dhauz

Desafio para área de Data Science

Parabéns pela aprovação na primeira etapa do processo seletivo para a área de Data Science da DHAUZ!

Seguindo com o nosso processo, nesta etapa você deverá realizar um desafio técnico, detalhado abaixo. O objetivo é avaliar a sua capacidade de exploração de dados, criatividade na elaboração de hipóteses, features e metodologia na resolução de problemas.

Você pode utilizar a linguagem, ferramentas e frameworks que se sentir mais confortável para elaborar a solução. Recomendamos o desenvolvimento do racional e exploração em uma ferramenta como por exemplo o Jupyter Notebook para facilitar a inclusão de comentários e gráficos explicativos.

Sinta-se livre para utilizar qualquer outra funcionalidade não listada acima que demonstre suas habilidades!

O projeto deve ser feito em um repositório no Github e o seu link enviado no final do desafio.

Análise de sentimentos em viagens aéreas

Você foi contratado pela DHAUZ como cientista de dados para participar em um projeto para um site de venda de passagens aéreas que deseja desenvolver um sistema de análise de sentimentos a partir dos reviews de passageiros. A ideia é que, a partir da classificação das informações colocadas pelos passageiros, as companhias aéreas possam entender quais os principais pontos a melhorar a experiência do cliente.

Informações do dataset:

- A base de dados das *reviews* foi coletada no site https://www.airlinequality.com/
- Link para baixar a base de dados:
    - https://dhauz-challenges.s3.amazonaws.com/Travel_Chalenge.zip

Comece respondendo as seguintes questões:

1. Faça uma etapa de processamento dos dados para verificar possíveis dados faltantes ou duplicados
2. Realize as etapas padrões de NLP nas colunas *Review* e *Review_title* (ex: Tokenização, remoção de stop-words, ...)
3. Exploração dos dados:
    a. Faça um gráfico para verificar a distribuição da feature *Overall_rating* pelas companhias aéreas. Faça um gráfico similar para verificar a distribuição dessa features pelos modelos de aeronaves (*Aircraft*)
    b. Utilize a visualização de nuvem de palavras para estudar quais palavras mais aparecem quando o *Overall_rating* é igual ou inferior a 3 e quando é igual ou superior a 8.
    c. Estude a correlação e, portanto, o possível impacto das colunas que contém notas separadas (*'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity'*) na nota final (*Overall_rating*)
4. Utilizando o critério abaixo para classificar o sentimento de cada review como **positivo, negativo** e **neutro**, faça dois modelos de classificação de sentimentos, sendo um deles utilizando os textos da *review* e *review_title* como inputs e o outro utilizando as notas das features separadas (*'Seat Comfort', 'Cabin Staff Service', 'Food & Beverages', 'Ground Service', 'Inflight Entertainment', 'Wifi & Connectivity'*) e compare os dois modelos.
    a. Nota final **menor** que 4: **Negativo**
    b. Nota final **entre** 4 e 7: **Neutro**
    c. Nota final **maior** que 7: **Positivo**
5. Com o modelo de classificação de sentimentos, faça uma análise sobre o impacto de **atrasos** de viagem no NPS de 3 companhias aéreas.
    a. NPS = %positivos-%negativos