Documentação do Web Crawler ProxyList

Sumário

1. Objetivo

O objetivo deste Web Crawler é acessar uma página pública de proxies, extrair informações sobre proxies disponíveis e salvar esses dados em arquivos locais e em um banco de dados. O sistema também oferece a capacidade de salvar a página HTML de cada execução.

2. Requisitos

Tecnologias Utilizadas

- **.NET Framework / .NET Core**: Framework para desenvolvimento da aplicação.
- **Banco de Dados (MariaDB, SqlLite, etc.)**: Para armazenar informações da execução e dados extraídos.

3. Funcionalidades

3.1. Acessar a Página de Proxies

O Web Crawler é configurado para acessar o site **"https://proxyservers.pro/proxy/list/order/updated/order_dir/desc"**. Este site oferece uma lista de proxies ordenada pela data de atualização. O crawler é responsável por fazer a requisição HTTP para esse endereço e carregar as páginas que contêm informações sobre os proxies.

3.2. Extrair Dados de Proxies

Durante a execução, o crawler coleta as seguintes informações de cada proxy disponível na página:

- **IP Address**: O endereço IP de cada proxy.
- **Port**: A porta associada ao proxy, geralmente usada para estabelecer a conexão.
- **Country**: O país de origem do proxy, que é importante para categorizar e classificar os proxies.

- **Protocol**: O tipo de protocolo utilizado pelo proxy, como HTTP, HTTPS, ou SOCKS, que determina o tipo de conexão possível através do proxy.

Esses dados são extraídos de todas as linhas presentes nas páginas acessadas, proporcionando uma visão completa dos proxies disponíveis.

3.3. Salvar Dados em Arquivo JSON

Após extrair os dados dos proxies, o Web Crawler salva as informações em um arquivo **JSON**. Esse arquivo contém todos os dados extraídos e pode ser utilizado para consulta posterior ou para alimentar outros sistemas. O arquivo JSON é armazenado localmente no sistema do usuário.

3.4. Salvar Informações da Execução no Banco de Dados

Além de salvar os dados extraídos, o crawler também registra informações sobre a execução no banco de dados. Os seguintes dados são salvos:

- **Data de Início e Término da Execução**: A data e hora de início e término da execução do crawler, permitindo o acompanhamento do tempo de processamento.
- **Quantidade de Páginas Processadas**: O número total de páginas visitadas durante a execução do crawler.
- **Quantidade de Linhas Extraídas**: O número total de proxies extraídos durante a execução.
- **Arquivo JSON Gerado**: O arquivo JSON gerado é salvo em banco, para fácil acesso e verificação.

3.5. Salvar Página HTML

Para cada página que é processada durante a execução, o conteúdo HTML completo é salvo em um arquivo local. Isso permite que, caso seja necessário, a página original seja analisada posteriormente para entender a estrutura da informação ou para verificar o estado da página na hora da coleta.

4.Conclusão

O Web Crawler ProxyList é uma solução eficiente para coletar proxies de uma página específica da web. Ele é capaz de acessar a página, extrair dados relevantes sobre os

proxies, salvar esses dados em arquivos JSON e HTML, e registrar informações sobre a execução em um banco de dados.