Número:_____ Nome: _____

INSTITUTO SUPERIOR TÉCNICO
Gestão e Tratamento de Informação

Exame 2 - Solution 2 Fevereiro 2011

- A duração deste exame é de **2 Horas**.
- É um exame com consulta, mas individual.
- Não é permitida a utilização de computadores nem telemóveis.
- O número total de valores é 20.
- Marque as suas respostas na folha de exame.
- Escreva o seu número e nome no topo de cada página.
- Apresente todos os cálculos realizados.
- Após o inicio da prova, poderá abandonar a sala ao fim de **1 hora** mediante a entrega do exame.

Para o uso oficial somente

1	**2**	**3**	**4**	**5**	**SUM**
4	4	4	4	4	20

Número:_____ Nome: _____

1. (4 vals) eXtensible Markup Language (XML), XPath and XQuery

Consider the following example XML document, encoding information about people and their activities.

```
<list-people-activities>
     <person id="1">
          <name>Maria Amélia</name>
          <city country="Portugal">Lisboa</city>
     </person>
     <person id="2">
          <name>João Mandrião</name>
          <city country="França">Paris</city>
     </person>
     <person id="3">
          <name>José Meireles</name>
          <city country="Portugal">Porto</city>
     </person>
     <!--     list of remaining persons   -->
     <activity>
          <person>1</person>
          <person>5</person>
          <person>3</person>
          <activity-name>Professor</activity-name>
     </activity>
     <activity>
          <person>2</person>
          <person>6</person>
          <activity-name>Plumber</activity-name>
     </activity>
     <!--  list of remaining activities   -->
</list-people-activities>
```

1.1. (2,5 pts) Present an XQuery expression for answering the following information need: *For each distinct country, list the number of different activities performed by inhabitants of that country.*

Answer:

```
for $pa in distinct-values($doc//@country)
let $pe := $doc//person[./city/@country=$pa]/data(@id)
let $ac := distinct-
values($doc//activity[data(person)=$pe]/activity-name)
return <country name="{$pa}" activities="{count($ac)}" />
```

Número:_____ Nome: _____

1.2. (1 pt) Present an XPath expression for answering the following information need: *List the names for people that perform an activity called "Professor" in any city from the country named "Portugal".*

Answer:

```
$doc//person[@id=$doc//activity[activity-
name="Professor"]/person[city/@country="Portugal"]]/name
```

1.3. (0,5 pt) Present an XQuery Update expression for modifying the "*activity*" elements, adding to each activity an attribute containing the number of people that perform it, and replacing the name of the element called "*activity-name*" by "*profession*".

Answer:

```
copy $doc := $doc1
modify (
 for $a in $doc//activity return
 ( insert node attribute count {count($a/person)} as last into
$a,
   rename node $a/activity-name as "profession" )
) return $doc
```

Número:_____ Nome: _____

2. (4 pts) Information Extraction

Consider the Hidden Markov Model represented by the following probabilities:

$$\Pi = [0.5 \quad 0.5] \qquad B = \begin{bmatrix} 0.2 & 0.1 \\ 0.5 & 0.1 \\ 0.3 & 0.8 \end{bmatrix} \qquad A = \begin{bmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{bmatrix}$$

where the symbols corresponding to each line in matrix *B* are *a*, *b*, and *c*.

2.1. (2,5 pts) Compute the probability of occurrence of the sequence *abcb*.

Answer:

Using the forward procedure, we have:

	A	B	C	B
1	0,10000000	0,02500000	0,00345000	0,00427750
2	0,05000000	0,01000000	0,01880000	0,00136950

P(O) = 0,00564700

Número:_____ Nome: _____

2.2. (1 pt) What would be the probability of the sequence *abcb* occuring if the sequence of states were 1122.

Answer:

P(abcb,1122) = Π(1)*B(a,1)*A(1,1)*B(b,1)*A(1,2)*B(c,2)*A(2,2)*B(b,2) = 0,00050400.

2.3. (0,5 pt) What is the most likely sequence of states for the sequence of symbols *abcb?*

Answer:

Using the Viterbi algorithm:

	A	B	C	B
	1	2	3	4
1	2,3025851	4,1997051	6,6076507	6,3889615
2	2,9957323	4,9618451	4,7795236	7,5929343

where the minimal path is shown in blue. The sequence of states is, therefore, 1121, with probability P(abcb,1121) = 0,0016800.

Número:_____ Nome: _____

3. (4 pts) Data Integration

3.1. (2,5 pts) Suppose that you have a relation Friends (x, y) where x and y refer to IDs of two friends. You also have a relation GTI(x, g) where g is the GTI grade of user x. Write Datalog rules for the following queries:

a) Find the GTI grade of the user whose ID is “1024”.
b) Find the IDs of users who have at least a friend with a GTI grade greater than 15.

Answer:

a) GTIGrade(g) : - GTI(1024, g)
b) FriendsGTI(x) :- Friends(x, y), GTI(y, g), g> 15

3.2. (1 pt) Assume the following two views and a query:

v1 (X, Y) : − s1 (X, Z), s2 (Z, Y)
v2 (X, Y) : − s3 (Y), s1 (X, Y)

q(X, Y) : − v1 (X, Z), v2 (Z, Y)

Unfold the query q so it only uses relations s1, s2 and s3 .

