Prova 1 - Respostas

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

a) A Lei de Zipf afirma que a frequência de uma palavra é inversamente proporcional à sua posição no ranking de frequência. Ou seja, a 1ª palavra mais frequente aparece aproximadamente o dobro da 2ª, o triplo da 3ª, e assim por diante. Neste texto:

A palavra mais frequente é "o", com 4 ocorrências.
A segunda é "rato", com 3.
A terceira é "gato" e "correu", com 2.
As demais aparecem 1 vez.

Embora o texto seja muito curto, já podemos perceber um padrão decrescente de frequência que segue aproximadamente a Lei de Zipf.

b) Textos após etapas de pré-processamento:
(i) Case folding (transformar tudo para minúsculas):
"o gato viu o rato. o rato correu. o gato correu atrás do rato."
(ii) Remoção de acentos:
"o gato viu o rato. o rato correu. o gato correu atras do rato."
(iii) Remoção de pontuação e caracteres especiais:
"o gato viu o rato o rato correu o gato correu atras do rato"
(iv) Remoção de stop words (como “o”, “do”):
"gato viu rato rato correu gato correu atras rato"
(v) Stemming (redução das palavras à raiz):
gato → gat
viu → vi
rato → rat
correu → corr
atrás → atras
Texto final:
"gat vi rat rat corr gat corr atras rat"

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

[1, 1, 0, 0], # d1
[0, 1, 1, 0], # d2
[1, 0, 0, 1] # d3

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

a) Porque o dataset possui textos (letras de música) associados a rótulos ("boa música" ou "má música"), o que é essencial para treinar um modelo supervisionado.

b) O desbalanceamento (9.900 vs. 100) pode levar o modelo a ignorar a classe minoritária, tornando a acurácia enganosa; métricas como F1-Score ou Recall são melhores alternativas.

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

a) Word embeddings fixas capturam relações semânticas e sintáticas entre palavras em vetores densos, enquanto Bag of Words (BoW) e TF-IDF representam palavras como vetores esparsos baseados apenas em frequência, ignorando contexto e significado.

b) O treinamento do Word2Vec é auto-supervisionado porque aprende embeddings a partir de padrões internos do texto (como palavras vizinhas em janelas de contexto), sem necessidade de rótulos externos.

c) Embeddings fixas falham em palavras polissêmicas (ex: "banco" como instituição ou assento), pois atribuem um único vetor independente do contexto. BERT (com embeddings contextuais) é superior nesses casos, adaptando a representação conforme o uso da palavra na frase.

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

a) Transformers superam RNNs/LSTMs porque evitam o problema de dependência de longa distância (vanishing gradients) através do mecanismo de self-attention, que captura relações entre todas as palavras da sequência em paralelo, permitindo processamento mais eficiente e contextualizado.

b) A principal estratégia é o fine-tuning: ajustar os parâmetros do modelo pré-treinado (ex: BERT) adicionando uma camada de saída específica para a tarefa (ex: classificação) e treinando-o com dados do domínio-alvo, aproveitando o conhecimento linguístico já aprendido.

c) Para tarefas seq2seq (ex: tradução), a abordagem encoder-decoder (ex: original do Transformer, T5) é a mais adequada, pois o encoder processa a entrada contextualmente e o decoder gera a saída passo a passo, mantendo a coerência entre as linguagens. Modelos como BART e T5 seguem essa arquitetura com sucesso.

Questões	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
GABARITO	C	B	B	A	D	B	E	A