# Guilherme Sousa Inácio

# Teste para Desenvolvedor Python Júnior, com técnicas de WebScraping.

# 1 - Já teve contato com a biblioteca beautifulsoup?

R: Sim, ano passado estudei por conta, baseado em meu interesse em extração de dados e automação, as técnicas de webscraping, e me deparei com essa biblioteca, bem como outras bibliotecas em outras linguagens de programação. Possuo familiaridade com esse tipo de biblioteca pois utilizo a biblioteca pupetter do NodeJS que faz um trabalho semelhante de webscraping.

# 2 - Se sim, poderia explicar seu funcionamento aplicado?

R: Sim. A biblioteca permite que o desenvolvedor extraia dados de uma página web em formato de árvore, sendo muito mais rápido trabalhar dessa forma onde cada elemento HTML é entendido como um objeto do que utilizar outras técnicas mais complexas como RegExp (utilizado para buscas em strings que são textos).

# 3 - Já teve contato com a biblioteca scrapy?

R: Sim, em paralelo ao estudo da biblioteca beautifulsoup, também encontrei conteúdo relacionado ao framework scrapy.

# 4 - Se sim, poderia explicar seu funcionamento aplicado?

R: Sim, a biblioteca scrapy, assim como a beautifulsoup permite o webcrawling de páginas web que não possuem api aberta. Utilizando os chamados "Spiders" da biblioteca se pode se definir como será a navegação pelas páginas e como será manipulado o resultado obtido dessas páginas.

# 5 - Poderia desenvolver um webcrawler para a coleta no site <https://www.melodybrazil.com/>, para todos os CDs disponíveis no site. Os resultados devem ser salvos em uma planilha xls e conter:

# - url da publicação,

# - o nome da publicação,

# - data da publicação,

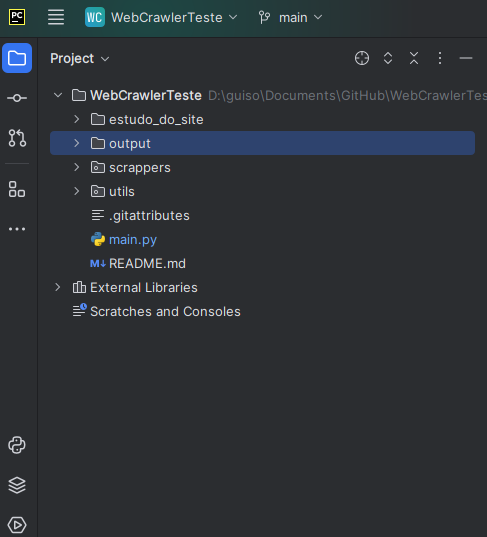
# - publicado por e

# - url de download do cd.

# 6 - Coloque em anexo na resposta deste e-mail, o web crawler desenvolvido e os resultados obtidos em formato xls.

# Sobre o WebScrapper:

## O programa é divido da seguinte forma.



Sendo “WebCrawlerTeste” o diretório raiz do projeto.

A pasta output é onde se encontram (ou se encontrarão após a execução do programa) os arquivos solicitados.

Dentro dessa pasta são gerados três arquivos, sendo eles:

* date\_info.json
* musics.json
* cds\_melody\_brasil.xls
* cds\_melody\_brasil.xlsx