Big Data Real-Time Analytics Com Python e Spark Versão 3.0

Projeto com Feedback 4

Análise de Sentimentos em Tweets Sobre o ChatGPT com PySpark

O ChatGPT é o sistema de conversação baseado em IA oferecido pela OpenAI. Foi lançado no final de 2022 e desde então tem recebido atenção do público em geral que finalmente percebeu como a IA evoluiu e já está entre nós.

O sentimento sobre o ChatGPT tem dividido as opiniões. Enquanto alguns reconhecem que o ChatGPT pode ser uma ferramenta útil para o dia a dia de diversas atividades, outros demonstram receio em ser substituídos pela IA.

Mas qual sentimento prevalece nas redes sociais? Responder a essa pergunta é o seu trabalho neste projeto. Através da análise de Tweets sobre o ChatGPT você deve construir um processo de análise que permita identificar o sentimento que predomina, especialmente no Twitter, sobre o ChatGPT. Esse projeto poderá ser estendido para refletir o sentimento sobre outros temas, por exemplo.

Para a construção desse projeto, recomendamos a utilização do PySpark. Não é necessário usar Machine Learning, embora seja uma possibilidade. O conhecimento para criar análise de sentimentos com o PySpark pode ser obtido no link abaixo (ou você pode pedir ajuda ao próprio ChatGPT):

https://spark.apache.org/docs/latest/ml-features

Para a fonte de dados você tem duas opções. Escolha a que considerar ideal para você.

Opção A: Dados Estáticos

Você pode trabalhar com dados estatísticos de tweets coletados sobre o ChatGPT a partir do link abaixo:

https://www.kaggle.com/datasets/tariqsays/chatgpt-twitter-dataset

Opção B: Extrair Dados em Tempo Real

Você pode extrair seus próprios dados em tempo real. Nesse caso você terá que criar uma API com o Twitter. O link abaixo tem a documentação e o procedimento:

https://developer.twitter.com/en/docs/twitter-api

Independente da fonte usada seu trabalho deve demonstrar qual sentimento prevalece sobre o ChatGPT: positivo, negativo ou neutro.

Quando concluir o projeto, envie o(s) script(s) que você criar para o seguinte e-mail: suporte@datascienceacademy.com.br.

Caso você tenha criado datasets auxiliares e esses sejam muito grandes, armazene em um diretório virtual (existem vários na internet, como Google Drive ou Dropbox) e envie o link para que nossa equipe possa baixar os datasets. Se os arquivos forem pequenos (uma amostra do dataset original), envie no anexo junto com o script. Documente seu script tanto quanto possível.

Caso prefira, disponibilize seu projeto no Github e envie o link do seu repositório para nossa equipe no e-mail suporte@datascienceacademy.com.br. Nesse caso, o Readme do repositório deve constar que este trata-se de um projeto da Formação Cientista de Dados da Data Science Academy, caso contrário não será avaliado.

Em até 72 horas, daremos o feedback respondendo seu e-mail. Caso não receba a resposta em até 72 horas entre em contato com a nossa equipe para verificar se recebemos seu projeto.

Bom trabalho!