UNIVERSITY OF LONDON

CM3010

BSc EXAMINATION

COMPUTER SCIENCE

Databases and Advanced Data Techniques

Release date: Wednesday 27 September 2023 at 12:00 midday British Summer Time

Submission date: Thursday 28 September 2023 by 12:00 midday British Summer Time

Time allowed: 4 hours to submit

INSTRUCTIONS TO CANDIDATES:

Part A of this assessment consists of a set of **TEN** Multiple Choice Questions (MCQs). You should attempt to answer **ALL** the questions in **Part A.** The maximum mark for Part A is **40**.

Candidates must answer **TWO** out of the **THREE** questions in **Part B**. The maximum mark for Part B is **60**.

Part A and Part B will be completed online together on the Inspera exam platform. You may choose to access either part first upon entering the test area but must complete both parts within **4 hours** of doing so.

Calculators are **not** permitted in this examination. Credit will only be given if all workings are shown.

Do not write your name anywhere in your answers.

PART A

Candidates should answer the **TEN** Multiple Choice Questions (MCQs) in Part A of the test area.

PART B

Candidates should answer any **TWO** questions from Part B.

Question 1: Linked Data Question

a. Consider the document below, retrieved from http://babelnet.org/rdf/post_n_EN:

```
@prefix bn:     <http://babelnet.org/rdf/> .

@prefix lemon: <http://www.lemon-model.net/lemon#> .

@prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#> .

bn:post_n_EN  a                lemon:LexicalEntry ;

       lemon:canonicalForm    <http://babelnet.org/rdf/post_n_EN/canonicalForm> ;

       lemon:language         "EN" ;

       lexinfo:partOfSpeech   lexinfo:noun .
```

Note: These servers sometimes respond to a request with an error message. You do not need to follow the links to answer this question.

i. What is the generic data model this information is represented in?

[1 mark]

ii. What is the serialisation format used for the data model?

[1 mark]

b. One friend says that it is, strictly speaking, impossible to know what word this RDF is talking about – let alone what language or part of speech it has – without requesting more triples. Another says that it's clear from these it is the English word post, being used as a noun. To what extent is either of them right? Explain your reasoning. What further information would help? [4 marks]

c. The document retrieved when you request the URL:

`http://babelnet.org/rdf/post_n_EN/canonicalForm` includes the following triple:

`<http://babelnet.org/rdf/post_n_EN/canonicalForm> lemon.writtenRep "post"`

i. Given a triplestore which follows the pattern of the data you have seen so far in this question, write a SPARQL query that finds the written representation and language for all nouns.

[6 marks]

ii. Write a SPARQL query that finds the language and part of speech for all words whose canonical form is written "post"

[4 marks]

d. The following document is (based on) an extract of the resource at:
http://www.lemon-model.net/lemon#

Note: These servers sometimes respond to a request with an error message. You do not need to follow the links to answer this question.

```
:LexicalSense
   a rdfs:Class, owl:Class ;
   rdfs:comment "Represents the intersection in meaning between the lexical
entry and the ontology entity. This is used as the ontology entity and lexical entry may not be in one-to-one
correspondence."@en ;
   owl:disjointWith :LexicalEntry :SenseDefinition.

:SenseDefinition
   a rdfs:Class, owl:Class ;
   rdfs:comment "A definition of a sememe, that is the a text describing the
exact meaning of the lexical entry when its sense is the given ontology reference"@en .

:definition
   a rdf:Property, owl:ObjectProperty ;
   rdfs:comment "Indicates a natural language definition. Note there is a pseudo-node to allow for further
description of the definition"@en ;
   rdfs:domain :LexicalSense ;
   rdfs:range :SenseDefinition .

 :value
   a rdf:Property, owl:DatatypeProperty ;
   rdfs:comment "This indicates the value of a pseudo-data node. An example of this is definition where the
value would generally be a string but it would not be possible to add further annotations, such as source or
creation date."@en .
```

i. What is the role of this document?

[1 mark]

ii. What format is it in?

[1 mark]

iii. To what does the ‘owl’ prefix refer?

[1 mark]

iv. This extract gives all relevant information for writing a textual definition for the objects we have been working with. Write triples (using a serialisation format of your choice) to provide one definition for the English noun “post” described above.

