Informatics 1: Data & Analysis
Session 2010–2011, Semester 2

Coursework Assignment: Past Exam Questions

This is the coursework assignment for Inf1-DA. The three questions are from recent exam papers, which covered the same syllabus as in this year's course. You should attempt all three questions, writing out your answers in full and submitting them in the coursework box outside the ITO office on Appleton Tower level 4. Please write your name, matriculation number, tutor name and tutorial group clearly on the first page of your solutions.

This coursework was distributed on Friday 25 February, at the end of week 7, which gives you two weeks to complete it. However, it is not intended to take all that time, and in previous years was issued as a one-week exercise. The earlier distribution is to help you schedule possibly overlapping assignment loads from your different courses.

Your tutor will grade your work, with written and verbal feedback, but these marks will not affect your final grade for Inf1-DA. Because this coursework is not for credit, you can freely share advice, ask your tutor about the questions, and discuss your work with other students. You are encouraged to do so.

Due date: 4pm Friday 11 March 2011 (2011-03-11 16:00Z)
Returned: Tutorial meetings on 22 and 23 March 2011 *Ian Stark*

1. (May 2010) A university wants to set up a database to record details about its staff, and the departments they belong to. They intend to record the following information.

   - For each member of staff, their staff identity number, name, job title, and salary.
   - For each department, its name, and address.
   - For each member of staff, all departments that they belong to. It is required that every member of staff belongs to at least one department.
   - For each department, the head of department. It is required that each department has exactly one head of department.

   (a) Draw an ER diagram that expresses the requirements for the database. Make sure that you capture all the constraints on the data mentioned above. [7 marks]

   (b) Are there any other natural constraints one might impose on the data that are not captured by the requirements above? For each such constraint, say whether it would be possible to modify your ER diagram to include the constraint, and, if so, explain how this would be done. [5 marks]

   (c) Use the SQL Data Definition Language to present relational schemata that implement your ER diagram. (You should implement the original diagram from question 1a, not any modification considered for question 1b.) [7 marks]

   (d) Explain which of the constraints in your ER diagram you have incorporated in your relational schemata and which you have not. [3 marks]

   (e) Using the relational schemata defined in question 1c above, formulate the following query three times; once each in relational algebra, tuple-relational calculus and SQL.

   - Find the name and salary of every head of department.

   [12 marks]

   (f) Formulate the following query in SQL.

   - Count the number of heads of department that belong to at least one department that they are not head of.

   [6 marks]

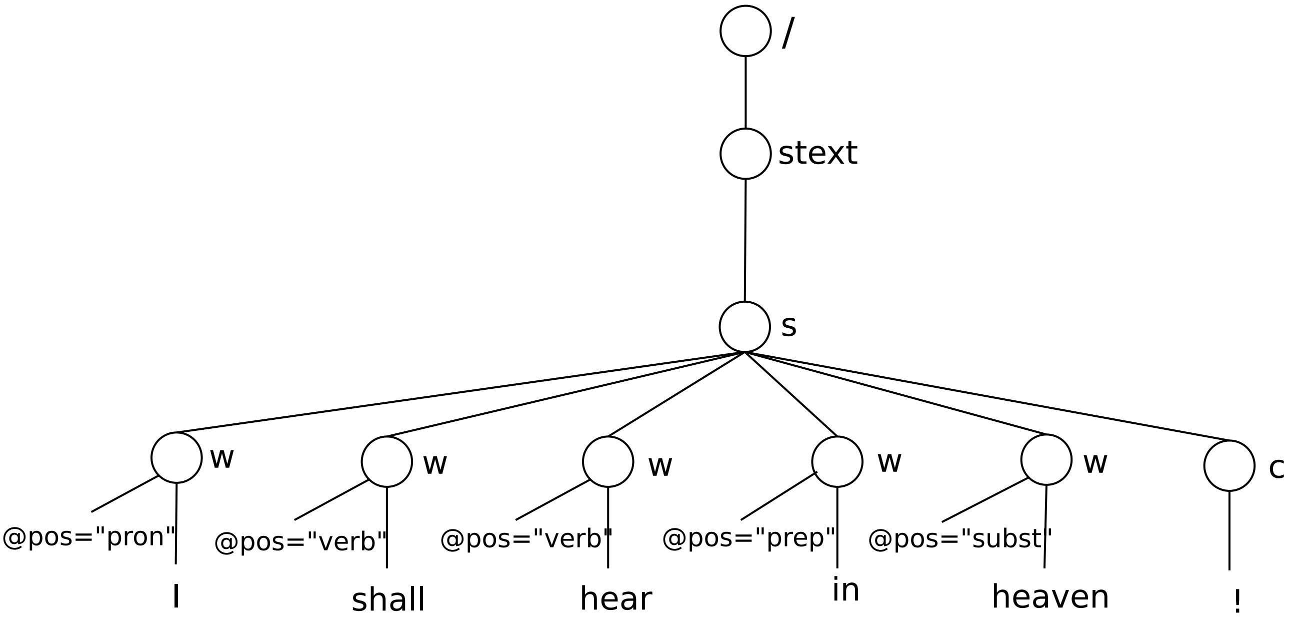

Figure 1: An XML tree

2. (May 2010) The tree in Figure 1 depicts an XML document, drawn according to the XPath data model. It represents a passage of spoken text annotated according to (a simplification of) the mark-up scheme of the British National Corpus.

   (a) Write out the tree in Figure 1 as an XML document. [8 marks]

   (b) In the tree in Figure 1, `<stext>` is an XML element that provides an annotation for marking up a passage of spoken text. For all the other annotations in Figure 1, explain both their XML status and also the meaning they carry in marking up the text. [10 marks]

   (c) Write a DTD to specify the XML structure of XML documents for spoken text in the format of Figure 1. You should allow for a passage of spoken text to possibly consist of several sentences. [8 marks]

   (d) Write XPath expressions to return the following lists of text strings from any XML document that is valid with respect to the DTD for Figure 1, and in which the annotations allowed by the DTD have been applied correctly.

      (i) All punctuation marks. [3 marks]

      (ii) All verbs. [3 marks]

      (iii) All verbs that appear in sentences that contain an exclamation mark “!”. [3 marks]

3. (August 2008)

   (a) What is the *information retrieval task*? What are the underlying assumptions on which it is based? Illustrate these assumptions using a suitably chosen example of an information retrieval task. [8 marks]

   (b) 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]

   (c) 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]

   (d) 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 1 returns 5 documents, 4 of which are relevant to the query. System 2 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 (you should not need a calculator for this). [8 marks]

   (e) 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]

   (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 3d (you should not need a calculator for this). Which of the two systems performs better according to the F-score? [4 marks]