

---

**UNIVERSIDADE FEDERAL DE ALAGOAS**  
**INSTITUTO DE COMPUTAÇÃO**

**Processamento de Linguagem Natural**  
**Professor: Thales Vieira**

---

**4a lista de exercícios**

**23 de setembro de 2025**

---

**Instruções:**

A lista deve ser respondida por grupos de até 2 pessoas (graduação).

Resoluções idênticas de grupos distintos serão desconsideradas.

O código e demais dados devem ser anexados a cada questão.

Data limite para entrega: 12/10/2025.

Usando sua base de textos após os pré-processamento realizados na lista 1, realize as seguintes tarefas:

**1.** Estude o notebook “A Visual Notebook to Using BERT for the First Time.ipynb” e o pdf “Text classification with Transformer”, ambos enviados em anexo a esta lista.

- a) Resolva o mesmo problema de classificação da primeira questão da segunda lista, usando uma combinação de DistilBERT com os três classificadores usados na segunda lista.
- b) Faça o mesmo adaptando a arquitetura Transformer do arquivo pdf para o seu problema.
- c) Compare todos os resultados.

**2.** A técnica de engenharia de prompt surgiu como uma possibilidade de treinar um grande modelo de linguagem com poucos exemplos (*few-shot learning*) sem precisar realizar *fine-tuning* nos pesos do modelo.

- a) Utilizando o *Ollama*, carregue o modelo *llama3.1:8b* e crie código python para extrair algum tipo específico de entidade nomeada usando engenharia de prompt. Você pode usar o artigo a seguir como referência para criar prompts: <https://arxiv.org/abs/2304.10428>.
- b) Rode o código em sua base de textos e gere um histograma com a frequência de ocorrência das entidades detectadas.

**3.** Os modelos GPT são treinados para gerar texto. Leia o tutorial em anexo para aprender a realizar fine-tuning com o GPT-2 usando o framework da Huggingface. Você deve conseguir rodar o modelo small em uma conta do Google Colab padrão. Realize o fine-tuning com sua base de textos, e depois gere 5 textos completos. Você pode começar de alguma tag, como descrito no tutorial, ou fornecer algumas palavras de entrada.