Answer:

Q(X, Y) :- s1(X, T), s2(T, Z), s3(Y), s1(Z, Y)

Número:_____ Nome: _____

3.3. (0,5 pt) Suppose the following source relations:

```
Professor(Id, Name, Sal)
Address(Id, Addr)
     Id: FK(Professor)
Student(Name, GPA, Year)
PayRate (Rank, HrRate)
Works (Name, Proj, Hrs, ProjRank)
     Name: FK(Student)
     ProjRank: FK(PayRate)
```

the target relation:

```
Personnel(Id, Name, Address, Sal)
```

the correspondences resulting from the schema matching phase:

```
C1: PayRate.HrRate, Works.Hrs -> Personnel.Sal
C2: Professor.Sal -> Personnel.Sal
C3: Professor.Id -> Personnel.Id
C4: Professor.Name -> Personnel.Name
C5: Address.Addr -> Personnel.Address
C6: Student.Name -> Personnel.Name
```

and the following SQL query that is generated as one possible solution for schema mapping by the fourth phase of CLIO query discovery algorithm:

```
SELECT P.Id, P.Name, P.Sal, A.Addr
FROM Professor P, Address A
WHERE A.Id = P.Id
UNION ALL
SELECT NULL as ID, S.Name, P.HrRate*W.Hrs, Null as Addr
FROM Student S, PayRate P, WorksOn W
WHERE S.name=W.name AND W.ProjRank = P.Rank
```

Show the corresponding set of covers returned by the third phase of the algorithm.

Answer:

{{C2, C3, C4, C5},{C6, C1}}

covers all correspondences and is minimal

Número:_____ Nome: _____

4. (4 pts) Data Cleaning

4.1. (2,5 pts) Suppose the following two strings:

```
Euler
Ellery
```

α) Compute the SOUNDEX code for each of them and conclude about their similarity.
β) Compute the edit distance between the same two strings and the corresponding similarity value.
χ) Conclude about the two similarity measures computed and their suitability to represent the real similarity between the two strings

Answer:

a) Both have the same SOUNDEX code so they are considered similar
b) TO DO
c) The two strings correspond to two different names, although they are phonetically similar.

Número:_____ Nome: _____

4.2. (1 pt) Consider the following table storing information about people:

Name	**Zip code**	**Local**
Maria Amélia Santos	3750-011	Agadão, Águeda
Francisco Martins	9400-025	Rua José António Almeida, nº 32, Funchal
Pedro Carvalho	1250-144	Madeira
Diogo Antunes	-	Águeda, Águeda
Pedro M. Carvalho	1250	Porto Santo
Joana Almeida	3750-031	Aguada de Baixo, Águeda
Maria Amelia	3751-011	Águeda
Joana Armeida Rodrigues	3750	Agueda

Give four distinct examples of dirty data and justify.

Answer:

1) missing data: tuple “Diogo Antunes”, because zip code is not filled in.
2) Approximate duplicate records : 1st and 7th tuples, because they have similar names, same zip code and similar locals
3) 7th tuple: misspelling in the name
4) 2nd tuple: the field local hás more information besides the local: free-field.

Número:_____ Nome: _____

4.3. (0,5 val) Assume that approximate duplicate records in the table of question 4.2 are assigned the same value of an additional field ID. Write in SQL, statements that fuse/merge the records of the table that refer to the same real-world entity.

Suggestions: Use grouping and aggregation. Assume that you can use user-defined aggregation functions.

Answer:

```
SELECT ID, Longest(Name),  Longest(Zip), Longest(Local)
FROM T
GROUP BY ID
```

Número:_____ Nome: _____

5. (4 vals) Miscelaneous

5.1. (1 pt) Consider the answers to questions 2.1, 2.2, and 2.3.

a) We know that the probability value for question 2.2 will always be lower than the probability value for question 2.3. Explain why?

Answer:

In question 2.3 we have found the sequence of states that maximizes the probability of generating the given sequence. Thus, it is the highest possible value.

b) Is the value obtained in question 2.1 higher or lower than that obtained in question 2.2.? Could it be othwerwise? Explain why (or why not).

Answer:

It is higher. The value in question 2.1 is the sum of the probabilities of every state generating the given sequence, thus it will always be higher (or equal to) the probability of any single specific state.

5.2. (1 pt) Consider the following telephone numbers (each of them indicating a unique sequence of characters):

```
8144658695
812 673 5748
812  453   6783
812-348 7584
(617) 536 6584
834-674-8595
```

Write an XPath expression that, through the usage of the replace function, allows one to reformat the numbers so that the results are as illustrated next:

```
814 4658695
812 673 5748
812 453 6783
812 348 7584
617 536 6584
834 6748595
```

Answer:

```
replace($telephone, "^\(?([0-9][0-9][0-9])\)?[- ]*([0-9][0-9]
[0-9])( ?)[- ]*", "$1 $2$$3)
```

Número:_____ Nome: _____

5.3.(1 pt) Suppose two wrappers over the mondial data set and over the geonames service, exporting country and city:

```
w1(city, country)
w2(city, country)
```

How would you express in Datalog a mediator `m(city, country)` using the GAV mapping language over these two wrappers?

Answer:

```
M(city, country) ⊃= w1(city, country)
M(city, country) ⊃= w2(city, country)
```

5.4. (1 pt) Explain why the *Sorted Neighborhood Join* method used for speeding up the data matching (or record linkage) process may loose duplicates. Does the method include any technique to overcome that inconvenient and decrease the number of missing duplicates? If yes, which one?

Answer:

SNM may loose duplicates if the key chosen to sort the records doesn't put duplicate records near each other. If this happens, these duplicate records will not belong to the same window and thus will never be compared.

One way to circumvent this is to run the algorithm several times, each one with a different key, and then combine the resulst using transitive closure.