FIAP

Data Science & Artificial Intelligence

NLP

in /thnogueira

PROJETO INTEGRADO

NATURAL LANGUAGE PROCESSING

CONTEXTO E DESAFIOS

A QuantumFinance tem um canal de atendimento via chat e precisar classificar os assuntos dos atendimentos para melhorar as tratativas dos chamados dos clientes. O canal recebe textos abertos dos clientes relatando o problema e/ou dúvida e depois é direcionado para algum uma área especialista no assunto para uma melhor tratativa.

O desafio de sua equipe é

Parte 1 - Criar um modelo classificador de assuntos aplicando técnicas tradicionais de NLP, que consiga classificar através de um texto o assunto conforme disponível na base de dados [1] para treinamento e validação do modelo seu modelo.

Parte 2 – Realizar a tarefa de classificação apresentada no ítem anterior com a utilização IA Generativa, , utilizando as técnicas de prompt engineering discutidas em sala.

Parte 3 (Extra) – Utilizar a IA Generativa para fazer uma classificação livre de assuntos e avaliar qualitativamente os resultados.

Observação: Nas partes 2 e 3 do trabalho utilizar a base de dados estratificada disponibilizada em [2]. Esta amostra possui 200 registros de cada categoria. Este ajuste deve-se à limitação do número de tokens disponibilizados no modo gratuito da API da openAI. Para fazer o desenvolvimento de prompt, busque rodar com poucos exemplos e ir incrementando aos poucos, a medida em que vá melhorando.

[1] - https://raw.githubusercontent.com/thiagonogueira/datasets/main/tickets_reclamacoes_classificados_one_line.csv
[2] - https://raw.githubusercontent.com/thiagonogueira/datasets/main/tickets_reclamacoes_classificados_one_line_generative.csv

ENTREGÁVEIS

Jupyter Notebook:

O notebook deverá estar bem organizado com seções claras e textos que facilitem a compreensão da análise e decisões tomadas e que permita a obtenção do resultado final a partir do dataset disponibilizado. Utilize o notebook de template para a entrega.

O modelo precisar atingir um score na métrica F1 Score superior a 75%. Utilize o dataset [1] para treinar e testar o modelo, separe o dataset em duas amostras (75% para treinamento e 25% para teste com o randon_state igual a 42).

Fique à vontade para testar e explorar as técnicas de pré-processamento, abordagens de NLP, algoritmos e bibliotecas, mas explique e justifique as suas decisões durante o desenvolvimento.

Importante: parte significativa da avaliação do resultado será feita com a execução do notebook. Desta forma, é importante que todas as células executem corretamente e que os resultados sejam integralmente reprodutíveis;

NOTAS E GRUPOS

Notas:

Parte 1 (70%), sendo:

- 35% - Demonstrações das aplicações das técnicas de PLN (regras, pré-processamentos, tratamentos, variedade de modelos aplicados, organização do pipeline, etc.)
- 35% - Baseado na performance (score) obtida com a amostra de teste no pipeline do modelo campeão (validar com a Métrica F1 Score). Separar o pipeline completo do modelo campeão conforme template.

Parte 2 (30%)

Parte 3 (Extra): Poderá acrescentar em até 2 pontos a nota do trabalho final.

Grupos:

O trabalho deverá ser feito, **necessariamente**, em grupo de 3 até 4 pessoas

OBRIGADO!