Universidade Federal do Rio Grande do Sul - Instituto de Informática
Departamento de Informática Aplicada - Processamento de Linguagem Natural
Prof. Dennis Giovani Balreira - 2024/2 - Turma U

Nome: _____ **Cartão:** _____ Data: 27/11/2024

Prova 1

	Parte Dissertativa	Parte Objetiva	Total
Pontuação	**6,0**	**4,0**	**10,0**
Nota			

Instruções

- A prova é individual e com consulta a uma folha A4 manuscrita de próprio punho frente e verso;
- O tempo de duração da prova é de 1h e 40 minutos;
- A prova possui questões dissertativas e objetivas, totalizando 10,0 pontos;
- É obrigatório o uso de caneta esferográfica azul ou preta ou lápis (apenas nas questões dissertativas);
- Responda às questões dissertativas com letra legível apenas nas folhas de rascunho fornecidas;
- Responda às questões objetivas diretamente na prova marcando sobre a própria letra dentre (a),(b),(c),(d),(e) apenas em uma única alternativa que julgar correta;
- Identifique todas as folhas de prova e rascunho com nome completo e número do cartão;
- As definições e o conteúdo desta avaliação são referentes ao material abordado em aula.

Parte I - Dissertativa [Questões 1 a 6] [6,0 pontos]

Questões	1	2	3	4	5	Subtotal
Pontuação	1,0	1,0	1,0	1,5	1,5	6,0
Nota						

Questão 1 **(Estatística de dados, pré-processamento) (1,0 ponto)**

Considere o seguinte texto de entrada:

“O gato viu o rato. O rato correu. O gato correu atrás do rato.”

a) Apresente a frequência das palavras. Por que é possível dizer que o texto segue em linhas gerais a Lei de Zipf? Justifique sua resposta.

b) Qual o texto resultante **final** após a aplicação da etapa de pré-processamento considerando, nesta ordem: (i) case folding, (ii) remoção de acentos, (iii) caracteres especiais e pontuação, (iv) remoção de stop words, (v) stemming.

Questão 2 **(Modelos tradicionais de representação textual) (1,0 ponto)**

Considere o seguinte conjunto de três documentos:

d1: "o sol brilha"
d2: "a lua brilha"
d3: "o céu tem sol"

Apresente a matriz resultante considerando o vetor de vocabulário como [“sol”, “brilha”, “lua”, “céu”] utilizando a abordagem Bag of Words (BoW). Assuma que as linhas devem corresponder aos documentos e as colunas aos termos.

Questão 3 (Avaliação de modelos supervisionados) (1,0 pontos)

Considere um dataset com 10.000 (dez mil) textos contendo letras de músicas e suas respectivas classificações como “boa música” e “má música”. Supondo que se deseja treinar um classificador a partir deste dataset, considere as perguntas abaixo.

a) Por que é possível utilizar técnicas de aprendizado de máquina supervisionado para este exemplo?

b) Suponha que o dataset possui 9.900 textos classificados como “boa música” e apenas 100 como “má música”. Como o desbalanceamento das classes pode impactar o desempenho do modelo e a confiabilidade da acurácia como métrica de avaliação? Qual métrica alternativa deve ser utilizada?

Questão 4 (Word embeddings fixas) (1,5 pontos)

Considere o uso de word embeddings fixas, como Word2Vec e GloVe, em tarefas de Processamento de Linguagem Natural (PLN). Responda às perguntas abaixo:

a) Como word embeddings fixas se diferem das representações tradicionais, como Bag of Words e TF-IDF.

b) Por que o treinamento de word embeddings fixas no modelo Word2Vec é considerado “auto-supervisionado”?

c) Apresente uma situação em que o uso de embeddings fixas pode ser limitado e justifique o uso de representações contextuais (como BERT) como alternativa.

Questão 5 (Word embeddings contextuais e LLMs) (1,5 pontos)

Word embeddings contextuais e modelos de linguagem de larga escala revolucionaram PLN nos últimos anos, principalmente com o advento da arquitetura Transformer. Sobre este contexto, considere as perguntas abaixo:

a) Explique, de forma geral, por que modelos baseados em Transformer apresentam desempenhos melhores do que outras arquiteturas neurais para texto, como RNN e LSTM.

b) Qual é a principal estratégia de transfer learning estudada para adaptar modelos pré-treinados como o BERT a uma tarefa específica, como análise de sentimentos para um dataset específico?

c) Qual abordagem da arquitetura Transformer (entre encoder-decoder, encoder-only e decoder-only) tende a produzir melhores resultados para tarefas ditas “seq2seq” (texto para texto), geralmente usadas em tarefas de tradução automática? Justifique brevemente.

Fim da parte dissertativa.

