Number:_________________ Name: ____

INSTITUTO SUPERIOR TÉCNICO
GESTÃO E TRATAMENTO DE INFORMAÇÃO

Exam 1 - Solution 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 vals) 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.

Answer:

```
for $a in $doc//artist[some $g in ./genre satisfies
matches($g,"[Ii]mpressionism")] let $works := $doc//work[@artist=$a/@id]
order by count($works) ascending return $a/name
```

Number:________________ Name: _______________________________________

1.2. (1 val) 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.

Answer:

```
$doc//work[@artist=$doc/artist[count(work) > 2]/@id]/text()
```

1.3. (0,5 val) 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 each 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.

Answer:

```
copy $doc := $doc1 modify (   ( for $a in $doc//artist return    insert node
<number-works>{ count($doc/work[@artist=$a/@id]) }</number-works> as last
into $a ,    for $a in $doc//work return    replace value of node
$a/@artist with $doc/artist[@id=$a/@artist]/name  ) ) return $doc
```

Number:__________________ Name: ____

2. (4 pts) 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.

Answer:

		V	O	D	K	A	L	I	M	A
	0	1	2	3	4	5	6	7	8	9
M	1	**1**	2	3	4	5	6	7	7	8
E	2	2	**2**	3	4	5	6	7	8	8
S	3	3	3	**3**	4	5	6	7	8	9
K	4	4	4	4	**3**	4	5	6	7	8
A	5	5	5	5	4	**3**	4	5	6	7
L	6	6	6	6	5	4	**3**	4	5	6
I	7	7	7	7	6	5	4	**3**	4	5
N	8	8	8	8	7	6	5	4	**4**	5
A	9	9	9	9	8	7	6	5	5	**4**

The edit distance is 4.

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.

Answer:

The path is shown above.

Alignment:

V	O	D	K	A	L	I	M	A
M	E	S	K	A	L	I	N	A

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.

Answer:

By considering the cost of replacement 2 and the cost of insertion/removal 1, we would have:

		V	O	D	K	A	L	I	M	A
	0	1	2	3	4	5	6	7	8	9
M	1	2	3	4	5	6	7	8	7	8
E	2	3	4	5	6	7	8	9	8	9
S	3	4	5	6	7	8	9	10	9	10
K	4	5	6	7	6	7	8	9	10	11
A	5	6	7	8	7	6	7	8	9	10
L	6	7	8	9	8	7	6	7	8	9
I	7	8	9	10	9	8	7	6	7	8
N	8	9	10	11	10	9	8	7	8	9
A	9	10	11	12	11	10	9	8	9	8

The edit distance is 8.

Number:_________________ Name: ____

3. (4 pts) Data integration

3.1. (2,5 pts) Consider a mediator schema with the follwoing 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?

Answer:

a) Q(C) :- Movie(T, -, TP), TP = "comedy", Schedule(C, T, -)

b) Q1(Y, 'Monumental') :- Movies70(TT, Y), MoviesMonumental(TT)

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.

Answer:

Q1(Y,'Monumental') :- Movie(TT,Y,-), Y >= 1970, Schedule(C,TT,-), C = "Monumental"

3.3. (0,5 val) 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.

Answer:

The purpose of schema matching is to find the correspondences between elements of a source schema and elements of a target schema.
Elements can be relation or attribute names.
One technique that can be used is to apply string matching functions to the element names in order to find similar element names.
Another one that can be used is to take into the account the values of the instances and infer whether the corresponding elements store information about the same subject.

Number:________________ Name: ____

4. (4 pts) Data cleaning

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

Answer:

JaroWinkler (x, y) = (1 – PL*PW) * jaro(x,y) + PL*PW
Jaro(x,y) = 1/3 [c/|x| + c/|y| + (c –t/2)/c]

X = Jonhatan
|x| = 8
y = Johnny
|y| = 6
Min(|x|, |y|) = 6; 6/2 = 3
C = (J, o, h, n, n) = 5
T = 2 (n, h)

Jaro(x, y) = 1/3(5/8 + 5/6 + (5-2/2)/5) = 1/3(0.625 + 0.833 + 4/5)
= 0.75

The longest common prefix is 'Jo" which has size 2, therefore:

JaroWinkler(x,y) = (1-0.2) * 0.75 + 2*0.1 = 0.8*0.75 + 0.2 = 0.8

4.2. (1 val) 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?

