Nome: _____ Número: _____

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Conhecimento e Linguagem

2025/26 – 1º Semestre

Exame
9 de janeiro de 2026 – 1 hora e 45 minutos

Mestrado em Engenharia Informática
Departamento de Engenharia Informática

Leia com atenção

- O exame tem a duração **máxima** de **1 hora e 45 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.
- Qualquer violação das regras definidas pode implicar a anulação da prova.
- Caso tenha alguma dúvida relativamente à interpretação de uma pergunta ou das regras, pergunte.

Boa sorte!

Pergunta 1 (32 pontos)

Atente ao seguinte excerto simplificado da WordNet, representado com recurso à linguagem RDF, usando os vocabulários RDFS e OWL, e responda às perguntas que se seguem.

```
@prefix wn: <http://www.w3.org/2006/03/wn/wn/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

wn:Synset rdf:type owl:Class ;
          rdfs:subClassOf owl:Thing .

wn:AdjectiveSynset rdf:type owl:Class ;
                   rdfs:subClassOf wn:Synset .

wn:NounSynset rdf:type owl:Class ;
              rdfs:subClassOf wn:Synset .

wn:VerbSynset rdf:type owl:Class ;
              rdfs:subClassOf wn:Synset .

wn:gloss rdf:type owl:DataTypeProperty ;
           rdfs:domain wn:Synset ;
           rdfs:range rdfs:Literal .

wn:hyponymOf rdf:type owl:ObjectProperty , owl:TransitiveProperty ;
             rdfs:domain wn:NounSynset ;
             rdfs:range wn:NounSynset .

wn:partMeronymOf rdf:type owl:ObjectProperty ;
                 rdfs:domain wn:NounSynset ;
                 rdfs:range wn:NounSynset .

wn:keyboard-n-1 rdf:type wn:NounSynset ;
                rdfs:label "keyboard" .

wn:keyboard-instrument-n-1 rdf:type wn:NounSynset ;
                           rdfs:label "keyboard instrument" .

wn:organ-n-1 rdf:type wn:NounSynset ;
             rdfs:label "organ" ;
             wn:gloss "a fully differentiated structural and functional unit in an animal that is
                  specialized for some particular function" .

wn:organ-n-2 rdf:type wn:NounSynset ;
             rdfs:label "organ" , "pipe organ" ;
             wn:gloss "wind instrument whose sound is produced by means of pipes arranged in sets supplied
                  with air from a bellows and controlled from a large complex musical keyboard" .

wn:pipe-n-5 rdf:type wn:NounSynset ;
            rdfs:label "organ pipe" , "pipe" , "pipework" .

wn:wing-n-1 rdf:type wn:NounSynset ;
            rdfs:label "wing" .

wn:foot-n-6 rdf:type wn:NounSynset ;
            rdfs:label "foot" ;
            wn:hyponymOf wn:organ-n-1 .

wn:hand-n-1 rdf:type wn:NounSynset ;
            rdfs:label "hand", "manus", "mitt", "paw" .

wn:keyboard-n-1 wn:partMeronymOf wn:organ-n-2 .
wn:organ-n-2 wn:hyponymOf wn:keyboard-instrument-n-1 .
wn:pipe-n-5 wn:partMeronymOf wn:organ-n-2 .
wn:wing-n-1 wn:hyponymOf wn:organ-n-1 .
wn:foot-n-6 wn:hyponymOf wn:organ-n-1 .
```

(a) (8 pontos) Indique, através do prefixo do seu namespace e do seu nome.