Nome: _____ **Cartão:** _____ Data: 27/11/2024

Parte II - Objetiva [Questões 6 a 13] [4,0 pontos]

Questões	6	7	8	9	10	11	12	13	Subtotal
Pontuação	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	4,0
Acertos									

Questão 6 (Processamento de Linguagem Natural) (0,5 ponto)

Processamento de Linguagem Natural (PLN) é uma área interdisciplinar que une conhecimentos de linguística, ciência da computação e inteligência artificial, sendo amplamente utilizada em aplicações práticas. Sobre esta área, é **INCORRETO** afirmar que:

a) Abrange tanto tarefas de análise de textos, como a classificação gramatical e extração de informações, quanto tarefas de síntese de textos, como a tradução automática e a geração de respostas em chatbots.

b) Modelos baseados em regras, como parsers sintáticos manuais, foram amplamente utilizados em etapas iniciais do PLN, mas possuem limitações ao lidar com variação linguística e ambiguidades contextuais.

c) A qualidade dos modelos de PLN depende unicamente do tamanho dos datasets utilizados para o treinamento, independentemente de sua diversidade ou representatividade.

d) As técnicas modernas de PLN, como redes neurais e embeddings contextuais, têm superado abordagens tradicionais ao capturar melhor as relações semânticas e contextuais das palavras, mas costumam ser custosas e complexas de interpretar.

e) A divisão do PLN em NLU (Interpretação de Linguagem Natural) e NLG (Geração de Linguagem Natural) reflete os objetivos específicos dessas subáreas: compreender o significado de textos de entrada e produzir respostas ou textos significativos, respectivamente.

Questão 7 (Pré-processamento) (0,5 ponto)

Sobre expressões regulares, pré-processamento de texto e distância mínima de edição, considere as asserções abaixo:

( ) Uma desvantagem das expressões regulares é que elas não podem ser usadas para encontrar padrões complexos, como palavras que aparecem em diferentes partes de um texto.

( ) A remoção de stopwords deve ser sempre realizada como parte do pré-processamento de texto, independentemente da aplicação-alvo.

( ) A tokenização, no pré-processamento de texto, consiste em dividir o texto em unidades menores chamadas "tokens".

( ) A distância de Levenshtein ignora a ordem dos caracteres nas strings comparadas, focando apenas na frequência dos caracteres.

Preenchendo cada asserção com V para verdadeiro e F para falso, a ordem correta, de cima para baixo, respectivamente, é:

a) F, F, F, V.
b) F, F, V, F.
c) F, V, V, F.
d) V, V, F, F.
e) V, F, V, V.

Questão 8 **(Representação de textos com técnicas tradicionais) (0,5 ponto)**

O que o cálculo do TF-IDF (Term Frequency - Inverse Document Frequency) representa no contexto do Processamento de Linguagem Natural?

a) A frequência absoluta de um termo em um único documento, sem considerar seu contexto em outros documentos.

b) A relação entre a frequência de um termo em um documento específico e sua frequência em todos os documentos do corpus, destacando termos mais relevantes e menos comuns.

c) A probabilidade de um termo aparecer em um documento, baseada na similaridade entre documentos do corpus.

d) A média ponderada da contagem de palavras em todos os documentos do corpus, utilizada para normalizar dados textuais.

e) A distância semântica entre palavras em um documento, considerando apenas a frequência relativa de termos.

Questão 9 **(Aprendizado supervisionado) (0,5 ponto)**

Considere os nomes e definições de conceitos na tabela abaixo relacionados à classificação de texto e avaliação de desempenho em aprendizado supervisionado:

I. Naive Bayes	A. Técnica que combina os resultados de múltiplos classificadores para melhorar a robustez e o desempenho do modelo.
II. Overfitting	B. Ocorre quando um modelo aprende padrões específicos do conjunto de treinamento, prejudicando seu desempenho em novos dados.
III. Generalização	C. Algoritmo probabilístico simples que assume independência entre as características.
IV. F-Score	D. Mede a capacidade de um modelo de aplicar o conhecimento adquirido no treinamento a dados nunca vistos.
V. Ensemble learning	E. Métrica que combina precisão e revocação em um único valor.

A alternativa que apresenta a correspondência correta entre cada nome (I-V) e sua respectiva definição (A-E) é:

a) I-C, II-B, III-D, IV-E, V-A.
b) I-B, II-C, III-E, IV-A, V-D.
c) I-E, II-A, III-C, IV-B, V-D.
d) I-C, II-E, III-D, IV-B, V-A.
e) I-D, II-A, III-C, IV-E, V-B.

Nome: _____ **Cartão:** _____ Data: 27/11/2024

Questão 10 **(Word Embeddings Fixas) (0,5 ponto)**

Sobre word embeddings fixas, qual das afirmações abaixo está correta?

a) Word embeddings fixas geram representações para cada ocorrência de palavra, onde o vetor de uma palavra muda dependendo da frase em que aparece.

b) Modelos como ELMo e BERT são exemplos de técnicas que produzem embeddings fixas, sendo baseadas na arquitetura Transformer.

c) Word embeddings fixas não conseguem capturar capturar quaisquer tipos de relações semânticas, como sinônimos ou palavras com significados semelhantes.

d) Embeddings fixas podem ser geradas utilizando redes neurais simples, não necessitando de redes especializadas em sequências, como RNNs ou LSTMs.

e) Uma limitação das embeddings fixas é que elas são treinadas individualmente para cada tarefa de PLN, sem reutilização entre diferentes modelos.