Answer

"Jonhatan" = {#j, jo, on, nh, ha, at, ta, an, n#}
"Johnny" = {#j, jo, oh, hn, nn, ny, y#}

Bx intersect By = {#j, jo}

J(x,y) = |Bx intersect By |/|Bx union By| = 2/ 14 = 1/7 = 0.142

The JaroWinkler method gives a weight to the common prefix between the two strings, so it assigns a higher similarity to these two strings which share a prefix.

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.

Answer:

The naïve method to compute the similarity between two sets of strings corresponds to the execution of a Cartesian Product between the two sets. This means to compare every string from one of the sets with every string of the oher set, which leads to an execution time unacceptable for a large number of strings.

One way of optimizing this process is to filter the strings by size and only compare those whose difference of sizes is less than the threshold used for the similariy measure.

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.

Answer:

To extract Web data using the string edit distance (SED) we could see a Web page as a sequence of HMTL tags, each tag represented by a symbol. We could then use the SED to compare the pages and determine which parts of the page structured are aligned. Those that are not aligned contain the values we want to extract.

For example, the pages

1: <html><strong>The Jungle Book</strong></html>

2: <html><strong>The Man Who Would Be King</strong></html>

Could be seen as sequences 1: ABCDE and 2: ABGDE. Using the SED we would get the alignment A-A, B-B, C replaced by G, D-D, and E-E. This tells us that C and G are likely to be information to be extracted, whereas A, B, D, and E are structure symbols.

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.

Answer:

```
jaccard-coefficient ($s1 as xs:string* , $s2 as xs:string*) as xs:double {
  count( $s1 intersect $s2 ) div count( $s1 union $s2 )
};
```

To implement a "soft-jaccard" function, the intersect operator should be replaced by a XQuery function that, when comparing the elements from the two entry sets, took into account their similarity through an internal function such as the edit distance. This function should return the set of elements from one of the sets whose similarity towards other element(s) of the other set is larger than a given threshold, instead of just returning the elements in common between both sets.

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.

Answer:

```
One could use the occurrences of individual words, together with the Bayes
theorem, to estimate the probability for each of the two classes:

P(positive | "really great movie") =
        P("really great movie" | positive) * P(positive)
      = P("really" | positive) * P("great" | positive) *
        P("movie" | positive)  * P(positive)
      = 1/15 * 2/15 * 3/15 * 3/5

P(negative | "really great movie") =
        P("really great movie" | negative) * P(negative)
      = P("really" | negative) * P("great" | negative) *
        P("movie" | negative)  * P(negative)
      = 1/8 * 0/8 * 2/8 * 2/5

Alternatively, the following solution would also be considered, using the
number of sentences where a word occurs:

P(positive | "really great movie") =
        P("really great movie" | positive) * P(positive)
      = P("really" | positive) * P("great" | positive) *
        P("movie" | positive)  * P(positive)
      = 1/3 * 2/3 * 3/3 * 3/5

P(negative | "really great movie") =
        P("really great movie" | negative) * P(negative)
      = P("really" | negative) * P("great" | negative) *
        P("movie" | negative)  * P(negative)
      = 1/2 * 0/2 * 2/2 * 2/5

Notice the zero probability in the case of “negative opinion”.

Since

P(positive | "really great movie") > P(negative | "really great movie")

the class for the sentence would be "positive opinion".
```

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

Answer:

The main steps of a data cleaning process are the following ones:

1. Extraction of the individual fields that are relevant
2. Standardization of record fields
3. Correction of data quality problems at attribute level (Missing values, syntax violation, etc)
4. Correction of data quality problems at attribute-set level and record level (Synonyms, homonyms, uniqueness violation, integrity constraint violation, etc)
5. Correction of data quality problems at relation level (Violation of functional dependencies, duplicate elimination, etc)
6. Correction of data quality problems problems at multiple relations level (Referential integrity violation, duplicate elimination, etc)

The integration of the user feedback is usually required to solve instances of data quality problems not addressed by the automatic methods used in any of the steps.

The effectiveness of the data cleaning and transformation process must be always measured for a sample of the data set to assess the quality of the data obtained.