Ano lectivo 2009/2010 – 1º semestre

Gestão e Tratamento de Informação

Exame 2

Regras

- O exame tem a duração de **2 horas**.
- O exame é **com consulta**, mas **individual**.
- Não é permitida a utilização de **qualquer material electrónico**, excepto calculadora. **Não é permitida a partilha de calculadoras**.
- Todas as folhas entregues devem ser identificadas com o **nome e número do aluno**.
- Após o início da prova, só poderá abandonar a sala ao fim de **30m**, **mediante a entrega do exame**.
- Deve **apresentar sempre os cálculos** que fez para todas as questões.

Cotação das questões

Questão	**1–4**			**5**
Alínea	(a)	(b)	(c)	(a)–(d)
Valor	2,5	1	0,5	1

1

Considere as seguintes sequências de caracteres:

(i) `CASPA` (ii) `COCAS`

Usando o algoritmo de programação dinâmica:

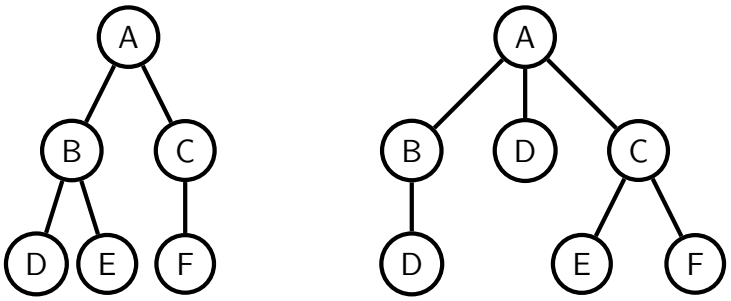
a) Calcule a distância de edição entre as sequências.

b) Indique um possível alinhamento mínimo.

c) Quantos alinhamentos mínimos existem para estas sequências? Explique como chegou a esse valor.

2

Considere as árvores apresentadas na figura seguinte.



a) Utilizando o algoritmo *Simple Tree Matching*, qual é a similaridade entre as árvores dadas?

b) O algoritmo *Simple Tree Matching* constrói uma matriz que conterá os valores de similaridade parciais das árvores a ser comparadas. O algoritmo termina retornando o valor na última posição da matriz, **somando-lhe 1**. Porque razão é feita esta soma?

c) No contexto de extracção de informação da Web, a que poderiam corresponder as árvores dadas e como seria usado o algoritmo? Dê um exemplo.

3

a) Responda às seguintes questões sobre integração de dados.

1. No que consiste o problema de *query containment*?
2. De que forma poderia o optimizador de um sistema wrapper-mediator tirar proveito da identificação rápida das relações de *query containment*?

b) Num sistema de integração de dados é frequente utilizarem-se interrogações SQL para descrever o mapeamento de esquemas/dados.

1. Nem todos os mapeamentos (i.e. transformações) são facilmente exprimíveis na linguagem SQL. Porquê?
2. Identifique um mapeamento (i.e. uma transformação) de dados que não seja facilmente exprimível em SQL.

   Sugestão: Identifique uma interrogação que não seja exprimível através da Álgebra Relacional ou através de Datalog sem recursão.

c) A operação de *approximate join* é uma operação fundamental com vista à detecção de duplicados. Esta operação é bastante exigente do ponto de vista computacional.