Questão 11 **(Redes neurais para textos) (0,5 ponto)**

Sobre redes neurais para textos, analise as afirmações abaixo:

I. Redes Neurais Recorrentes (RNNs) são capazes de processar sequências de texto, mas sofrem de limitações como o desaparecimento do gradiente em sequências longas.

II. Redes LSTM utilizam uma célula de memória para armazenar informações relevantes por curtos períodos, enquanto ignoram dependências de longo prazo.

III. Na arquitetura Transformer, o mecanismo de atenção é aplicado de forma paralela a todas as palavras da sequência, permitindo maior eficiência computacional em comparação com RNNs e LSTMs.

IV. A arquitetura Transformer não utiliza a etapa de *embeddings*, dispensando representações vetoriais para palavras.

São corretas as afirmações:

a) apenas I e II.

b) apenas I e III.

c) apenas II e III.

d) apenas II e IV.

e) apenas III e IV.

Questão 12 **(Word embeddings contextuais e LLMs) (0,5 ponto)**

Sobre word embeddings contextuais e Modelos de Linguagem de Grande Escala (LLMs), como BERT e GPT, qual das afirmações abaixo está correta?

a) O pré-treino do modelo BERT consiste em prever a próxima palavra com base apenas nas palavras anteriores, de forma unidirecional.

b) Modelos BERT utilizam arquitetura Transformer muito próxima da original proposta, sendo muito bom para tarefas de geração de texto.

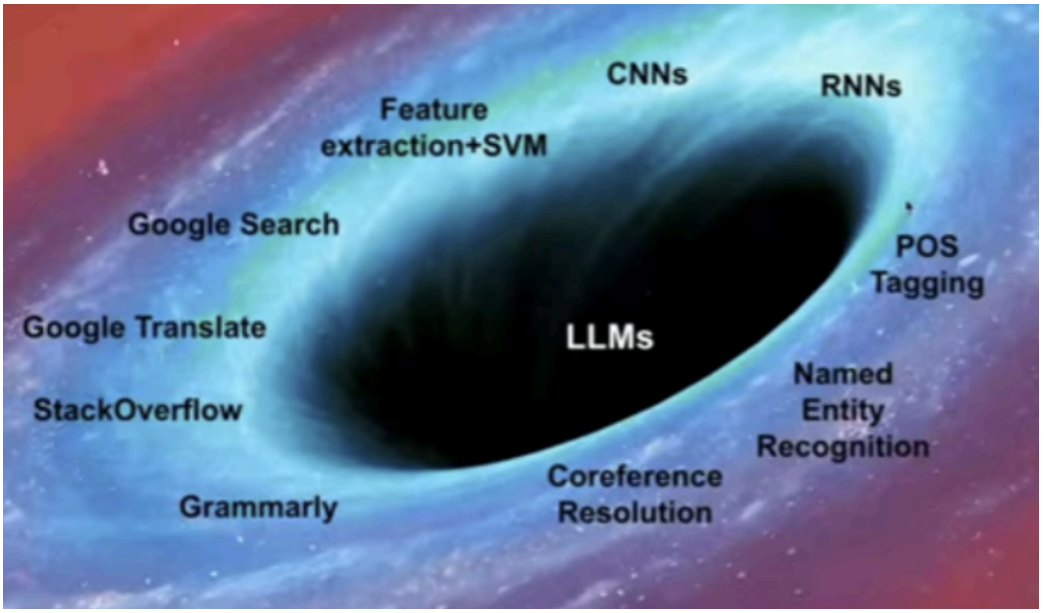
c) Para que sejam geradas word embeddings contextuais é obrigatório o uso de parte da arquitetura Transformer.

d) O GPT utiliza o mecanismo de atenção para processar entradas de forma bidirecional, analisando simultaneamente o contexto anterior e posterior de uma palavra.

e) Estudos mostram que, para determinadas tarefas específicas, modelos BERT ainda produzem resultados melhores que modelos da família GPT.

Questão 13 (Word embeddings contextuais e LLMs) (0,5 ponto)

Considere a seguinte figura abaixo, indicando o "estado atual" de LLMs, apresentada e defendida pelo pesquisador Rodrigo Nogueira, um dos sócios fundadores da Maritaca AI:



A alternativa abaixo que melhor descreve a principal mensagem da figura acima é:

a) LLMs vêm melhorando e substituindo aplicações e tarefas específicas que antes continham resultados incomparáveis e inatingíveis.

b) Softwares como o Grammarly e Google Translate ainda continuam extremamente relevantes como eram há anos atrás, garantindo que eles jamais serão substituídos por LLMs.

c) LLMs dependem de abordagens tradicionais, como Feature Extraction + SVM, para melhorar a eficiência e a precisão em aplicações modernas.

d) LLMs não conseguem resolver nem mesmo problemas simples, como tradução automática e correção gramatical.

e) LLMs atingirão brevemente a Inteligência Artificial Geral, termo usado para descrever um tipo de IA que tem a capacidade de realizar qualquer tarefa intelectual que um ser humano possa fazer.

Fim da parte objetiva.

Boa prova!