Number:_____ Name: _____

---

# INSTITUTO SUPERIOR TÉCNICO
# GESTÃO E TRATAMENTO DE INFORMAÇÃO

Exam 1 16 January 2011

---

- The duration of this exam is **2 Hours**.
- You can access your own paper material, but the exam is to be done individually.
- You are not allowed to use computers nor mobile phones.
- The maximum grade of the exam is 20 pts.
- Write your answers below the questions.
- Write your number and name at the top of each page.
- Present all calculations performed.
- After the exam starts, you can leave the room **one** hour after and after delivering the exam.

To be used by instructors, ONLY:

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

Number:_____ Name: _____

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

Consider the following example XML document, encoding information about plastic artists and their works of art.

```
<list-artists-works>
 <artist id="1">
  <name>Vincent van Gogh</name>
  <genre>Expressionism</genre>
  <genre>Post-impressionism</genre>
 </artist>

 <artist id="2">
  <name>Claude Monet</name>
  <genre>Impressionism</genre>
 </artist>

 <artist id="3">
  <name>Pablo Picasso</name>
  <genre>Modernism</genre>
  <genre>Cubism</genre>
  <genre>Expressionism</genre>
 </artist>

 <!-- list of remaining artists -->

 <work artist="1">Os Comedores de Batata</work>
 <work artist="1">Os Girassóis</work>
 <work artist="1">A Noite Estrelada</work>
 <work artist="2">Impressão, nascer do sol</work>
 <work artist="3">Dora Maar au chat</work>

 <!-- list of remaining works -->

</list-artists-works>
```

**1.1. (2,5 pts)** Present an XQuery expression for returning the following: List the names of artists from a genre related with "impressionism" (i.e., with a genre containing "impressionism" as part of its name), sorted, in ascending order, by the number of their works.

Number:_____ Name: _____

**1.2. (1 pt)** Present an XPath expression for answering the following information need: List the names of all the works of art authored by artists that have worked in more than two different genres.

**1.3. (0,5 pt)** Present an XQuery Update expression for modifying the XML document in order to:

- add, to each artist element, a new element entitled "*number-works*" containing the number of works authored by that artist, and
- change each of the "*work*" elements in order to replace the value of the "*artist*" attribute by the name of the corresponding artist.

The same XQuery Update expression should perform both modifications.

Number:_____ Name: _____

**2. (4 pts) Web data and information extraction**

Consider the following character strings:

VODKALIMA MESKALINA

**2.1. (2,5 pts)** Using the dynamic programming algorithm, compute the edit distance between both strings.

**2.2. (1 pt)** Present a minimal alignment between the strings. You must present the corresponding backtracking path on the distance matrix computed for the previous question.

Number:_____ Name: _____

**2.3. (0,5 pt)** Assume the cost of replacing a character is twice the cost of inserting/removing a character. What would be the edit distance in this case? Justify your answer.

Number:_____ Name: _____

## 3. (4 pts) Data integration

**3.1. (2,5 pts)** Consider a mediator schema with the following two relations:

```
Movie(title, year, type)
Schedule(cinema, title, hour)
```

a) Write in Datalog and using the mediator schema the query: which cinemas show movies of type comedy?

b) Now suppose the following views were defined over the mediator:

```
Movies70(TT,Y) :- Movie(TT,Y,-), Y >= 1970
MoviesMonumental(TT) :- Schedule(C,TT,-), C = "Monumental"
```

How would you express the following query using the views: In which year were produced the movies that are shown in Monumental?

**3.2. (1 pt)** Unfold the query of question 3.1 b) so that the query is expressed only in terms of the base relations.

Number:_____ Name: _____

**3.3. (0,5 pt)** Explain what is the purpose of the schema matching phase of a schema mapping process. Give an example of two techniques that can be used for schema matching.

Number:_____ Name: _____

## 4. (4 pts) Data cleaning

**4.1. (2,5 pts)** Compute the value of the Jaro-Winkler measure for the two names: `Jonhatan` and `Johnny`. Use 0.1 as the weight to be given to the prefix.

**4.2. (1 pt)** Compute the value of the Jaccard measure for the same two names, assuming 2-grams. What can you conclude about the use of these two measures for computing the similarity between these two names?

Number:_____ Name: _____

**4.3.(0,5 val)** Explain why methods for optimizing the similarity join between two sets of strings, when the number of strings in each set is very large, were proposed. Give an example of one such method.

Similarity join is an operation that accepts as input two sets of objects (e.g., strings), a similarity measure sim(), and a threshold value t. It returns all pairs of objects whose similarity value is greater than or equal to t.

Number:_____ Name: _____

## 5. (4 pts) Miscelaneous

**5.1. (1 pt)** Explain how the string edit distance measure can be used to extract data from Web pages. Illustrate your explanation with an example.

**5.2. (1 pt)** Write an XQuery function for computing the Jaccard coefficient between two sets of words, considering the following signature:

```
jaccard-coefficient ( $s1 as xs:string* ,
                      $s2 as xs:string* ) as xs:double
```

Explain how you could change the function in order to compute a "*soft*" similarity score between sets of words (i.e., a "*soft-jaccard*" score that would take into account similar words in the sets), using an internal comparator that you consider to be appropriate to this task.

Number:_____ Name: _____

**5.3. (1 pt)** Consider the following sequence of example sentences and the corresponding classifications:

```
"the movie is great"              -> positive opinion
"it's a crap movie"               -> negative opinion
"really wonderful movie"          -> positive opinion
"really awful movie"              -> negative opinion
"I liked the movie, it is great" -> positive opinion
```

Use the Naive Bayes algorithm to classify the sentence "*really great movie*" as expressing a positive or a negative opinion, with basis on the example sentences that were presented.

Number:_____ Name: _____

**5.4. (1 pt)** Explain briefly the main steps of a data cleaning process that are usually required.