Universidade Federal do Ceará
Campus Quixadá
QXD0099 - Desenvolvimento de Software para Persistência
Prof. Francisco Victor da Silva Pinheiro

Lista 3 - Extração e Manipulação de Dados Estruturados, Semiestruturados e Não Estruturados

1. Scraping de Websites com BeautifulSoup

- **Objetivo**: Praticar a extração de dados de um site usando scraping.
- **Tarefa**: Usando a biblioteca BeautifulSoup, escreva um código que extraia e imprima o título e todos os links de uma página web. A URL pode ser qualquer página pública, como https://example.com.

2. Extração de Texto de Imagens com OCR

- **Objetivo**: Extrair texto de imagens usando OCR.
- **Tarefa**: Usando pytesseract e PIL, escreva um código para carregar uma imagem, extrair o texto nela contido e salvar o resultado num arquivo txt..

3. Implementação Completa de um Extrator de Dados Estruturados e Não Estruturados

- **Objetivo**: Integrar conhecimentos e simular um fluxo completo de extração de dados.
- **Tarefa**: Escreva um código que possa extrair dados de um site (HTML), de um PDF e de uma imagem. O código deve identificar o tipo de cada arquivo, extrair as informações relevantes e exibi-las em um formato organizado.

"A melhor maneira de predizer o futuro é inventá-lo."
Alan Kay