Nome: _____ Número: _____

Sistemas Baseados em Conhecimento
2021/22 – 1º Semestre

Exame de Recurso
4 de fevereiro 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 (44 pontos)

Atente ao seguinte excerto RDF Turtle e responda às perguntas que se seguem.

```
@prefix lres:   <http://lres.xyz/> .
@prefix rdfs: <https://www.w3.org/TR/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix schema: <https://schema.org/> .
@prefix dbr:    <https://dbpedia.org/resource/> .

lres:enwn
        a       dcat:Dataset ;
        lres:lang       "English";
        rdfs:label      "Princeton WordNet";
        rdfs:label      "English WordNet";
        dct:creator     lres:CFellbaum;
        dct:licence     <https://www.dcat-ap.de/def/licenses/bsd> ;
        owl:sameAs      dbr:WordNet .

lres:ontopt
        a       dcat:Dataset ;
        lres:lang       "Portuguese";
        rdfs:label      "Onto.PT" ;
        dct:creator     lres:HGO;
        dct:licence     <https://www.dcat-ap.de/def/licenses/bsd> .

lres:tales
        a       dcat:Dataset ;
        lres:lang       "Portuguese";
        rdfs:label      "TALES";
        dct:creator     lres:HGO;
        dct:licence     <http://dcat-ap.de/def/licenses/cc-zero> .

lres:HGO
        a       schema:Person ;
        rdfs:label      "Hugo Gonçalo Oliveira" .

lres:CFellbaum
        a       schema:Person ;
        rdfs:label      "Christiane Fellbaum";
        owl:sameAs      dbr:Christiane_Fellbaum ;
        owl:sameAs <http://viaf.org/viaf/56454343> .
```

(a) (8 pontos) Indique dois, e **apenas dois**, vocabulários conhecidos que são utilizados neste excerto e diga para que são normalmente usados.

Resposta: Vocabulários usados (escolher dois):

- RDFS, uma extensão do RDF que permite representar conhecimento mais abstrato, nomeadamente classes e propriedades de propriedades.
- OWL, um vocabulário para representar ontologias, i.e., conhecimento rico e complexo acerca de coisas, grupos de coisas e relações entre elas;
- Dublin Core Terms (dct), que faz parte da iniciativa Dublin Core, para descrever propriedades de documentos;
- Data Catalog Vocabulary (dcat), um vocabulário para facilitar a interoperabilidade e organização de dados;
- Schema, um vocabulário para representar dados na Internet;
- DBPedia resource (dbr), um vocabulário para representar instâncias na DBPedia.

(b) (10 pontos) Apenas com base no excerto apresentado, indique duas propriedades: uma que **pudesse ser** formalizada como *owl:DatatypeProperty* e outra como *owl:ObjectProperty*. Explique a sua decisão.

> **Resposta:** Uma *owl:DatatypeProperty* é estabelecida entre um recurso e um tipo literal. Neste excerto, temos duas propriedades entre recursos e *strings*: *lres:lang* e *rdfs:label*.
> Uma *owl:ObjectProperty* é estabelecida entre dois recursos. Neste excerto temos quatro propriedades deste tipo: *dct:creator*, *dct:license*, *owl:sameAs*, e *rdf:type* (a).

(c) (8 pontos) Explique o significado da propriedade *rdfs:label* e a utilidade de, para o mesmo recurso, haver **dois ou mais valores possíveis** para ela.

> **Resposta:** Esta propriedade serve para indicar uma representação dos recursos a ser consumida por humanos (linguagem natural). Quando a mesma coisa pode ter mais do que um nome / ser referida de diferentes formas em linguagem natural (sinonímia), faz sentido ter mais do que um valor para a propriedade *rdfs:label*.

(d) (10 pontos) Qual o resultado da seguinte query SPARQL neste excerto? Apresente-o sob a forma de tabela.

```
PREFIX dcat:  <http://www.w3.org/ns/dcat#>
PREFIX rdfs:  <https://www.w3.org/TR/rdf-schema#>
PREFIX lres:  <http://lres.xyz/>
SELECT DISTINCT ?x ?y
WHERE {
?dataset a dcat:Dataset .
?dataset rdfs:label ?x .
?dataset lres:lang ?y
}
```

Resposta:

?x	?y
‘‘Princeton WordNet’’	‘‘English’’
‘‘English WordNet’’	‘‘English’’
‘‘Onto.PT’’	‘‘Portuguese’’
‘‘TALES’’	‘‘Portuguese’’

(e) (8 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		

Se este excerto fosse publicado na *Web*, de forma aberta, **quantas estrelas** teria nesta escala? Justifique a sua resposta.

Resposta: Teria 5 estrelas, porque estaria publicado na Web (1), num formato consumível por outros sistemas (2), não proprietário (3) e seguindo o standard RDF Turtle (4), para além de estar ligada a outros conjuntos de dados (5), nomeadamente à DBPedia, através da propriedade *owl:sameAs*, ou a licenças representadas no Data Catalog Vocabulary, através da propriedade *dct:licence*.

Pergunta 2 (56 pontos)

Considere a seguinte coleção com quatro documentos:

ID	Conteúdo
D1	Tomás está com Covid19.
D2	O fim da pandemia.
D3	O início da pandemia.
D4	Covid19: o fim?