Duas **classes definidas**:

Resposta: Duas das seguintes: wn:Synset, wn:AdjectiveSynset, wn:VerbSynset, wn:NounSynset

Duas **propriedades usadas**:

Resposta: Duas das seguintes: wn:hyponymOf, wn:partMeronymOf, wn:gloss, rdf:type, rdfs:label, rdfs:subClassOf, rdfs:range, rdfs:domain

(b) (8 pontos) Escreva o resultado da seguinte query SPARQL, quando executada sobre este excerto, sem realizar a inferência de conhecimento implícito.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX wn:  <http://www.w3.org/2006/03/wn/wn/>
SELECT DISTINCT ?a ?b
WHERE {
    ?x wn:hyponymOf ?z.
    ?x rdfs:label ?a .
    ?z rdfs:label ?b .
}
```

Resposta:

?a	?b
''pipe organ''	''keyboard instrument''
''organ''	''keyboard instrument''
''wing''	''organ''
''foot''	''organ''

(c) (8 pontos) Suponha que está a pensar em adicionar um novo recurso, `wn:RTV`, que, em lógica de predicados de primeira ordem, seria descrito da seguinte forma:
$wn{:}Synset(x) \wedge \exists_y(owl{:}ObjectProperty(x, y) \wedge wn{:}VerbSynset(y))$
Descreva este tipo de recurso em **linguagem natural**.

Resposta: Um wn:RTV é um wn:Synset que tem pelo menos uma relação com um wn:Synset do tipo wn:VerbSynset.

(d) (8 pontos) Suponha que quer enriquecer o excerto com mais relações de hiponímia (`wn:hyponymOf`) obtidas diretamente a partir de um LLM causal esquerda-direita. Escreva uma prompt *two-shot learning* para obter o hiperónimo (inverso de hipónimo) de *keyboard*, recorrendo a exemplos do excerto RDF.

Resposta:
```
Complete the following list:
foot hyponymOf organ
organ hyponymOf keyboard instrument
keyboard hyponymOf
```

Pergunta 2 (20 pontos)

Responda às seguintes perguntas relativamente à criação e utilização de *Large Language Models* (LLMs).

(a) (8 pontos) Descreva duas possíveis funções do humano no desenvolvimento de LLMs.

> **Resposta:** Duas das seguintes: recolha de dados para pré-treino; escolha de hiperparâmetros e treino; anotação / curadoria de dados para fine-tuning; alinhamento através do ordenamento de respostas / feedback; avaliação manual.

(b) (12 pontos) Considere a afirmação: *Os LLMs podem gerar explicações plausíveis para as suas decisões.*

1. Indique como seria classificado um método de Inteligência Artificial Explicável baseado na afirmação: Intrínseco *vs* extrínseco? Agnóstico ao modelo *vs* dependente do modelo? Justifique a sua resposta

> **Resposta:** O método é intrínseco, porque é o LLM que classifica e que também gera as explicações, e dependente do modelo, porque só funciona para LLMs.

2. Tendo em conta a flexibilidade da interação e o tipo de explicações que é possível obter, compare a utilização dos seguintes para resposta automática a perguntas: (i) um LLM; (ii) um *Knowledge Graph*.

> **Resposta:** Ao contrário de um LLM, com o qual é possível interagir em linguagem natural e fazer as perguntas que entendermos, a interação com um Knowledge Graph implica a utilização de uma linguagem formal, e está limitada aos seus conteúdos e estrutura. No entanto, a resposta de um Knowledge Graph vai sempre derivar da sua estrutura, o que torna o processo inerentemente interpretável, algo que não acontece num LLM, que funciona como uma caixa-negra.

Pergunta 3 (48 pontos)

Considere o seguinte corpo com cinco frases e os seus respetivos embeddings.

ID	Text	Embedding
S1	Louis sings and plays the trumpet well.	[1 1 0 0 1]
S2	Joe plays the guitar beautifully.	[0 0 1 0 0]
S3	Ray plays the organ with his hands.	[0 1 0 1 1]
S4	Jimi plays the guitar.	[0 0 1 0 0]
S5	Dave plays the drums, the guitar and he sings.	[1 1 1 0 1]

