rank the three documents in order of relevance to the query. [*10 marks*]

(b) One way to evaluate the performance of an information retrieval system is to assess its *precision* P and *recall* R. Informally, P can be defined as the proportion of the documents returned by the system which do match the objectives of the original search. Give a similar informal definition of R.

Here is the mathematical formula for calculating precision.

$$P = \frac{TP}{TP + FP}$$

Name and define the terms TP and FP here. Give the formula for recall R, explaining any other new terms that appear. [*8 marks*]

(c) You have been given two different information retrieval systems to compare: *Hare* and *Tortoise*. Each one is tested on the same query for a collection of 4000 documents, of which 200 are relevant to the query. *Hare* returns 1200 documents, including 150 that are relevant; while *Tortoise* returns just 160, with 120 of them being relevant.

Tabulate the results for each system and calculate their precision and recall on this test. Show your working.

One way to combine precision and recall scores is to use their *harmonic mean*. Give the formula for this, and calculate its value for each of *Hare* and *Tortoise*. [*12 marks*]