Suponha que é indexada por um sistema de Recuperação de Informação (*Information Retrieval*) em que os documentos são representados de acordo com um modelo vetorial baseado na **contagem de ocorrências** dos termos, depois de um pré-processamento onde todos os carateres são convertidos em minúsculas e os sinais de pontuação são removidos (nada mais).

(a) (10 pontos) Quais as **dimensões** da matriz termo-documento com as condições indicadas? Justifique.

> **Resposta:** 4 documentos × 9 tokens diferentes

(b) (10 pontos) Suponha que a função utilizada para o ordenamento de documentos é a similaridade do coseno, calculada da seguinte forma:

$$\cos(\vec{u},\vec{v}) = \frac{\vec{u}.\vec{v}}{|\vec{u}||\vec{v}|} = \frac{\sum_{i=0}^{n} u_i \times v_i}{\sqrt{\sum_{i=0}^{n} u_i^2} \times \sqrt{\sum_{i=0}^{n} v_i^2}}$$

Nestas condições, que documento vai aparecer na **primeira posição** para a *query*: *pelo fim da covid19* ? Justifique com cálculos.

> **Resposta:** O documento D4 vai aparecer na primeira posição, porque é ele que maximiza a similaridade com a *query*.
>
> | token | D1 | D2 | D3 | D4 | q |
> |---|---|---|---|---|---|
> | tomás | 1 | 0 | 0 | 0 | 0 |
> | está | 1 | 0 | 0 | 0 | 0 |
> | com | 1 | 0 | 0 | 0 | 0 |
> | covid19 | 1 | 0 | 0 | 1 | 1 |
> | o | 0 | 1 | 1 | 1 | 0 |
> | fim | 0 | 1 | 0 | 1 | 1 |
> | da | 0 | 1 | 1 | 0 | 1 |
> | pandemia | 0 | 1 | 1 | 0 | 0 |
> | início | 0 | 0 | 1 | 0 | 0 |
>
> $cos(q, D1) = \frac{\sqrt{3}}{6}$, $cos(q, D2) = \frac{\sqrt{3}}{3}$, $cos(q, D3) = \frac{\sqrt{3}}{6}$, $cos(q, D4) = \frac{2}{3}$

(c) (26 pontos) Um sistema como o apresentado terá um conjunto de limitações conhecidas. Por exemplo, ele vai considerar que todos os termos têm a mesma relevância, não vai conseguir associar diferentes formas da mesma palavra, nem vai considerar o significado das palavras e das frases. Neste contexto, responda às seguintes:

1. Uma técnica comum para calcular a relevância dos termos num documento é denominada $TF.IDF$. Explique **em que consiste** esta técnica e qual a sua **relação com as chamadas** *stopwords*.

> **Resposta:** Esta técnica considera que a relevância de um termo num documento é dada pela sua frequência no documento (TF) e pelo número de documentos em que esse termo ocorre (DF), sendo que a primeira contribui positivamente e a segunda negativamente ($\frac{TF}{DF}$). Por exemplo, por serem muito

usadas, é expectável que as *stopwords* tenham uma frequência alta em qualquer documento. No entanto, elas também vão ocorrer em vários documentos, o que diminuirá a sua relevância.

2. Explique uma tarefa que poderia incluir no pré-processamento para permitir a associação de **diferentes formas da mesma palavra**.

Resposta: Duas tarefas comuns para o fazer são o *stemming* e a lematização. (explicar uma ou outra) Stemming consiste em aplicar um conjunto de regras para remover as terminações das palavras, o que se aplica bem a flexões regulares, mas nem sempre resulta numa palavra válida. No caso da lematização, todas as formas são convertidas na forma em que aparecem num dicionário (lema).

3. Indique uma tarefa do domínio do Processamento de Linguagem Natural que permita comparar o **significado de pares de frases**. Explique detalhadamente como poderia ser aplicada ao cenário desta pergunta, referindo, se necessário, dados a usar, *features*, representações e métricas.

   Resposta: A Similaridade Semântica Textual é uma tarefa que tem como objetivo calcular a similaridade semântica entre sequências de texto, e que poderia ser aplicada a este cenário. Ela pode basear-se num conjunto de features lexicais (palavras, lemas), sintáticas (parts-of-speech, dependências), semânticas formais (sentidos, relações) e semânticas distribucionais (word embeddings). Uma abordagem possível passa por treinar uma rede neuronal para calcular *embeddings* de frases, e usar esses embeddings para representar os documentos da coleção. Uma vez representadas através de *embeddings*, a similaridade entre frases pode ser calculada através do coseno dos seus *embeddings*.

(d) (10 pontos) Considere agora uma tarefa diferente, a Extração de Informação (*Information Extraction*). Descreva duas das suas **sub-tarefas** e exemplifique o possível resultado de cada uma quando aplicada ao documento D1.

Resposta: A Extração de Informação tem como objetivo extrair informação estruturada a partir de texto, não estruturado, de forma a que seja possível a sua utilização para popular bases de conhecimento. Duas das suas sub-tarefas são:

- O Reconhecimento de Entidades Mencionadas, que tem como objetivo identificar e classificar menções a entidades feitas no texto. No caso de D1, poderiam ser identificadas as entidades "Tomás" e "Covid19", e classificadas respetivamente como PESSOA e como OUTRO ou DOENÇA (se esta classe existir na ontologia adoptada).
- Extração de relações, que tem como objetivo identificar e classificar relações entre conceitos referidos no texto. No caso de D1, poderia ser extraída a seguinte relação: *estáCom(Tomás, Covid19)*.