Nome: _____ Número: _____

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Sistemas Baseados em Conhecimento

2021/22 – 1º Semestre

Exame
25 de janeiro de 2022 – 1 hora e 30 minutos

Mestrado em Engenharia Informática
Departamento de Engenharia Informática

Leia com atenção

- O exame tem a duração **máxima** de **1 hora e 30 minutos**.
- Tem um total de três questões, todas com mais de uma alínea.
- Escreva as respostas no espaço destinado.
- Como material de consulta pode **apenas** usar uma folha A4 que traga consigo. Não é permitida a utilização de meios eletrónicos.
- Caso tenha alguma dúvida pergunte.
- Qualquer violação das regras definidas pode implicar a anulação da prova.

Boa sorte!

Pergunta 1 (36 pontos)

Atente ao seguinte excerto de uma ontologia, representada com recurso à linguagem OWL, e responda às perguntas que se seguem.

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://medonto.xy#"
   xml:base="http://medonto.xy"
   xmlns:medonto="http://medonto.xyz#"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:xml="http://www.w3.org/XML/1998/namespace"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <owl:Ontology rdf:about="http://medonto.xyz"/>

  <owl:Class rdf:about="http://medonto.xyz#Disease">
    <rdfs:label>Disease</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://medonto.xyz#Drug">
    <rdfs:label>Drug</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://medonto.xyz#Symptom">
    <rdfs:label>Symptom</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://medonto.xyz#LongTermSymptom">
    <rdfs:subClassOf rdf:resource="http://medonto.xyz#Symptom"/>
    <rdfs:label>Long-term Symptom</rdfs:label>
  </owl:Class>
  <owl:Class rdf:about="http://medonto.xyz#ShortTermSymptom">
    <rdfs:subClassOf rdf:resource="http://medonto.xyz#Symptom"/>
    <rdfs:label>Short-term Symptom</rdfs:label>
  </owl:Class>
  <owl:ObjectProperty rdf:about="http://medonto.xyz#hasSymptom">
    <rdfs:domain rdf:resource="http://medonto.xyz#Disease"/>
    <rdfs:range rdf:resource="http://medonto.xyz#Symptom"/>
  </owl:ObjectProperty>
  <owl:ObjectProperty rdf:about="http://medonto.xyz#usedFor">
    <rdfs:domain rdf:resource="http://medonto.xyz#Drug"/>
    <rdfs:range rdf:resource="http://medonto.xyz#Disease"/>
  </owl:ObjectProperty>
  <rdf:Description rdf:about="http://medonto.xyz#Aestudose">
    <rdf:type rdf:resource="http://medonto.xyz#Disease"/>
    <rdfs:label>Aestudose</rdfs:label>
    <medonto:hasSymptom rdf:resource="http://medonto.xyz#SuddenMemoryLoss"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://medonto.xyz#Distractolepsia">
    <rdf:type rdf:resource="http://medonto.xyz#Disease"/>
    <rdfs:label>Distractolepsia</rdfs:label>
    <medonto:hasSymptom rdf:resource="http://medonto.xyz#SuddenMemoryLoss"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://medonto.xyz#Estudozepan">
    <rdf:type rdf:resource="http://medonto.xyz#Drug"/>
    <rdfs:label>Estudozepan</rdfs:label>
    <medonto:usedFor rdf:resource="http://medonto.xyz#Aestudose"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://medonto.xyz#SuddenMemoryLoss">
    <rdf:type rdf:resource="http://medonto.xyz#ShortTermSymptom"/>
    <rdfs:label>Sudden Memory Loss</rdfs:label>
    <rdfs:label>SML</rdfs:label>
  </rdf:Description>
</rdf:RDF>
```

(a) (12 pontos) Indique, através do prefixo do seu namespace e do seu nome:

- Duas classes definidas:

  Resposta: Duas das seguintes: medonto:Disease, medonto:Drug, medonto:Symptom, medonto:ShortTermSymptom, medonto:LongTermSymptom

- Duas instâncias definidas:

Resposta: Duas das seguintes: medonto:Aestudose, medonto:Distractolepsia, medonto:Etudozepan, medonto:SuddenMemoryLoss

- Duas propriedades definidas:

Resposta: medonto:hasSymptom, medonto:usedFor

(b) (7 pontos) Quando executada sobre este excerto, qual é o resultado da seguinte query SPARQL?

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX medonto:  <http://medonto.xyz#>
SELECT ?x
WHERE {
?y rdf:type medonto:Disease .
?y rdfs:label ?x .
?y medonto:hasSymptom medonto:SuddenMemoryLoss
}
```

