Informatics 1B Data & Analysis

Coursework assignment

Handed out: 28th February 2008

Your answers to these questions must be handed in to the **Informatics Teaching Office**, room 4.01 Appleton Tower, by **12 noon** on **Friday 7th March**. Please ensure that your name is clearly written on every page. Marked solutions will be returned in Data & Analysis tutorials in week 11.

Please answer all questions: 1(a)–(d) and 2(a)–(f).

1. An art dealer wishes to set up a database containing information on her clients, and on the paintings she buys from and sells to clients. She wishes to record the following information.

   - For each client: the name and address.
   - For each painting: an artist and a title.
   - Reference numbers for every client and painting.
   - For each painting: the client from which the dealer bought the painting (there should be exactly one such client), together with the price at which the painting was purchased.
   - For each painting sold by the dealer: the client who bought the painting together with the price they paid for it. (A painting can be sold to at most one buyer.)
   - No special information needs to be stored about artists.

   (a) Draw an ER diagram that represents the above information, and designate a primary key for each entity set. Incorporate, as far as possible, the constraints on the relationships in your diagram. Explain the conventions you use for representing constraints in the diagram. Are there any constraints that you have not been able to include in the diagram? If so, describe these.

   [10 marks]

(b) Use the SQL Data Definition Language to give relational schemata that implement the above ER diagram, using just two tables:

- `Clients`, with 3 fields: client reference; name; and address.
- `Paintings`, with 7 fields: painting reference; title; artist; client bought from; price bought at; client sold to; price sold at.

You may choose your own names for the fields, but make sure it is clear which field each name is for. Make sure that you capture all the constraints on the data.

[8 marks]

(c) Using the relational schemata defined in part (b) above, formulate each of the following two queries three times: once in relational algebra, once in tuple-relational calculus, and once in SQL.

i. Find the names of all clients who have bought a Picasso.

ii. Find the names of all clients who have sold paintings to the dealer, but not bought paintings from the dealer.

[24 marks]

(d) Formulate the following two queries in SQL.

i. Find the average amount of money paid out by the dealer when buying paintings by Picasso.

ii. Find the number of paintings bought from the dealer by Charles Saatchi.

[8 marks]

[Question 1: 50 marks total]

2. The extract below is taken from (a simplified presentation of) the British National Corpus (BNC) XML Edition.

```
<wtext type="FICTION>
<p>
<s>
<w pos="PRON">It</w> <w pos="VERB">is</w>
<w pos="ART">a</w> <w pos="SUBST">place</w>
<w pos="CONJ">that</w> <w pos="VERB">promotes</w>
<w pos="ADJ">deep</w> <w pos="SUBST">thought</w>.
</s>
...
</p>
...
</wtext>
```

We call this simplified document `"bnc.xml"`.

(a) Write a DTD to specify the XML format of the document `"bnc.xml"`. (You need only include the attribute values that apear in the example.)

[8 marks]

(b) Write path expressions to return the following lists of elements from `"bnc.xml"`.

i. All `w` elements for words whose part of speech is `VERB`.

ii. All `s` elements for sentences containing the word "thought".

[8 marks]

(c) Write an XQuery query to produce an XML document consisting of all sentences, as already annotated, that contain the word "deep". Your resulting document should be valid according to the DTD below.

```
<!DOCTYPE deepSentences [
<!ELEMENT deepSentences (s*)>
...
]>
```

The omitted part is the specification for the `s` element, which should be understood as being the same as the specification in your answer to part (a) above.

[8 marks]

(d) Write an XQuery query to remove most of the structure from the document, producing an XML document consisting of just a sequence of word elements with no annotation. Your resulting document should be valid according to the DTD below.

```
<!DOCTYPE wordsOnly [
<!ELEMENT wordsOnly (w*)>
<!ELEMENT w (#PCDATA)>
]>
```

[8 marks]

(e) In practice, corpora are constructed by starting with raw text and adding structure. Describe the various stages of processing that must be undertaken in order to go from a text file to an annotated entry in a corpus with structure similar to that illustrated in `"bnc.xml"`. What are the difficulties that arise in automating this procedure?

[12 marks]

(f) Write out all the bigrams that occur in the fragment of `"bnc.xml"` displayed above. Which, if any, are likely collocations? Justify your answer. (You should make use of your knowledge of the English language.)

[6 marks]

[Question 2: 50 marks total]