Projeto de Coleta de Dados - Processo Seletivo para Estágio Proffer

Objetivo do Projeto

Desenvolver uma solução para coletar informações de preços de produtos do site Preço da Hora Bahia (https://precodahora.ba.gov.br/) para as cidades de Salvador e Feira de Santana.

Detalhes do Projeto

Dados de Entrada

Serão fornecidas duas listas em formato JSON:

- **Lista de EANs**: Contendo 15.000 códigos de barras de produtos
- **Lista de Descrições**: Contendo 367 descrições de produtos

O que é EAN?

EAN (European Article Number) é um código de barras padrão utilizado internacionalmente para identificar produtos. No Brasil, é comumente chamado de código de barras ou GTIN (Global Trade Item Number):

- Geralmente possui 13 dígitos (EAN-13)
- Cada produto comercializado possui um código EAN único
- É o número impresso abaixo do código de barras nas embalagens dos produtos
- Exemplo: 7896004713274

Parâmetros de Busca

- **Cidades**:
    - Salvador (código do município: 2927408)
    - Feira de Santana (código do município : 2910800)
- **Estado**: Bahia (UF: BA, código do estado: 29)

Sobre os códigos de município e estado

Os códigos de município e estado são baseados no sistema de codificação do IBGE (Instituto Brasileiro de Geografia e Estatística):

- **Código de Estado (UF)**: É um número de dois dígitos atribuído pelo IBGE a cada unidade federativa do Brasil. Para a Bahia, o código é 29.
- **Código de Município**: É um número de 7 dígitos atribuído pelo IBGE a cada município brasileiro. Os dois primeiros dígitos representam o código do estado ao qual o município pertence.

- Salvador: 2927408 (29 = Bahia, 27408 = código específico do município)
- Feira de Santana: 2910800 (29 = Bahia, 10800 = código específico do município)

Para mais informações, você pode consultar: https://www.ibge.gov.br/explica/codigos-dos-municipios.php

Dados a Serem Coletados

Obrigatórios:

- EAN do produto
- Descrição do produto
- Preço
- Data da coleta
- Código do município
- Código do estado

Bônus (opcionais, mas serão considerados diferenciais):

- Rede de onde foi coletada (estabelecimento)
- Bairro
- Cidade
- UF
- CNPJ da rede coletada (estabelecimento)

Requisitos Técnicos

1. Você tem liberdade para escolher as tecnologias e frameworks que preferir.
2. A linguagem deve ser Python.
3. O resultado final deve ser um dataset estruturado contendo todos os dados obrigatórios e, se possível, os dados bônus.
4. Sua solução deve ser capaz de lidar com possíveis erros como:
    - Produtos não encontrados
    - Falhas na conexão
    - Limites de requisição do site

Entrega

Prazo

A entrega deverá ser feita até segunda-feira, às 14:00 do dia 10/03/2025.

O que entregar

Você deve entregar:

1. O código-fonte completo da sua solução

2. Um arquivo README contendo:
    - Instruções detalhadas para executar o código
    - Descrição da arquitetura e tecnologias utilizadas
    - Explicação sobre como os dados são coletados e processados
    - Desafios encontrados e como foram solucionados
    - Possíveis melhorias que poderiam ser implementadas
3. O dataset final com os dados coletados

Critérios de Avaliação

Sua solução será avaliada considerando:

1. **Funcionalidade**: A solução atende aos requisitos e coleta corretamente os dados.
2. **Código**: Organização, legibilidade e boas práticas de programação.
3. **Performance**: Eficiência na coleta dos dados, considerando o volume solicitado.
4. **Tratamento de Erros**: Como a solução lida com exceções e casos de erro.
5. **Documentação**: Clareza e completude da documentação.

Observações Importantes

- Este projeto simula um desafio real que você pode encontrar trabalhando em nossa equipe de coleta de dados.
- Valorizamos soluções criativas e eficientes, mesmo que não estejam 100% completas.
- A vaga é para estágio, então não esperamos conhecimento avançado em todas as tecnologias mencionadas, mas buscamos candidatos com potencial de aprendizado e raciocínio lógico.

Dicas Úteis

- Analise o site antes de começar a codificar para entender como os dados estão estruturados.
- Verifique se o site possui uma API pública que possa ser utilizada.
- Considere a possibilidade de paralelizar a coleta para otimizar o tempo.
- Implemente mecanismos para lidar com possíveis limitações de requisições (rate limiting).
- Organize bem seus dados desde o início para facilitar o processamento posterior.

Boa sorte! Estamos ansiosos para ver sua solução.