1. Explique no que consiste esta operação.
2. Explique de que forma se pode melhorar o desempenho desta operação.

4

Considere os dois documentos XML apresentados abaixo, de nome *ex1.xml* e *ex2.xml*. Estes correspondem a uma colecção de registos bibliográficos e a um conjunto de tópicos de pesquisa sobre essa colecção, respectivamente.

```
<livros>
   <livro id="1" titulo="Semantic Web for Dummies">
     <autores><autor>Jeffrey Pollock</autor></autores>
     <editora>John Wiley and Sons</editora>
     <resumo>This book covers the architectures, strategies, and
             standards involved in Semantic Web technology.</resumo>
   </livro>
   <livro id="2" titulo="Querying XML">
     <autores>
      <autor>Jim Melton</autor>
      <autor>Stephen Buxton</autor>
     </autores>
     <editora>Elsevier Science and Technology</editora>
     <resumo>This book provides a background from fundamental
             concepts of XML Query Languages.</resumo>
   </livro>
   <livro id="3" titulo="Data Quality and Record Linkage Techniques">
     <autores><autor>Thomas N. Herzog</autor></autores>
     <editora>Springer-Verlag</editora>
     <resumo>Helps practitioners understand at an applied level, of the
             issues involved in improving data quality through
             editing, imputation, and record linkage.</resumo>
   </livro>
   <!-- O documento XML completo contém mais livros da colecção -->
</livros>
```

Listing 1: Documento XML com colecção de registos bibliográficos.

```
<topicos>
   <topico tipo="easy" id="1" query="semantic web">
     <descricao>Encontrar livros sobre o tópico da Semantic Web<descricao>
   </topico>
   <topico tipo="hard" id="2" query="xml">
     <descricao>Encontrar livros sobre a linguagem XML<descricao>
   </topico>
   <!-- O documento XML completo contém mais tópicos de pesquisa -->
</topicos>
```

Listing 2: Tópicos de pesquisa sobre colecção de registos bibliográficos.

a) Apresente expressões XPath que, com base no documento XML com a informação bibliográfica (i.e., o documento *ex1.xml* da Listagem 1), permitam responder às seguintes necessidades de informação:

   1. Quais os editoras dos livros que contêm as palavras "XML"ou "MySQL"no título, e em que um dos autores tem o nome "Herzog"?

2. Quais os títulos dos livros cujo resumo refere o preço (i.e., um valor numérico seguido da palavra "Euros") e que têm vários autores?

b) Apresente uma expressão XQuery Full-Text que liste, por ordem decrescente dos respectivos *ids*, os dois livros mais relevantes para cada um dos tópicos de pesquisa classificados como "easy"(i.e., cada um dos tópicos descritos no documento *ex2.xml* da Listagem 2, onde o atributo *tipo* toma o valor "easy"). Apresente a resposta no formato exemplificado abaixo:

```
1  <resultados>
2   <topico query="xquery"~>
3     <livro editora="John Wiley and Sons" id="9">
4     <livro editora="Morgan Kaufman" id="5">
5   </topico>
6   <topico query="data integration">
7     <livro editora="Springer-Verlag" id="8">
8     <livro editora="Springer-Verlag" id="4">
9   </topico>
10 </resultados>
```

Listing 3: Documento XML exemplificando o formato de saída.

c) Apresente um XML Schema que permita validar o documento XML com informação sobre os tópicos de pesquisa (i.e., o documento *ex2.xml* da Listagem 2). Tenha em atenção que o atributo *id* do elemento *topico*, caso esteja definido, toma um valor único no documento. Além disso, o atributo *tipo* é obrigatório e toma apenas os valores "hard"ou "easy".

5

a) Apresente um algoritmo para cálculo da distância de edição entre sequências de caracteres que funcione apenas com sequências do mesmo tamanho, mas que seja mais eficiente do que o algoritmo de programação dinâmica aprendido nas aulas.

b) Desenhe um *Hidden Markov Model* para o problema de descobrir *links* para download de software em páginas HTML. Responda desenhando uma representação gráfica do modelo e indicando:

- O que representam os estados;
- O que representam os símbolos;
- Valores que ache apropriados para as probabilidades.

c) Escreva uma função em XQuery que aceite como entrada uma árvore XML e devolva, do conjunto de elementos filhos cujo nome apenas contém letras minúsculas, aquele(s) que tem o maior conteúdo textual.

d) Indique duas das dimensões da qualidade de dados e descreva sucintamente no que consistem.