Análise e Modelagem dos dados do Datasus

25 DE ABRIL

Projeto Integrador III

Aluno: Vinícius de Paula

IESB
Centro Universitário

Coleta de Dados

O objetivo deste trabalho é criar uma base de dados com os dados das internações hospitalares no período de 2019 a 2023 na unidade temporal de mês, geograficamente agregados por Município.

Para isso, foi necessário o desenvolvimento de um sistema de webscraping com o objetivo de coletar sistematicamente e de forma automatizada todos os dados que envolvam a produção hospitalar, isso engloba o registro detalhado de cada atendimento, procedimento diagnóstico, cirúrgico ou terapêutico realizado nos hospitais e demais instituições de saúde conveniadas ao SUS.

O Sistema de Informações Hospitalares do SUS (SIH/SUS) é a ferramenta utilizada para gerenciar esses dados.

Desenvolvimento do Sistema

O objetivo desta etapa foi automatizar a coleta de dados do DATASUS e realizar o tratamento desses dados para posterior análise. Para isso, foi utilizada a linguagem Python por meio das bibliotecas Selenium em conjunto com o WebDriver do Google Chrome para acessar e extrair informações. A coleta foi feita em um período de 16 anos.

Bibliotecas Utilizadas

- pandas (importada como pd): Utilizada para manipulação e análise dos dados tabulares.
- selenium: Utilizada para automatizar a interação com o navegador web.
- webdriver e chrome.service: Parte do pacote selenium, utilizadas para configurar e controlar o navegador Google Chrome.

Antes de iniciar a coleta de dados, é configurada uma instância do navegador Google Chrome. São utilizadas opções para evitar a exibição visual do navegador (--headless), o que torna a execução mais rápida e menos intrusiva.

Coleta Sistemática dos Dados

A coleta de dados ocorreu da mesma maneira que ocorreria caso feita de forma manual, por uma pessoa. Os passos exatos executados pelo robô para coleta sistemática dos dados foram:

1. **Acesso ao Site do DATASUS**: O programa acessa o site do DATASUS por meio de uma URL específica.

2. **Seleção dos Parâmetros de Consulta**: São feitas seleções nos campos do site para configurar a consulta, incluindo a escolha do grupo de procedimentos e do período.

3. **Iteração sobre os Períodos**: O programa itera sobre os períodos desejados, coletando os dados para cada um deles. Para cada período, são realizadas as seguintes etapas:

    i. Seleção do período específico.
    ii. Configuração das opções de exibição dos dados.
    iii. Extração dos dados da nova aba aberta e armazenamento em um DataFrame do Pandas.
    iv. Fechamento da aba e retorno à aba principal.
    v. Deseleção do período utilizado.

4. **Concatenação dos Dados**: Após a coleta de dados para todos os períodos desejados, os DataFrames resultantes são concatenados em um único DataFrame.

5. **Exportação dos Dados**: O DataFrame consolidado é exportado para um arquivo CSV.

Tratamento dos dados

Após a coleta, os dados foram tratados da seguinte maneira:

1. **Remoção de Aspas**: É aplicada uma função para remover as aspas presentes nos valores das colunas do DataFrame.

2. **Manipulação das Colunas**:

    i. Adição da coluna 'Cod_Município', que contém o código do município.
    ii. Remoção dos primeiros seis caracteres da coluna 'Município', que correspondem ao código do município.
    iii. Divisão da coluna 'Data' em duas colunas separadas 'Mês' e 'Ano'.
    iv. Remoção de vírgulas na coluna 'Ano'.
    v. Eliminação da coluna 'Data'.

3. **Renomeação das Colunas**: As colunas são renomeadas para remover as aspas e garantir consistência nos nomes.

4. **Seleção das Colunas Relevantes**: Apenas as colunas de interesse são mantidas no DataFrame final.

5. **Exportação dos Dados Tratados**: O DataFrame tratado é exportado para um novo arquivo CSV.

Importação de Dados

Nesta etapa do projeto foi desenvolvido um programa com o objetivo de importar os dados coletados anteriormente para uma tabela em um banco de dados PostgreSQL. Para isso, foi utilizada a biblioteca Pandas para ler o arquivo gerado como um novo dataframe e a biblioteca SQLAlchemy para criar uma conexão com o banco de dados e salvar os dados na tabela correspondente.

As passos exatos desenvolvidos pelo pragrama foram:

1. **Configurações de Conexão com o Banco de Dados PostgreSQL**: As configurações de conexão com o banco de dados PostgreSQL foram definidas no início do código, incluindo o nome do banco de dados, usuário, senha e host.

2. **Criação da Conexão com o Banco de Dados:** foi criada uma conexão com o banco de dados PostgreSQL utilizando o SQLAlchemy. A string de conexão foi construída utilizando as configurações definidas anteriormente.

3. **Importação do Arquivo CSV:** O programa leu o arquivo CSV especificado no caminho fornecido utilizando a função read_csv() do Pandas. O DataFrame resultante contém os dados que serão importados para o banco de dados.

4. **Definição do Nome da Tabela:** Foi especificado o nome da tabela no banco de dados onde os dados serão armazenados.

5. **Importação dos Dados para o Banco de Dados PostgreSQL:** Os dados contidos no DataFrame são salvos na tabela do banco de dados PostgreSQL utilizando o método to_sql() do Pandas. Se a tabela já existir, os dados são substituídos (if_exists='replace'). A opção index=False é utilizada para não incluir o índice do DataFrame como uma coluna no banco de dados.

6. **Fechamento da Conexão:** Após a importação dos dados, a conexão com o banco de dados é encerrada utilizando o método dispose() do SQLAlchemy.

O programa demonstrou-se como uma maneira eficiente de importar os dados coletados para o banco de dados PostgreSQL utilizado no projeto. A utilização das bibliotecas Pandas e SQLAlchemy simplificaram o processo de leitura do arquivo CSV, manipulação dos dados e interação com o banco de dados, permitindo uma integração rápida dos dados.

Conclusão

O programa automatizado demonstrou uma abordagem eficiente para a coleta e tratamento dos dados do DATASUS, permitindo análises posteriores mais ágeis e precisas. A utilização do Selenium em conjunto com o WebDriver do Google Chrome possibilitou a interação automatizada com o site, enquanto a manipulação dos dados com o Pandas simplificou o processo de tratamento.

O resultado final é um conjunto de dados tratados e prontos para análise, contribuindo para estudos e tomadas de decisão na área da saúde pública.