

### **Lista de Exercícios - Módulo 1**

- 1. (Processamento de Linguagem Natural)** O que é processamento de linguagem natural?
- 2. (Subdivisões de PLN)** A área de PLN geralmente é subdividida em duas partes: compreensão e geração. Explique brevemente cada parte e apresente dois exemplos de aplicações para cada área.
- 3. (Teste de Turing)** Por que não podemos afirmar que “passamos no teste de Turing” apesar de conseguirmos replicar seu experimento com sucesso?
- 4. (Histórico)** Explique brevemente cada uma das “ondas” históricas de PLN apresentadas abaixo:
  - Modelos baseados em regras (1950-1984)
  - Modelos estatísticos (1985 - 2012)
  - Word embeddings fixas + Deep learning (2013-2017)
  - Word embeddings contextuais + Transformer - NLU (2018 - 2022)
  - Word embeddings contextuais + Transformer - NLG (2023 - ?)
- 5. (Modelos baseados em regras)** O que são modelos baseados em regras? Apresente duas desvantagens desses modelos.
- 6. (Fatores de crescimento de PLN)** Cite e explique brevemente dois fatores que impulsionaram o crescimento de PLN nos últimos anos.
- 7. (Dado e informação)** Qual a diferença entre dado e informação?
- 8. (Tipos de dados)** Explique brevemente os três tipos de dados: (i) estruturados, (ii) semi-estruturados e (iii) não-estruturados. Qual dos tipos de dados é o mais “abundante”? Qual é o mais difícil de processar?
- 9. (Níveis de conhecimento da linguagem)** Dentre os principais níveis de análise linguística, pode-se destacar a sintaxe e a semântica. Qual dos níveis apresenta um maior desafio para modelos de linguagem? Por quê?
- 10. (Propriedades estatísticas dos textos)** O que diz a Lei de Zipf? Qual sua relação com stopwords?
- 11. (Propriedades estatísticas dos textos)** O que diz a Lei de Heaps? É possível afirmar que, em geral, a taxa de novos termos em um texto se mantém constante?
- 12. (Modelos de linguagem probabilísticos)** Em que se baseiam os modelos probabilísticos para calcular a previsão da próxima palavra?

**13. (N-gramas)** Por que modelos N-gramas podem ser vistos como um caso particular da definição geral de modelos probabilísticos? Qual a diferença para o caso geral?

**14. (N-gramas)** Qual o problema da utilização de n-gramas (considerando valores pequenos de n) para calcular dependências entre palavras neste exemplo? “O bolo que Maria fez enquanto trabalhava fora durante o período da tarde estava delicioso”.

**15. (Expressões regulares)** Explique o conceito de expressões regulares e sua importância no pré-processamento de texto?

**16. (Pré-processamento de texto)** É correto afirmar que todo texto deve passar pelas técnicas de pré-processamento estudadas em aula antes de serem utilizados para um propósito específico?

**17. (Pré-processamento de texto)** Qual estratégia é comumente utilizada para segmentação de sentenças para línguas ocidentais como o inglês e português? Por que esta estratégia nem sempre funciona para outras línguas?

**18. (Pré-processamento de texto)** Cite um exemplo de tarefa onde a remoção de *stopwords* pode acabar “atrapalhando” o desempenho do modelo no seu processamento.

**19. (Pré-processamento de texto)** Quais problemas podem ocorrer ao utilizar um stemming para as palavras “bebê” e “bebendo”?

**20. (Distância mínima de edição)** Qual a distância de Levenshtein considerando custos (1,1,2) para as operações (inserção,deleção,substituição) entre as palavras “gato” e “ratos”?

**21. (Representação textual)** Qual problema em representar caracteres por números da tabela ASCII no contexto de modelos de linguagem?

**22. (Modelos tradicionais de representação textual)** Explique brevemente o funcionamento do modelo Bag of Words (BoW). Qual problema é mitigado posteriormente por TF-IDF?

**23. (Modelos tradicionais de representação textual)** Explique brevemente os termos TF e IDF do modelo TF-IDF. Qual termo está relacionado à raridade do termo?

**24. (Modelos tradicionais de representação textual)** Considere o seguinte conjunto de três documentos:

d1: “gato e cachorro”

d2: “gato gosta de leite”

d3: “cachorro gosta de osso”

Considere as seguintes questões:

a) Apresente o vetor representando o vocabulário (monte o vetor a partir da ordem em que as palavras aparecem e não remova palavras).

b) Apresente a matriz BoW, onde linhas são documentos e colunas são palavras.

c) Apresente a matriz TF-IDF, onde linhas são documentos e colunas são palavras. A resposta final pode ficar indicada em log.

