Product Data Science - NLP Test

Teste de experiência na posição

Descrição do problema

Você é um cientista de dados que trabalha em uma empresa de e-commerce. Essa empresa oferece produtos em uma plataforma que possui um sistema em que os usuários fazem comentários sobre os produtos oferecidos.

Após ter coletado um volume considerável de comentários, a empresa gostaria de desenvolver um sistema capaz de inferir qual o sentimento dos comentários dos usuários e também obter *insights* sobre quais são as características de produtos e perfis de usuários que geralmente tendem a ser mais positivos ou negativos.

A base com os comentários e pontuações atribuidas aos produtos pelos usuários da plataforma está disponível para *download* em: https://github.com/americanas-tech/b2w-reviews01/blob/main/B2W-Reviews01.csv

Detalhes da base de dados

A base possui os seguintes metadados:

#	Field	Data type	Description
1	submission_date	date/time	review submission date (format YYYY-MM-DD hh:mm:ss)
2	reviewer_id	string	unique reviewer id
3	product_id	integer	unique product id
4	product_name	string	product name
5	product_brand	string	product brand
6	site_category_lv1	string	product category - first level
7	site_category_lv2	string	product category - second level
8	overall_rating	integer	overall customer rating, from 1 to 5
9	recommend_to_a_friend	string	answer to "would you recommend this product to a friend?" ("Yes"/"No")
10	review_title	text	review title, introduces or summarizes the review content
11	review_text	text	main text content of the review
12	reviewer_birth_year	integer	reviewer's birth year
13	reviewer_gender	string	reviewer's gender ("F" for female; "M" for male)
14	reviewer_location	string	reviewer's Brazilian State, according to the delivery address

Fonte: https://github.com/americanas-tech/b2w-reviews01/blob/main/b2wreviews01_stil2019.pdf

O desafio

Como cientista de dados, você deverá utilizar técnicas de NLP (*natural language processing*), de aprendizado de máquina e análise de dados para desenvolver (treinar) um algoritmo capaz de classificar o sentimento dos comentários feitos pelos usuários.

Obs: Caso você não possua os recursos computacionais necessários é permitido utilizar uma amostra da base de dados para o desenvolvimento do projeto.

Entregáveis

1. Uma apresentação com os principais resultados oriundos da sua análise e desenvolvimento do algoritmo.
2. O código usado, contendo uma breve descrição de como utilizar/rodar e uma justificativa da metodologia utilizada (podem ser feitos gráficos, tabelas ou qualquer análise que ajude a explicar melhor a solução).

Critérios de avaliação

Iremos avaliar o projeto da seguinte maneira, do maior para o menor em termos de relevância:

1. Construção do pipeline de modelagem.
2. Qualidade da documentação
3. Clareza e capacidade analítica nos relatórios.
4. Qualidade e arquitetura do código, por exemplo:
    - Modularização
    - Funções
    - Testes
    - Logging
5. Reprodutibilidade e instruções para o uso, por exemplo:
    - Docker
    - Conda
    - CLI
    - Virtualenv
6. Escalabilidade de código (processamento) não é um obrigatório, mas será um plus.

O projeto pode ser desenvolvido em inglês ou português e em um repositório GitHub/GitLab. O repositório pode ser aberto ou privado (se for privado, terá que adicionar um de nossos integrantes como Master). Não iremos avaliar você até o final do período de 9 dias (contando a partir da data combinada entre você e a equipe de recrutamento), mas gostaríamos de acompanhar o desenvolvimento do trabalho.

Github para adicionar:
https://github.com/Igorcortez
https://github.com/jhosoume
https://github.com/vhdeluca

Por último, mas não menos importante:

- Não existe apenas uma resposta certa: estamos procurando um profissional criativo capaz de fornecer uma solução eficiente para este problema.

- A linguagem de programação mais usada em nossa equipe é Python, mas fique à vontade para usar outra de sua preferência.