Informatics 1B Data & Analysis
Sample Mock Exam 2008

ANSWER ALL THREE QUESTIONS

THE LENGTH OF ACTUAL EXAM IS TWO HOURS

1. A bank wishes to set up a database to record information about its bank accounts and their holders. They intend to record the following information.

   - For each account holder, their: name, address, telphone number, and date-of-birth.
   - For each account: the account number, the account holder (each account has exactly one account holder), the account type (each account has exactly one account type), and the current balance of the account.
   - For each account type: the name of the account type (e.g., "current", "savings"), the interest rate, and the overdraft limit.
   - A reference number for every account holder.

   (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 implements the above ER diagram, using three tables:

   - `Holders` — with 5 fields: reference; name; address; telephone number; and date-of-birth.
   - `AccountTypes` — with 4 fields: reference; name of account type; interest rate; and overdraft limit.
   - `Accounts` — with 4 fields: account number; account holder; account type; and current balance.

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. List, for every account owned by Gordon Brown, the account number, the account type, and the balance of the account.

[12 marks]

(d) Formulate the following two queries in SQL.

i. Find the total sum of money that Gordon Brown has stored in accounts of types "savings". (You may assume that there is only one person on the system called Gordon Brown.)

ii. List the account number and name of account holder for all accounts that are overdrawn with an overdraft exceeding the overdraft limit. (Assume that overdrawn balances are represented as negative integers, and overdraft limits are also represented as negative integers.)

[12 marks]

2. The XML document in Figure 1 illustrates how speech is represented in (a simplified version of) the British National Corpus (BNC) XML edition. In it, the root `sp` element stands for an individual speech in a performance text, e.g., a play, and the `stage` element represents a stage direction.

(a) Write a DTD to specify the XML format for speech elements illustrated by the example in Figure 1. In your DTD, you should allow the `pos` attribute to take only values that appear explicitly in Figure 1.

[8 marks]

```
<sp>
  <p>
    <s>
      <w pos="ADJ">Seven </w>
      <w pos="SUBST">books </w>
      <w pos="ART">a </w>
      <w pos="SUBST">week</w>
   </s>
 </p>
 <stage>
   <s>
      <w pos="PRON">He </w>
      <w pos="VERB">dances</w>
   </s>
 </stage>
 <p>
   <s>
      <w pos="SUBST">Library </w>
      <w pos="SUBST">books</w>
   </s>
 </p>
</sp>
```

Figure 1: A speech from the (simplified) BNC

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

i. All `stage` elements for stage directions containing at least one verb.

ii. All `w` elements for words of speech that are pos-tagged as nouns. (Note that nouns that appear in stage directions should not be included.)

[6 marks]

(c) Write an XQuery query to produce an XML document consisting of the text of the speech in `"speech.xml"` presented as a list of words. Your document should be valid with respect to the DTD below.

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

Note that, in this, DTD the `w` element does not have any associated attributes.

[6 marks]

(d) Describe the various processes that need to be undertaken in going from the plain text of a single speech in a play to an annotated entry in a corpus such as the BNC. Illustrate your answer using the speech presented in Figure 1 as an example. What further issues are likely to arise when annotating an entire play rather than a single speech?

[10 marks]

3. You are interested in comparing the annual rainfall across different countries in Europe, so you search on the terms: **European**, **country**, **annual** and **rainfall** using an information retrieval system. Among the documents found by the system are three with the following frequencies for each of the above terms.

	European	country	annual	rainfall
Document A	0	0	3	4
Document B	4	12	3	0
Document C	3	3	3	3
Query	1	1	1	1

(a) Write out the formula for calculating the *cosine* between two 4-value vectors (x_1, x_2, x_3, x_4) and (y_1, y_2, y_3, y_4).

[3 marks]

(b) Apply the above formula to calculate the cosines between the query vector in the table above, and the frequency vectors for each of the three documents (you should not need a calculator for this). Then rank the three documents in order of suitability according to the cosine similarity measure.

[7 marks]

(c) Comment briefly both on the choice of query and on the use of the cosine similarity test for ranking it. Are these chosen appropriately given the information the user wishes to find?

[3 marks]

(d) In Document B above, the term **European country** is investigated for its significance as a bigram. The data is tabulated below.

O_{ij}	European	¬European
country	4	8
¬country	0	36

Explain what each of the four entries in this table represents.

[4 marks]

(e) You are told that the χ^2 value for the bigram **European country** in Document C is 13.09 (rounded to two decimal places). Show in detail how this result is calculated. (You should not need a calculator for this.)

[10 marks]

(f) The critical value for a χ^2 test with one degree of freedom is 6.64 for $p = 0.01$ and 10.83 for $p = 0.01$. Use this information to interpret the result of the χ^2 test of part (d).

[3 marks]