**25. (Similaridade de textos)** Como métodos baseados em distância de edição, como Levenshtein, diferem de medidas de similaridade vetorial, como distância de cosseno? Em quais casos cada um seria mais apropriado?

**26. (Similaridade de textos)** Por que mesmo utilizando a técnica de distância de cosseno para capturar similaridade dos textos, os resultados com BoW e TF-IDF não tendem a ser tão bons quanto embeddings fixas?

**27. (Aprendizado supervisionado para textos)** Por que a construção de datasets para modelos de aprendizado de máquina supervisionado geralmente são tão custosos de serem elaborados?

**28. (Aprendizado supervisionado para textos)** Caso se deseje aplicar um modelo tradicional de representação textual, como o BoW, sobre um dataset de letras de músicas, quais dos passos abaixo poderiam ser feitos sobre o texto antes da codificação e da aplicação de um algoritmo de aprendizado de máquina? Justifique.

- (i) Tokenização
- (ii) Remoção de pontuação
- (iii) Case folding
- (iv) Remoção de stopwords

**29. (Aprendizado supervisionado para textos)** Por que é dito que as probabilidades calculadas no algoritmo Naive Bayes não são “exatas” no sentido estatístico sobre todo o texto? Por que essa “falta de exatidão” não é um problema para tarefas de aprendizado de máquina?

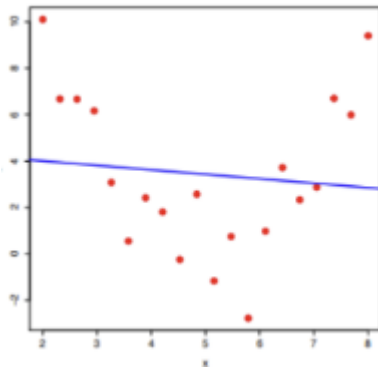
**30. (Avaliação de modelos supervisionados)** O que é generalização em modelos de aprendizado de máquina? Como “garantir” que o modelo proposto está generalizando de forma correta?

**31. (Avaliação de modelos supervisionados)** Como geralmente é feita a divisão dos dados com holdout assumindo conjuntos de treino e teste? O que é divisão estratificada e porque devemos priorizá-la sempre que possível?

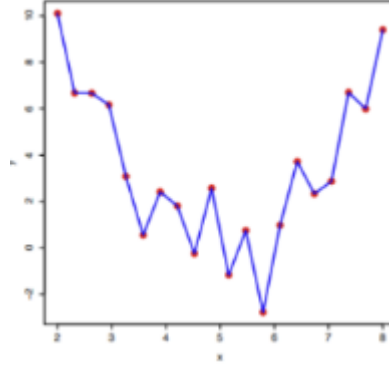
**32. (Avaliação de modelos supervisionados)** Por que a abordagem k-fold cross validation é “sempre que possível” indicada em relação à abordagem com holdout?

**33. (Avaliação de modelos supervisionados)** Qual a importância da utilização do conjunto de validação (development test set) na divisão dos dados?

**34. (Avaliação de modelos supervisionados)** Considere as imagens (A) e (B) abaixo no contexto de generalização de modelos, onde os pontos são os dados e as linhas são as funções encontradas descrevendo o comportamento destes dados.



(A)



(B)

- Quais problemas elas representam?
- O que ocorreu com os modelos para que estes problemas tenham surgido?
- Como seria a curva possível de “um bom modelo” que descreva estes pontos?

**35. (Avaliação de modelos supervisionados)** Por que, em geral, é mais utilizada a técnica conhecida como F1-Score para avaliação de modelos do que acurácia?

**36. (Word embeddings fixas)** Quais problemas as abordagens de representação de texto tradicionais (BoW e TF-IDF) apresentam?

**37. (Word embeddings fixas)** O que diz a hipótese distribucional? Apresente um exemplo que ilustre este processo.

**38. (Word embeddings fixas)** Considerando a afirmação “O significado de uma palavra é o seu uso na linguagem”, dada por Wittgenstein em 1953, é possível definir o significado de uma palavra sem seu contexto?

**39. (Word embeddings fixas)** Explique o conceito de word embeddings fixas e como elas diferem de representações tradicionais de texto, como Bag of Words e TF-IDF. Por que essas embeddings são consideradas mais avançadas?

**40. (Word embeddings fixas)** Quais são as principais limitações das word embeddings fixas na representação de palavras em diferentes contextos? Como essas limitações podem impactar aplicações em PLN?

**41. (Word embeddings fixas)** Explique como o treinamento de embeddings como Word2Vec utiliza grandes corpora de texto. Qual é o impacto do tamanho e da diversidade do corpus na qualidade das embeddings geradas?

