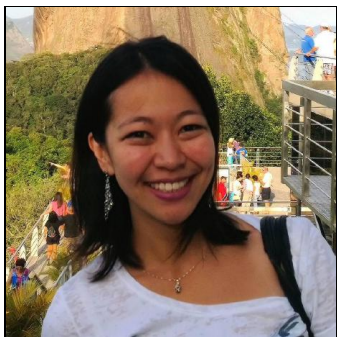


Web Crawling e Scraping com Scrapy e Scrapy Cloud

>whoami



Lidiane Taquehara



- Tecnóloga em Análise e Desenvolvimento de Sistemas pela [FATEC Jundiaí](#)
- Back-end developer na [Love Mondays](#)

Web Scraping

Web Crawling

Web Crawling e Scraping no Love Mondays

The screenshot displays the Love Mondays website, which is a Glassdoor company. The header includes the logo, navigation links for 'Empresas', 'Cargos ou empresa' (with a search bar containing 'Python'), 'Cidade', and a search icon. A 'Minha Conta' link is also present. The main content area shows job listings for 'Python' with a 'CRIAR ALERTA' button and a 'FILTRAR' button. The results are listed as '1 - 10 de 67 resultados'.

Logo	Job Title	Company	Location	Time Ago
	Analista Desenvolvedor Python	GFT	São Paulo, SP	Há 7 dias
	Programador FULL Stack Python	Mazzatech	São Paulo, SP	Há 9 dias
	Analista Desenvolvedor Python Pleno	Verx Consulting	São Paulo, SP	Há 14 dias
	Desenvolvedor Python Pleno / Senior	Verx Consulting	São Paulo, SP	Há 14 dias

Scrapy

- Framework Python voltado para scraping e crawling
- Open Source
- Construído em cima do [Twisted](#)
 - Eficiente
 - Assíncrono

Exemplo de uma Spider

- Acessa a lista dos [100 filmes mais populares no IMDB](#)
- Coleta o título de cada filme
- Acessa a página específica de cada filme
- Coleta também a sinopse e o nome do diretor

[Link para o exemplo no GitHub](#)

0 código

```
name = 'most_popular_movies'
```

O nome que identifica a spider

```
start_urls =
```

```
['https://www.imdb.com/chart/moviemeter']
```

O ponto de partida da spider

```
parse()
```

Método que manipula a resposta recebida por cada requisição feita.

Execução da spider

Gerenciamento dos dados na nuvem

Scrapinghub

- Criação e manutenção do Scrapy
 - Data on Demand
 - Scrapy Cloud

Scrapy Cloud

- Execução de web crawling em nuvem
- Armazenamento dos dados em um banco de alta disponibilidade

Tecnologias



PORTIA



SCRAPY

Dashboard

The screenshot shows the Scrapyhub dashboard interface. At the top, there's a navigation bar with the Scrapyhub logo, a search bar, and links to Scrapy Cloud, Portia, Crawlers, Datasets, and Help. A user profile for Lidiane Taquehara is visible on the right. The main content area is titled 'Most_popular_movies' and shows a job with 100 items, 102 requests, 18 logs, and 19 stats. Below this, there's a 'Job Items' section with filters and buttons for EXPORT, PUBLISH, and SAMPLES. The job details show the director as Ruben Fleischer and the title as Venom. The item list shows three items, each with a director, summary, and title. The first item is 'Venom' by Ruben Fleischer, the second is 'A Star Is Born' by Frank Pierson, and the third is 'A Star Is Born' by Frank Pierson.

scrapinghub Search → Scrapy Cloud ▾ Portia ▾ Crawlers ▾ Datasets ▾ Help ▾ Lidiane Taquehara

Most_popular_movies Job **Items 100** Requests 102 Log 18 Stats 19 Console

1 spider, 0 members

Job Items EXPORT ▾ PUBLISH ▾ SAMPLES ▾

Filter by Field: Choose field... ▾ Choose action... ▾ All Items ▾ CLEAR UPDATE SHOW SCRAPED FIELDS

Item 0 2018-10-18 15:58:27 UTC DOWNLOAD COMMENT

director	Ruben Fleischer
summary	When Eddie Brock acquires the powers of a symbiote, he will have to release his alter-ego "Venom" to save his life.
title	Venom

Item 1 2018-10-18 15:58:42 UTC DOWNLOAD COMMENT

director	Frank Pierson
summary	A has-been rock star falls in love with a young, up-and-coming songstress.
title	A Star Is Born

Item 2 2018-10-18 15:58:42 UTC DOWNLOAD COMMENT

SPIDERS

- Dashboard
- Periodic Jobs

PROJECT

- Usage Stats
- Activity
- Members
- Settings

ADDONS

- Addons Setup

Utilização dos dados

Possibilidades:

- Download (CSV, JSON JSON Lines, XML)
- Publicação como um dataset público no Scrapinghub
- Consumo dos dados através da Scrapy Cloud API

Scrapy Cloud API

- Permite interagir com as spiders e os dados coletados.
- Endpoints:
 - `app.scrapinghub.com`
 - `storage.scrapinghub.com`

python-scrapinghub

- Client Python: python-scrapinghub
- Exemplo simples: <https://github.com/lidimayra/scrapinghub-api-demo>

Para saber mais:

- [Scrapy - Site oficial](#)
- [Scrapy - GitHub](#)
- [Scrapinghub - Site oficial](#)
- [Scrapinghub - GitHub](#)

Muito obrigada!!

Apresentação disponível em:

<https://scrapy-slides.netlify.com>