Digesto
Rua Butantã, 434, cj. 34
Pinheiros, São Paulo - SP
+55 (11) 2626-0350

Desafio Digesto

Desenvolvedor(a) Backend - Versão 1.2 - 10/07/2019

OVERVIEW

O Digesto entrega valor para os seus clientes através do fornecimento de dados de forma rápida, acessível e organizada.

Para atingir esse objetivo, trabalhamos diariamente no desenvolvimento, teste e monitoramento de dezenas de *web crawlers* que são capazes de capturar dados da web e salvá-los de forma estruturada.

Neste desafio você irá desenvolver um *web crawler* simples, para que possamos avaliar o seu padrão de código, leitura de documentação da linguagem e bibliotecas, e uso do sistema de versionamento *git*.

O desafio deve ser realizado em etapas, mantendo através dos seus commits no *git* sempre uma versão funcional (mesmo que simplificada) de cada etapa do seu progresso.

OBJETIVOS

1. Desenvolver um *web crawler* em Python que extrai das páginas-alvo e salva em disco os dados especificados.
2. Manter em sistema *git* hospedado de sua preferência (github, gitlab, etc.) o progresso do seu código de acordo com as melhores práticas que regem o uso de tal sistema. Ao término do desafio, compartilhar conosco o endereço público do repositório.
3. Refletir sobre o código conforme ele cresce e propor abstrações. Tais abstrações só devem ser usadas caso ajudem a manter esta pequena aplicação mais coesa, concisa e com baixo nível de acoplamento entre seus componentes internos.

ESPECIFICAÇÃO

Dadas as opções de máquinas nas páginas-alvo, o *crawler* deve extrair os seguintes atributos de cada opção de máquina:

- CPU / VCPU
- MEMORY
- STORAGE / SSD DISK
- BANDWIDTH / TRANSFER
- PRICE [$/mo]

Páginas-alvo:

1. https://www.vultr.com/pricing/ (apenas aba *Vultr Cloud Compute (VC2))*
2. https://www.digitalocean.com/pricing/ (apenas aba *Standard)*

Ao executar um crawler, devem ser disponíveis as seguintes opções não-excludentes:

- --print
  - Imprime resultados na tela
- --save_csv
  - Salva dados em arquivo csv
- --save_json
  - Salva dados em arquivo json

Não deve ser usado o framework Scrapy ou semelhante, mas sinta-se livre para usar alguns conceitos como inspiração, bem como bibliotecas menores utilizadas pelo mesmo (requests, xpath, regex, etc.).

ETAPAS

1 página-alvo, imprime na tela

A primeira etapa exige que o seu *crawler* funcione para a **página alvo 1**, capturando as informações e sendo capaz de imprimi-las na linha de comando em formato arbitrário.

1 página-alvo, imprime na tela, salva em json

A segunda etapa exige que o seu *crawler* funcione para a mesma página-alvo da etapa anterior, **tendo as mesmas funcionalidades da etapa anterior,** mas também sendo capaz de salvar os dados **em um arquivo em formato json**.

1 página-alvo, imprime na tela, salva em json, salva em csv

A terceira etapa exige que o seu *crawler* funcione para a mesma página-alvo da etapa anterior, **tendo as mesmas funcionalidades da etapa anterior,** mas também sendo capaz de salvar os dados **em um arquivo em formato csv**.

2 páginas-alvo

A quarta etapa exige que você extraia as informações também da **página-alvo 2**, **tendo as mesmas funcionalidades da etapa anterior.**

PRAZO

O prazo de entrega é de 7 dias corridos a partir da apresentação do enunciado. Sinta-se livre para nos fazer perguntas caso não tenha entendido completamente o enunciado.