Resposta:

?x
Aestudose
Distractolepsia

(c) (7 pontos) Escreva uma query SPARQL para obter todas as propriedades definidas na ontologia que se podem estabelecer entre qualquer coisa e uma doença (*medonto:Disease*). Pode ignorar a definição dos prefixos.

Resposta:

```
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX medonto:  <http://medonto.xyz#>
SELECT ?p
WHERE {
?p rdfs:range medonto:Disease .
}
```

(d) (10 pontos) Recorde-se do conceito de *Linked Data* e da seguinte escala de estrelas atribuídas a conjuntos de dados:

*	On the Web	****	RDF Standards
**	Machine-readable	*****	Linked RDF
***	Non-proprietary format		

Suponha que o excerto apresentado é publicado na *Web*, onde existem várias bases de conhecimento RDF sobre doenças, sintomas e medicamentos. Especificamente, existe uma taxonomia de sintomas de doenças, DST, e um dataset de sintomas caraterizados com base nessa taxonomia, DSD. Indique o que faltaria a este excerto para ter **cinco estrelas**, e dê um exemplo de como isso poderia ser feito, referindo que propriedades, entre as normalmente utilizadas para esse fim, poderia usar.

Resposta: Para ter cinco estrelas, este excerto teria de ser ligado a outros datasets. Isso poderia ser feito, por exemplo, se a propriedade *rdf:type* fosse usada para ligar os sintomas aqui definidos ao seu tipo na DST. No caso da DSD incluir sintomas equivalentes aos aqui definidos (e.g., Sudden Memory Loss), a propriedade *owl:sameAs* também poderia ser usada para os ligar.

Pergunta 2 (34 pontos)

Suponha que quer obter mais informação acerca das palavras mencionados num conjunto de publicações num rede social, tais como a seguinte:

Os ZZZZ têm uma música que é interessante, mas que tem uma letra de treta.

E que tem acesso a num repositório que inclui as seguintes entradas (notação N-triples):

```
letra ex:sense letra_1
letra ex:sense letra_2
letra ex:sense letra_3
letra rdfs:label "letra"

letra_1 ex:pos "Noun"
letra_1 ex:gloss "cada um dos caracteres do alfabeto"
letra_2 ex:pos "Noun"
letra_2 ex:gloss "sentido expresso claramente pelo que se escreve"
letra_3 ex:pos "Noun"
letra_3 ex:gloss "palavras que acompanham música"

letra_4 ex:pos "Verb"
letra_4 ex:p3ps letrar_2

letrar_1 ex:pos "Verb"
letrar_1 ex:gloss "adquirir conhecimento literários"
letrar_2 ex:pos "Verb"
letrar_2 ex:gloss "compor a letra de"

música_1 ex:gloss "arte e técnica de combinar sons de forma harmoniosa"
música_1 ex:pos "Noun"
música_2 ex:gloss "concerto vocal ou instrumental"
música_2 ex:pos "Noun"
música_3 ex:gloss "treta, lábia"
música_3 ex:pos "Noun"

treta_1 ex:gloss "destreza na luta ou no jogo da esgrima"
treta_1 ex:pos "Noun"
treta_2 ex:gloss "palavras para enganar; lábia"
treta_2 ex:pos "Noun"
```

(a) (8 pontos) Ao nível do pré-processamento, indique uma **sub-tarefa** do domínio do Processamento de Linguagem Natural essencial ao início deste processo e **explique** brevemente em que consiste.

