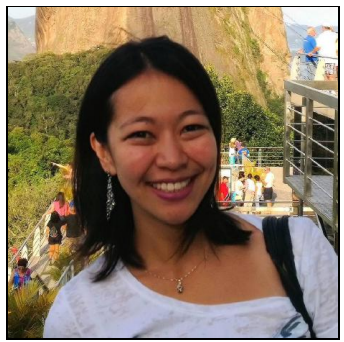
Web Crawling e Scraping com Scrapy e Scrapy Cloud

>whoami



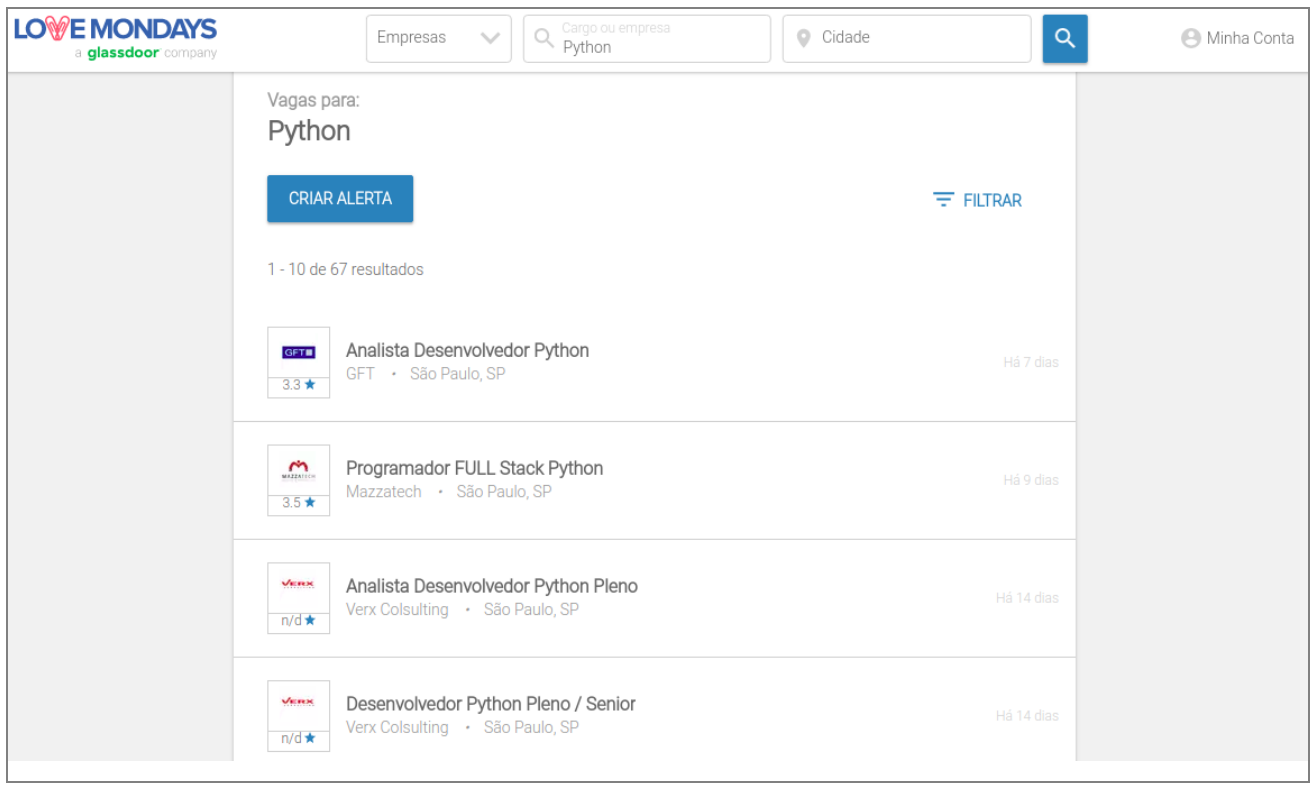
Lidiane Taquehara

- 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



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

O 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

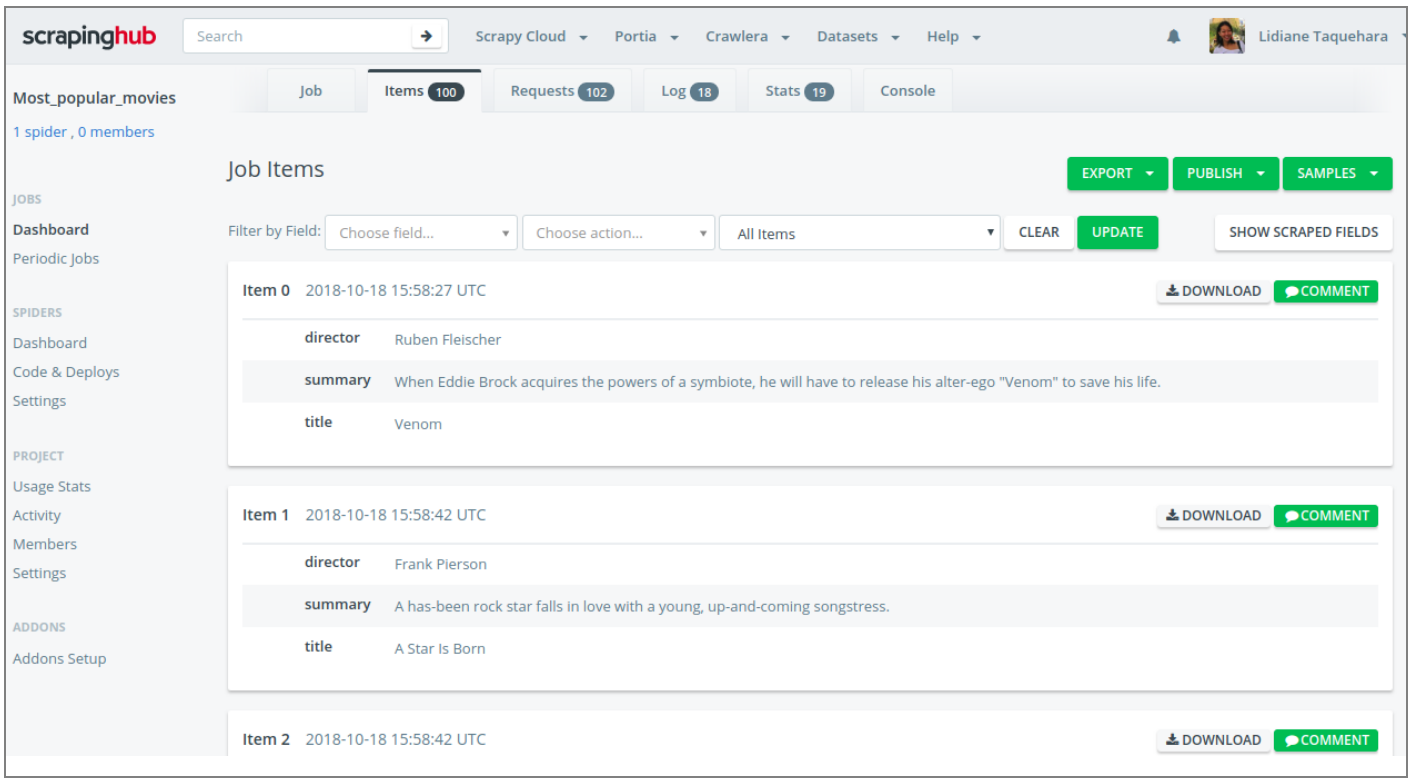
- Execução de web crawling em nuvem
- Armazenamento dos dados em um banco de alta disponibilidade

Tecnologias

PORTIA

SCRAPY

Dashboard


scrapinghub
Search
Scrapy Cloud
Portia
Crawlera
Datasets
Help
Lidiane Taquehara
Most_popular_movies
1 spider , 0 members
Job
Items 100
Requests 102
Log 18
Stats 19
Console
Job Items
EXPORT
PUBLISH
SAMPLES
JOBS
Dashboard
Periodic Jobs
Filter by Field:
Choose field...
Choose action...
All Items
CLEAR
UPDATE
SHOW SCRAPED FIELDS
SPIDERS
Dashboard
Code & Deploys
Settings
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
PROJECT
Usage Stats
Activity
Members
Settings
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
ADDONS
Addons Setup
Item 2 2018-10-18 15:58:42 UTC
DOWNLOAD
COMMENT

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