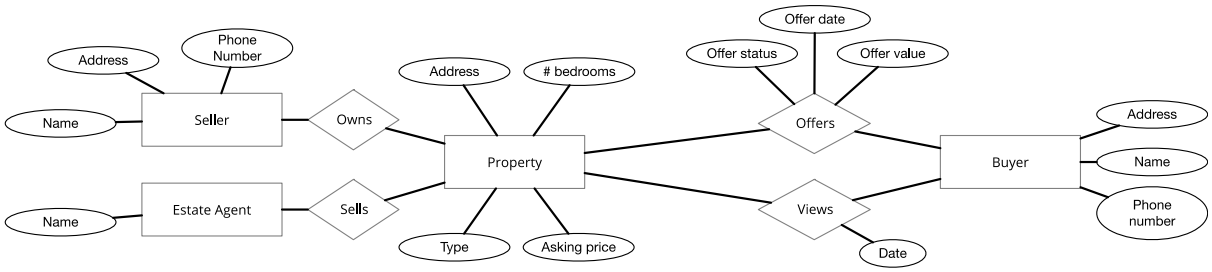
[4 marks]

e. Sketch an ER diagram for a relational implementation of this model. Include cardinality where known.

[7 marks]

Question 2: ER Question

An estate agency selling residential houses and flats is building a database to track its business activity. The ER diagram below shows the main elements. Sellers approach the agency with a property to sell. The property has a type (e.g. ‘terraced house’, or ‘flat’), a number of bedrooms and an asking price. A single agent is assigned to each property. The agent co-ordinates viewings of the property by potential buyers and, eventually, handles offers to buy them. The offer goes through a series of stages – ‘offer made’, ‘rejected’ or ‘offer withdrawn’, ‘accepted’ and ‘sale completed’ (not all stages will happen for any given offer).



a. Add cardinality indications for this diagram.

[3 marks]

b. How would you adapt this to a relational model? Be specific, naming any new entities, relations or attributes.

[5 marks]

c. List the tables, primary and foreign keys for a relational implementation of this database.

[6 marks]

d. Give the MySQL command for creating one of those tables.

[3 marks]

e. Agents are paid a commission on property where the offer gets to ‘sale completed’ status. The commission is 1% of the sale price.

   i. Write a MySQL query to calculate and list the commission earned since 1 January 2023 for each Estate Agent.

   [6 marks]

   ii. Modify your query from (i) above to list just the top earning agent.

   [2 marks]

f. The IT specialist at the agency is considering using a document database instead of a relational database. Give reasons **specific to this use case** for why this might be a good or bad idea (general observations about the difference between models do not receive marks).

[5 marks]

Question 3: IR/doc db question

The Hathi Trust Digital Library is a repository that gathers books from research libraries primarily digitised by Google Books or the Internet Archive. It contains a vast number of texts, but not all of them are perfectly catalogued. Researchers trained a Machine Learning system to recognise the language of works of fiction in the Hathi Trust collection.

The classifier identifies books as being in German with 80% precision and 88% recall.

a. If the system lists 2,200,000 books as being in German, how many of these are likely to be in German?

[2 marks]

b. Given your answer to (a), how many books in the whole collection are likely to be in German (including those that haven't been classified as German)? [3 marks]

c. Danish language fiction is identified with 100% precision and 76% recall. If texts from the labelled books are going to be used for machine learning systems, why might this performance be more useful than the accuracy of German classification?

[5 marks]

d. The F1 measure for these are (German) and (Danish). What is an F1-measure?

[2 marks]

e. The researcher has made a local document database to store a selection of transcribed books. They run the following command:

```
db.books.find({ lang: "German" })
```

What does this command do?

[1 mark]

f. Books in the database have a '**year**' field. Rewrite the command to include only volumes published in the nineteenth century [5 marks]

g. Book contents are included in a single textual field called "**text**". How would you adjust your query to include only books containing the word "Strudel"?

[2 marks]

h. In order to represent the books' content and structure more accurately, the researcher is trying to choose between enriching this database or switching to an XML database using TEI to encode book content and catalogue information. What factors should the researcher take into account in their decision?

[10 marks]

END OF PAPER