

Bootcamp: Engenheiro(a) de Dados

Trabalho Prático

Módulo 4: Pipeline de Dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Extração de Dados;
2. Desenvolvimento de *crawlers*;
3. Transformações e análise de dados.

Enunciado

Você é uma pessoa da área de Engenharia de Dados em uma empresa de consultoria educacional. Os consultores da empresa vão auxiliar um grupo educacional de ensino médio a desenvolver seu planejamento pedagógico e suas metodologias de ensino. Nesse sentido, eles fizeram a você uma requisição de preparar dados do ENEM para análise.

Você deve, portanto, extrair os dados do ENEM, de maneira automatizada, selecionar apenas os alunos que estudaram em alguma escola em Minas Gerais e apresentar algumas análises que vão ajudar o trabalho dos consultores.

Atividades

Você deverá desempenhar as seguintes atividades:

1. Extrair, de maneira programática, os dados do ENEM 2020 (<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>);

2. Filtrar apenas os alunos que estudaram em escolas no estado de Minas Gerais;
3. Salvar um arquivo CSV com os dados já limpos em seu computador;
4. Escrever os dados numa tabela relacional em uma base de dados de sua escolha (MySQL, PostgreSQL, SQL Server etc.);
5. Realizar consultas para responder às perguntas da atividade.