Informatics 1B Data & Analysis

Coursework assignment

4th March 2009

Your answers to these questions must be handed in to the **Informatics Teaching Office**, room 4.01 Appleton Tower, by **12 noon** on **Friday 13th 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 three questions.

1. The organisers of the Edinburgh Fringe want to set up a database to record all shows being performed, and also the venues at which they take place. They intend to record the following information.

   - For each show, the name of the show and the name of the company performing the show.
   - For each venue, the venue number, name and address.
   - A reference number for every show.
   - The venue in which each show takes place. (Each show has exactly one venue.)

   (a) Draw an ER diagram that represents the above information. Make sure that your diagram accurately reflects the constraints on the relationships. Designate a primary key for each entity set.

   [8 marks]

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

   - `Venues` — with 3 fields: venue number; venue name; address.
   - `Shows` — with 4 fields: show reference; show name; performing company; and venue.

You may choose your own field names, but make sure that it is clear which of the above fields each name refers to. Also, ensure that you capture all the constraints on the data.

[8 marks]

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

i. Find the name of every show performed at the venue called "Gilded Balloon".

[12 marks]

(d) Formulate the following two queries in SQL.

i. Count the number of different venues at which a shows entitled "Hamlet" is being performed.

ii. Find the numbers and names of all venues at which there is no show being performed.

[12 marks]

2. The XML document in Figure 1 illustrates a possible representation of scripted dialogue (for example, from a play) in a corpus.

(a) Draw the XML tree (following the XPath data model) corresponding to the document of Figure 1.

[8 marks]

(b) Write a DTD to specify the XML format for scripted dialogue illustrated by the example in Figure 1.

[8 marks]

(c) Let `"script.xml"` be an XML document containing a dialogue script in the illustrated format. (The script need not be the same as the one in Figure 1.) Write path expressions to return the following lists of elements from `"script.xml"`.

i. All `speaker` elements where the speaker is Hamlet.

```
<dialogue>
<speaker name="Polonius">
<s>
<w>What</w> <w>do</w> <w>you</w> <w>read</w>
<punc>,</punc> <w>my</w> <w>lord</w> <punc>?</punc>
</s>
</speaker>
<speaker name="Hamlet">
<s>
<w>Words</w> <punc>,</punc> <w>words</w> <punc>,</punc>
<w>words</w> <punc>.</punc>
</s>
</speaker>
</dialogue>
```

Figure 1: A fragment of scripted dialogue

ii. All `s` elements for sentences that contain a question-mark (i.e., that contain an element `<punc>?</punc>`).

iii. All words spoken by Polonius. (Return the text of the words rather than the `w` elements.)

[9 marks]

(d) The example in Figure 1 does not contain any part-of-speech information. Briefly illustrate how the XML structure of the document could naturally be extended to include such information.

[2 marks]

(e) Describe the main processes that need to be undertaken in going from the plain text of a written play to an annotated entry in a corpus that includes part-of-speech information.

[8 marks]

3. (a) The performance of an information retrieval system can be evaluated in terms of its *precision*, P, and *recall*, R. Informally, P may be defined as the proportion of those documents returned by the

system that match the original objectives of the search. Give a similar informal definition of R.

[2 marks]

(b) The mathematical formulae defining precision and recall are:

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

Name and define and the three values TP, FP, and FN appearing in these formulae.

[6 marks]

(c) Suppose a document collection contains 100 documents, 20 of which are relevant to a particular query, which is submitted to two different information retrieval systems. System A returns 5 documents, 4 of which are relevant to the query. System B returns 20 documents, 15 of which are relevant to the query. Calculate the precision and recall for each of these two systems, showing the details of your calculations.

[8 marks]

(d) Briefly explain why both precision and recall need to be taken into account when evaluating an information retrieval system, i.e., why it is not sufficient to consider just one of these values on its own.

[2 marks]

(e) Give one example of an information retrieval task for which both precision and recall are important, and say why each is important for this example. Say which of Systems A and B from question 3c should be preferred for your chosen example. Justify your answer.

[4 marks]

(f) Sometimes, the *harmonic mean* is used to calculate an *F-score* which combines precision and recall into a single value. Calculate this F-score for each of the two systems in question 3c. Which of the two systems performs better according to the F-score?

[3 marks]