(a) (8 pontos) Apresente os cálculos necessários para, recorrendo à **similaridade do cosseno**, encontrar a frase mais próxima de *Who sings and plays?*, com embedding $X = [1\ 1\ 0\ 0\ 0]$. Considere que $\sqrt{3} \approx 1.7$.

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x}.\vec{y}}{||\vec{x}|| \times ||\vec{y}||} = \frac{\sum_{i=0}^{n} x_i \times y_i}{\sqrt{\sum_{i=0}^{n} x_i^2} \times \sqrt{\sum_{i=0}^{n} y_i^2}}$$

Resposta: $|X| = \sqrt{1+1} = \sqrt{2}$
$cos(X, S2) = cos(X, S4) = 0$
$cos(X, S1) = \frac{2}{\sqrt{2}.\sqrt{3}}$
$cos(X, S3) = \frac{1}{\sqrt{2}.\sqrt{3}}$
$cos(X, S5) = \frac{2}{\sqrt{2}.\sqrt{4}}$
Ignorando $\sqrt{2}$ no denominador, verifica-se que o cosseno mais elevado é obtido com S1.

(b) (10 pontos) Suponha que as frases do corpo são os únicos dados usados no treino de um modelo de linguagem baseado em trigramas ($n = 3$).

1. Apresente os cálculos necessários para calcular a palavra mais provável para completar a frase: *I know that Joan plays the...*

   Resposta: $P(guitar|plays, the) = \frac{2}{5}$
   $P(trumpet|plays, the) = \frac{1}{5}$
   $P(organ|plays, the) = \frac{1}{5}$
   $P(drums|plays, the) = \frac{1}{5}$
   A palavra será *guitar*.

2. O corpo tem 20 tokens diferentes. Sem qualquer optimização, **quantos parâmetros** são necessários para representar o modelo? Justifique.

   Resposta: $|V| = 20$. O número de parâmetros é igual a $|V|^n$, ou seja, $20^3 = 8000$.

(c) (8 pontos) A Análise de Subjetividade é uma sub-tarefa da Análise de Sentimentos. Indique quais frases do corpo são **subjetivas** e explique porquê.

Resposta: As frases subjetivas são aquelas que transmitem opiniões. neste caso, $S1$ e $S2$.

(d) (8 pontos) A frase S3 utiliza a palavra *organ* que pode ter mais do que um sentido, aqui denominados por `organ-1` e `organ-2`. Utilize o método Naive Bayes para desambiguar o seu sentido nesta frase, após lematização, e considerando as seguintes probabilidades:

- À priori: $P(organ\text{-}1) = 0.6$, $P(organ\text{-}2) = 0.4$
- Condicionais:
  - $P(play|organ\text{-}1) = 0.9$, $P(play|organ\text{-}2) = 0.1$
  - $P(hand|organ\text{-}1) = 0.3$, $P(hand|organ\text{-}2) = 0.7$

Naive Bayes generalizado: $P(H, E_1, ..., E_n) = \alpha P(H).\prod_{i=0}^{n} P(E_i|H)$

Resposta:
$P(o1|S3) = \alpha P(o1).P(play|o1).P(hand|o1) = 0.6 \times 0.9 \times 0.3 = \alpha 0.162$
$P(o2|S3) = \alpha P(o2).P(play|o2).P(hand|o2) = 0.4 \times 0.1 \times 0.7 = \alpha 0.028$
O sentido será `organ-1`.

(e) (14 pontos) O Reconhecimento de Entidades Mencionadas (NER) pode ser realizado com recurso a raciocínio probabilístico com tempo. Considere o reconhecimento de entidades **Pessoa** e:

1. Indique o nome de uma **tarefa de inferência Bayesiana** adequada, em inglês ou português, e o seu objetivo neste contexto.

   Resposta: *Most likely explanation*, que teria como objetivo atribuir a melhor sequência de etiquetas para uma sequência de tokens de entrada.

2. Descreva um **formato de dados** de treino adequados e exemplifique-o para a frase S2.

   Resposta: Se optarmos pelo formato BIO, teremos três etiquetas: O para marcar tokens que não interessam, B-PER a indicar o início do nome de uma pessoa, I-PER a indicar um token que faz parte do nome de uma pessoa.
   Aplicando à frase S2, teremos: `[B-PER O O O O]`

3. Indique o nome de um **algoritmo de aprendizagem** ou descreva uma **arquitetura** adequada a este objetivo.

   Resposta: Um dos seguintes: Hidden Markov Model; Conditional Random Fields; classificação da representação de cada token, obtida por um encoder (e.g., BERT).