**42. (Word embeddings fixas)** Qual a ideia geral do algoritmo Word2Vec (skip-gram)? Quais valores da rede neural são usados como embeddings finais para cada palavra?

**43. (Word embeddings fixas)** Por que embeddings fixas possuem muitas vezes vieses e estereótipos “maldosos”?

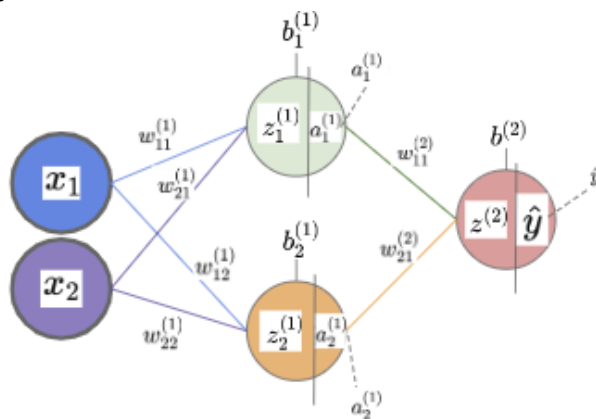
**44. (Word embeddings fixas)** Word embeddings fixas são “estáticas” no sentido de que cada palavra possui apenas uma representação vetorial, independentemente do contexto. Explique por que isso é uma limitação e como embeddings contextuais superam esse problema.

**45. (Redes neurais para textos)** Explique brevemente o que são modelos de redes neurais e como estes modelos em suas versões “base” podem ser utilizados com textos.

**46. (Redes neurais para textos)** Defina brevemente os seguintes conceitos sobre redes neurais:

- Perceptrons, pesos e bias
- Camadas de entrada, ocultas e de saída
- Função de ativação (activation function)
- Função de perda (loss function)
- Gradiente descendente
- Backpropagation
- Deep learning

**47. (Redes neurais para textos)** Considere uma rede neural simples com as seguintes informações e estrutura:



**Entradas ( $x_1, x_2$ ):** Representam os valores fornecidos à rede como dados de entrada.

**Pesos ( $w_{ij}^{(1)}, w_{ij}^{(2)}$ ):**

- $w_{ij}^{(1)}$ : Pesos conectando as entradas ( $x_1, x_2$ ) aos neurônios da camada oculta.
- $w_{ij}^{(2)}$ : Pesos conectando os neurônios da camada oculta ( $a_1^{(1)}, a_2^{(1)}$ ) ao neurônio de saída.

**Bias ( $b_1^{(1)}, b_2^{(1)}, b^{(2)}$ ):**

- $b_1^{(1)}, b_2^{(1)}$ : Bias adicionados à entrada dos neurônios da camada oculta.
- $b^{(2)}$ : Bias adicionado à entrada do neurônio da camada de saída.

**Camada Oculta ( $z_1^{(1)}, z_2^{(1)}$ ):** Representam os valores intermediários calculados para os neurônios da camada oculta antes de aplicar a função de ativação.

**Ativações Ocultas ( $a_1^{(1)}, a_2^{(1)}$ ):** Saídas dos neurônios da camada oculta após a aplicação da função de ativação (ReLU, neste caso).

**Camada de Saída ( $z^{(2)}$ ):** Valor calculado no neurônio da camada de saída antes da aplicação da função de ativação.

**Saída ( $\hat{y}$ ):** Predição final da rede neural após a função de ativação ser aplicada no neurônio de saída.

**1. Entradas:**

$$x_1 = 2.0 \text{ e } x_2 = -1.0$$

**2. Pesos iniciais (da camada de entrada para a camada oculta):**

- $w_{11}^{(1)} = 0.5, w_{12}^{(1)} = -0.3$
- $w_{21}^{(1)} = 0.8, w_{22}^{(1)} = -0.7$

**3. Pesos iniciais (da camada oculta para a camada de saída):**

- $w_{11}^{(2)} = 0.2, w_{21}^{(2)} = -0.5$

**4. Biases:**

- $b_1^{(1)} = 0.1, b_2^{(1)} = 0.2$  (para os neurônios da camada oculta)
- $b^{(2)} = -0.1$  (para o neurônio de saída)

**5. Função de ativação:** ReLU ( $f(x) = \max(0, x)$ )

**6. Função de perda:** MAE (Erro Absoluto Médio):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_{\text{target}} - y_{\text{predito}}|$$

O alvo da rede é  $y_{\text{target}} = 1.5$ .