> **Resposta:** Uma sub-tarefa comum a muitas tarefas PLN é a *tokenização*, que consiste em separar o texto pelas suas unidades mais pequenas que podem ter significando, neste caso, palavras.
>
> Outra tarefa que poderia ser útil para a desambiguação seria a *part-of-speech tagging*, em que cada *token* teria atribuída uma etiqueta correspondente à sua função gramatical, algo que poderia ser considerado para eliminar sentidos correspondentes à mesma palavra, mas com outra função (e.g., *letra* como forma do verbo *letrar*).

(b) (8 pontos) Depois do pré-processamento, descreva a forma mais **simples e direta** de identificar que palavras do texto estão representadas no repositório.

> **Resposta:** A forma mais simples de o fazer será selecionar todos os recursos no repositório cuja *rdfs:label* seja igual a uma *token* do texto.

(c) (10 pontos) Considerando a publicação exemplo e os sentidos representados no repositório, indique quais os sentidos das palavras *letra*, *música* e *treta* que seriam atribuídos pelo **algoritmo de Lesk**. Justifique a sua resposta.

Resposta:

letra_3, porque tem na definição *música*, que aparece no texto, e *palavras*, que aparece na definição de treta_2.

música_3, porque tem na definição *treta*, que aparece no texto.

treta_2, porque tem na definição *palavras*, que aparece na definição de letra_3, e *lábia*, que aparece na definição de *música_3*.

(d) (8 pontos) Escreva uma representação da frase recorrendo a predicados de lógica de primeira ordem.

Resposta: $\exists_{x,y} Musica(x) \wedge autorDe(ZZZZ, x) \wedge interessante(x) \wedge temLetra(x, y) \wedge treta(y)$

Pergunta 3 (30 pontos)

Considere o paradigma de *Open Information Extraction*...

(a) (10 pontos) Qual a principal diferença entre este paradigma e a extração de relações mais tradicional?

> **Resposta:** Em *Open Information Extraction* o objetivo é extrair todas as relações identificadas no texto, com base em padrões que não dependem diretamente das palavras, mas sim de outras caraterísticas (e.g., *parts-of-speech*, dependências), enquanto que na extração de relações tradicional é necessário definir ou aprender padrões para cada relação que se pretende extrair.

(b) (10 pontos) Discuta uma das limitações deste paradigma quando aplicado ao enriquecimento de bases de conhecimento.

> **Resposta:** Como se pode extrair qualquer tipo de relação, identificadas por palavras no texto, é normal extrair muitas relações que não estão definidas na base de conhecimento a enriquecer. Uma forma de ultrapassar essa limitação será tentar normalizar o nome das relações extraídas, mas pode não ser suficiente.

(c) (10 pontos) Imagine agora que tem acesso a:

- Uma grande coleção de textos;
- Um sistema de *Information Retrieval* tradicional, onde a coleção está indexada, de forma a permitir pesquisas com base em palavras-chave;
- Um sistema de *Open Information Extraction* (OIE).

Explique como poderia tirar partido do sistema de OIE, de modo a permitir pesquisas semânticas, referindo, a existir, algum pré-processamento adicional necessário, e eventuais alterações ao nível da representação dos dados, das *queries* e dos resultados.

> **Resposta:** Um sistema de *Open Information Extraction* poderia ser utilizado para extrair relações a partir dos documentos. Essas relações podem ser representadas em triplos RDF, e assim interrogáveis via SPARQL. Uma forma de integrar pesquisas semânticas passaria por permitir precisamente pesquisas SPARQL, sobre uma representação dos dados em RDF. Em vez de documentos, os resultados poderiam passar a ser apresentados como uma lista de entidades.