UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO

Processamento de Linguagem Natural
Professor: Thales Vieira

4a lista de exercícios **22 de outubro de 2024**

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: 11/11/2024.

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", anexo a essa 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) Compare todos os resultados.

2. Estude e pesquise sobre o BERTopic, uma adaptação do BERT para modelagem de tópicos. O código fonte está disponível em https://github.com/MaartenGr/BERTopic.

a) Extraia os tópicos de sua base, exibindo as informações dos tópicos (palavras mais relevantes).
b) Exiba visualizações com gráficos de barra e usando visualize_topics().

3. 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. Experimente criar um prompt no ChatGPT ou outro grande modelo de linguagem para extrair algum tipo específico de entidade nomeada. Exiba seu prompt e pelo menos 5 exemplos de texto com as respectivas entidades identificadas pelo modelo. Você pode usar o artigo a seguir como referência: `https://arxiv.org/abs/2304.10428`.

4. 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.