Ano lectivo 2009/2010 – 1º semestre

Gestão e Tratamento de Informação

Exame 1

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.
- 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 **1 hora**, **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) `VINHOS` (ii) `LINHA`

Usando o algoritmo de programação dinâmica:

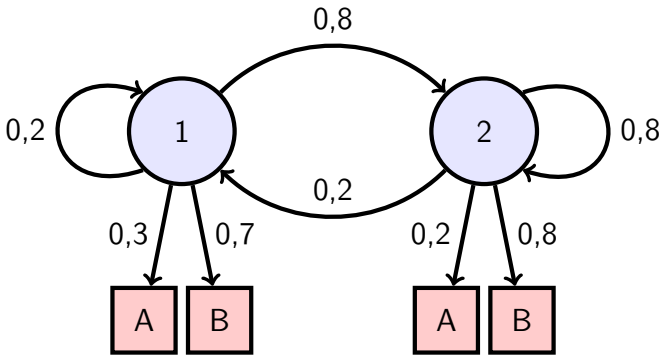
a) Calcule a distância de edição entre as sequências.

b) Indique um possível alinhamento mínimo.

c) Suponha que não é possível substituir caracteres, ou seja, que a distância deve ser calculada apenas com inserções e remoções. Qual a distância de edição entre as sequências, neste caso?

2

Considere o seguinte *Hidden Markov Model* (HMM):



em que a probabilidade inicial do estado 1 é 0,3 e a probabilidade inicial do estado 2 é 0,7.

a) Calcule a probabilidade de ocorrência da sequência AABAB.

b) A utilização de HMMs para extracção de informação, tal como aprendeu na disciplina, corresponde a um método *supervisionado* ou *não supervisionado*? Justifique a sua resposta.

c) Sabendo que a sequência de estados foi 11121, qual é a probabilidade de observar a sequência da alínea anterior?

3

a) Responda às seguintes questões sobre integração de dados.

   1. Num sistema de integração virtual *wrapper-mediator*, sempre que o mapeamento entre o esquema global e as fontes seja efectuado de acordo com o modelo *global-as-view*, que estratégia de optimização deve ser utilizada?
   2. Indique sucintamente no que consiste a estratégia que identificou anteriormente.

b) Nos processos de integração de dados, uma das principais dificuldades da etapa de mapeamento entre esquemas são as *heterogeneidades semânticas*.

   1. O que entende por heteronegeneidades semânticas?
   2. Exemplifique sucintamente uma heterogeneidade semântica entre duas tabelas.

c) Num sistema de integração de dados, após a detecção de duplicados, efectua-se tipicamente uma operação de consolidação dos tuplos identificados como duplicados.

1. Como se designa esta operação?
2. Nos casos em que esta operação é implementada em SQL num SGBD relacional, é frequente a necessidade de utilizar funções definidas pelo utilizador (UDF's). Justifique a razão desta necessidade.

4

Considere o documento XML que se apresenta abaixo, de nome *ex.xml*, o qual lista informação sobre trabalhadores e as seus respectivos locais de trabalho.

```
<trabalhadores>
   <pessoa id="12345" idade="30"~>
     <nome>ricardo esforçado</nome><cidade>lisboa</cidade>
   </pessoa>
   <pessoa id="54321" idade="25"~>
     <nome>joão mandrião</nome><cidade>porto</cidade>
     <funcao>pedeiro</funcao>
     <funcao>ladrilhador</funcao>
   </pessoa>
   <exerce>
     <id-pessoa>12345</id-pessoa>
     <local>restaurante faz-te à febra</local>
   </exerce>
</trabalhadores>
```

Listing 1: Documento XML com informação sobre trabalhadores

a) Apresente expressões XPath que, com base no documento XML com informação sobre os trabalhadores, permitam responder às seguintes necessidades de informação:

   1. Quais os nomes dos trabalhadores residentes em Lisboa, cuja idade representa um número par.
   2. Quais os locais de trabalho para os quais existe um funcionário cujo último nome é "mandrião".

b) Apresente uma expressão XQuery FLWOR que liste, por ordem crescente da idade, todas as pessoas com mais de 27 anos que trabalhem num local com a palavra "restaurante" no nome. A resposta deve ser apresentada no formato exemplificado abaixo:

```
1  <ul>
2   <li>
3    <nome>simão calão</nome>
4    <local>restaurante bota abaixo</local>
5   </li>
6   <li>
7    <nome>jamal pontual</nome>
8    <local>passas-fome restaurante</local>
9   </li>
10 </ul>
```

Listing 2: Documento XML exemplificando formato de saída

c) Apresente o fragmento de um XML Schema que permita validar a parte do documento XML com informação sobre os trabalhadores, tomando em atenção que atributo *id* do elemento pessoa permite identificar univocamente um indivíduo (i.e., é uma chave), que o atributo idade é um número inteiro entre 18 e 65, e que o elemento *funcao* se pode repetir um número indefinido de vezes.

5

a) Considere duas sequências de caracteres arbitrárias s_1 e s_2, de tamanho l_1 e l_2, respectivamente. Suponha que $l_1 \neq l_2$. Qual a menor distância de edição possível entre as duas sequências? Justifique.

b) Considere o HMM da Questão 2. No contexto de Extracção de Informação, que tipo de problemas poderia resolver este modelo? Para responder, explique a lógica da sua proposta e diga a que corresponderiam os estados e os símbolos observáveis.

Nota: Indique apenas um caso plausível, sem se preocupar se o uso deste HMM seria efectivamente a melhor solução ou se estas seriam as probabilidades adequadas.

c) Escreva uma função em XQuery que, aceitando como entrada uma cadeia numérica com um máximo de seis algarismos, verifique se a mesma é uma capicua através de uma expressão regular. Em caso afirmativo, a função deve retornar o factorial do número fornecido à entrada. No caso contrário, a função deve retornar zero.

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