Número:_____ Nome: _____Versão A

Ciências ULisboa

Processamento de Dados

Exame Modelo 2019/20
Duração 60 minutos

Grupo I – Assinale a resposta correta com um X na tabela seguinte.
Respostas assinaladas justo das questões serão ignoradas. Cada reposta errada terá uma penalização de -0.5.

	1	2	3	4	5
a					X
b		X	X		
c				X	
d	X				

1. [2.0] Formatos de ficheiros de texto para armazenar dados:
   (a) TSV, CSV, e XLS
   (b) TSV, XLS, e XML
   (c) XLS, CSV, e XML
   (d) TSV, CSV, e XML

2. [2.0] Linguagem comum para especificar ontologias biomédicas:
   (a) TXT
   (b) OWL
   (c) XLS
   (d) CSV

3. [2.0] Comando que cria um ficheiro com o conteúdo da página web:
   (a) `curl 'https://www.ebi.ac.uk/...'`
   (b) `curl 'https://www.ebi.ac.uk/...' > file.txt`
   (c) `curl 'https://www.ebi.ac.uk/...' < file.txt`
   (d) `curl 'https://www.ebi.ac.uk/...' | file.txt`

4. [2.0] Comando que seleciona as linhas do ficheiro XML que indicam o organismo *Homo sapiens*:
   (a) `sed 's/<name type="scientific">Homo sapiens<\/name>//' file.xml`
   (b) `gawk -F'[<>]' '{ print $3 }' < file.xml`
   (c) `grep '<name type="scientific">Homo sapiens</name>' file.xml`
   (d) `xmllint --xpath '<name type="scientific">Homo sapiens</name>' file.xml`

5. [2.0] Sequência que emparelha com a expressão regular `A?T*G+C`:
   (a) `GGGC`
   (b) `TTTG`
   (c) `AAAT`
   (d) `ATTC`

Regras: - Todos os exames precisam ser identificados de forma clara e inequívoca; - Não serão aceites respostas a lápis; - Papel de rascunho será distribuído durante o exame; - As respostas podem ser em português ou inglês; - Todos os dispositivos eletrónicos devem ser desligados, exceto relógios regulares; - Nenhuma pergunta semântica será respondida pelo corpo docente; - Cartão de identificação com foto é obrigatória; - Sair da sala só é possível após 30 minutos após o entrega ou desistência do exame; - Os alunos não inscritos para o exame serão aceites no local estando sujeitos às vagas extras disponíveis; - Serão aplicadas as regras éticas académicas, ou seja, não use material não autorizado, não comunique com outros alunos, etc.

Grupo II

6. [2.5] Considere o ficheiro *test.txt* com as seguintes duas linhas:

```
Caffeine is a compound.
caffeine drinks are great.
```

Indique na coluna da direita o resultado dos seguintes comandos:

`cat  test.txt | wc -l`	2
`sed -E 's/^c/\n/'  test.txt | wc -l`	3
`sed -E 's/^c/\n/i'  test.txt | wc -l`	4
`sed -E 's/[^c][ao]/\n/g'  test.txt | wc -l`	7
`sed -E 's/[^c][ao]/\n/ig' test.txt | wc -l`	6

7. [2.5] Considere o ficheiro *test.xml* com o seguinte conteúdo:

```
<entry>
  <reference key="1">
    <title>Molecular cloning of cDNA...</title>
    <dbReference type="PubMed" id="27586648"/>
    <source>
      <tissue>Skeletal muscle</tissue>
    </source>
  </reference>
  <reference key="2">
     <title> Polymorphisms and deduced...</title>
     <dbReference type="PubMed" id="1354642"/>
     <dbReference type="DOI" id="10.1016/0888-
7543(92)90042-Q"/>
  </reference>
</entry>
```

Indique na coluna da direita o resultado dos seguintes comandos:

`xmllint --xpath '//reference' test.xml | grep -c '<[^/]'`	9
`xmllint --xpath '//reference[@key=2]' test.xml | grep -c '<[^/]'`	4
`xmllint --xpath '//dbReference/@id' test.xml | grep -c '<[^/]'`	0
`xmllint --xpath '//dbReference[@id=27586648]/..' test.xml | grep -c '<[^/]'`	5
`xmllint --xpath '//dbReference[@id=27586648]/../source' test.xml | grep -c '<[^/]'`	2

8. [2.5] Considere a versão final do script *getalllabels.sh*:

```
1.  OWLFILE=$1
2.  xmllint --xpath "//*[local-name()='Class']/*[local-name()='hasExactSynonym'
        or local-name()='hasRelatedSynonym' or local-name()='label']" $OWLFILE | \
3.  tr '<>' '\n' | \
4.  grep -v -e ':label' -e ':hasExactSynonym' -e 'hasRelatedSynonym' -e '^$' | \
5.  tr -d '[](){}' | \
6.  sed -E 's/[,:;] .*$//; s/^ *//; s/ *$//' | \
7.  sort -u
```

Indique as alterações necessárias para que este remova os números decimais do início de cada descritor:

```
6. sed -E 's/[,:;] .*$//; s/^ *//; s/ *$//; s/^[0-9]+\.[0-9]+//' | \
```

9. [2.5] Ambos os comandos grep e gawk foram usados para selecionar dados, indique a razão para serem usados ambos os comandos em vez de apenas um deles?

O grep foi usado para selecionar linhas,
enquanto o gawk foi usado para selecionar valores singulares (colunas) nas linhas.