Assumindo os valores iniciais fornecidos acima, calcule os valores intermediários dos neurônios da camada oculta ( $a_1(1)$  e  $a_2(1)$ ), da camada final ( $\hat{y}$ ), e o erro gerado durante a primeira passada na rede feedforward.

**48. (Redes neurais para textos)** Defina brevemente os seguintes conceitos sobre redes neurais textuais:

- a) RNN
- b) LSTM
- c) Bi-LSTM
- d) Encoder-decoder
- e) Attention e self-attention
- f) Transformers

**49. (Redes neurais para textos)** Discorra brevemente sobre os seguintes problemas de RNN e como eles podem ser resolvidos ou mitigados:

- a) Desaparecimento e explosão de gradientes
- b) Dificuldade em aprender dependências de longo prazo
- c) Treinamento lento (sequencial)

**50. (Redes neurais para textos)** Apresente modelos de linguagens conhecidos que ilustrem o uso de arquiteturas Transformer e suas respectivas principais tarefas:

- a) encoder-decoder
- b) encoder-only
- c) decoder-only

**51. (Redes neurais para textos)** Qual a ideia geral do mecanismo de atenção, com ênfase em self-attention?

**52. (Redes neurais para textos)** Por que a arquitetura Transformers possibilitou que uma quantidade enorme de dados fosse treinada em relação à arquiteturas prévias, como RNNs ou LSTMs?

**53. (Word embeddings contextuais)** Por que word embeddings contextuais conseguem lidar bem com polissemia (palavra com vários significados)? Qual a diferença principal em relação a word embeddings fixas? Apresente um breve exemplo.

**54. (Large Language Models)** Não há um “consenso” na comunidade com relação à classificação de modelos como LLMs. Entretanto, assumindo a definição vista em aula de que LLMs são “redes neurais profundas treinadas em quantidades massivas de dados que implementam embeddings contextuais”, podemos considerar que todas LLMs são baseadas em transformers?

**55. (Word embeddings contextuais)** O processo geral de treinamento de LLMs pode ser dividido em quatro fases: (1) coleta de dados, (2) pré-processamento (tokenização), (3) pré-treino e (4) *fine-tuning*. Explique brevemente cada uma destas fases tendo enfoque nos seguintes modelos estudados:

- (a) BERT
- (b) GPT-3

**56. (Word embeddings contextuais)** Qual papel a arquitetura Transformer exerce sobre o tipo de aplicação que os modelos exercem? Assumindo unicamente esta característica, quais camadas Transformer (encoder-decoder, encoder, decoder) tendem a apresentar melhores resultados para as tarefas: (1) tradução automática, (2) classificação, (3) análise de sentimentos, (4) geração de texto, (5) sumarização.

**57. (Large Language Models)** Por que LLMs como o BERT e GPT podem ser treinados de forma “auto-supervisionada”?

**58. (Tokenização)** Qual o principal objetivo da utilização de estratégias de tokenização específicas para LLMs, como a Byte Pair Encoding (BPE) e WordPiece, por exemplo, em relação à técnicas tradicionais, como segmentação de palavras por espaçamentos?

**59. (BERT)** O pré-treino do modelo BERT é feito com base em duas tarefas principais: (i) Masked Language Modeling (MLM) e (ii) Next Sentence Prediction (NSP). Explique brevemente a ideia geral de cada uma delas.

**60. (BERT)** Qual a saída do modelo BERT após o pré-treino? Como esta saída pode ser adaptada para realizar tarefas específicas, como classificação, por exemplo?

**61. (Transfer learning)** Com relação à técnicas de transfer learning, explique brevemente os conceitos:

- a) Pré-treino
- b) Continuação de pré-treino

c) Fine-tuning

**62. (Geração de texto)** Cite e explique brevemente dois motivos que tornam o modelo BERT “pouco adequado” para tarefas de geração de texto e fins.

**63. (Geração de texto)** Qual a ideia geral de modelos de geração de texto autorregressivos, usado pela família de modelos GPT? No que atuam as estratégias de decodificação?

**64. (GPT)** Qual a intuição geral para o modelo GPT-3 ser unidirecional, em vez de bidirecional, como o modelo BERT? Qual a ideia geral do mecanismo “Masked Self Attention” nesta parte (ou seja, quais partes foram mascaradas)?

**65. (Aprendizado com poucos dados)** Por que é interessante utilizar abordagens alternativas ao “*fine-tuning*” para LLMs generativas?

**66. (Aprendizado com poucos dados)** No que consiste as abordagens zero-shot e few-shot learning?

**67. (Aprendizado com poucos dados)** Fale brevemente sobre os seguintes problemas de LLMs generativas:

- a) “Papagaios estocásticos”
- b) Halucinações
- c) Raciocínio

Bom Trabalho!