网站采集工具firecrawl

参考

反正就是一个很牛逼的网站爬取工具，支持纯JS网站，也就是现在流行的VUE等没有html的网站，原理是集成了一个无头chrome浏览器，等页面渲染了才爬取。

特性

- 整站爬取、单个页面爬取、纯JS网站爬取
- 提取为LLM支持的markdown格式，当然了，直接爬取HTML是基本操作
- 只抓取main页面，无意义的重复内容

相关文档和参考地址

- https://docs.firecrawl.dev/introduction
- 网上[illegible]firecrawl[illegible]已经是最新版本啦，而且[illegible]https://docs.firecra[illegible].dev/v1-welcome
- github源代码地址: https://github.com/mendableai/firecrawl

安装

下[illegible]代码, **docker-compose build** 生[illegible]**docker-compose up -d** 运行

playwright-service

这是处理纯JS网站的服务，例如现在几乎所有的网站都是动态生成的，所以这个服务是必须的

极简.env文件，其实不要也可以跑

```
# 核心配置
NUM_WORKERS_PER_QUEUE=8
PORT=3002
HOST=0.0.0.0
REDIS_URL=redis://redis:6379
REDIS_RATE_LIMIT_URL=redis://redis:6379
PLAYWRIGHT_MICROSERVICE_URL=http://playwright-service:3000/html

# 数据库及其他可选配置
USE_DB_AUTHENTICATION=false
```

API□□□□□GPT□□□□□□□□□GPT□

- □□□□

```
curl -X POST http://localhost:3002/v1/crawl \
    -H 'Content-Type: application/json' \
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "url": "https://docs.firecrawl.dev",
      "limit": 100,
      "scrapeOptions": {
        "formats": ["markdown", "html"]
      }
    }'
```

- □□□□□□□□□□□□□□□□□

```
curl -X GET http://localhost:3002/v1/crawl/<jobid>
```

- □□□□ URL

```
curl -X POST http://localhost:3002/v1/scrape \
    -H 'Content-Type: application/json' \
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "url": "https://docs.firecrawl.dev",
      "formats": ["markdown", "html"]
    }'
```

- □□□□□□

```
curl -X POST http://localhost:3002/v1/map \
    -H 'Content-Type: application/json' \
```

```
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "url": "https://firecrawl.dev"
    }'

``

- □□□□

```shell

curl -X POST http://localhost:3002/v1/search \
    -H 'Content-Type: application/json' \
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "query": "AI tools",
      "limit": 5,
      "scrapeOptions": {
        "formats": ["markdown"]
      }
    }'
```

- □□□□□

```
curl -X POST http://localhost:3002/v1/extract \
    -H 'Content-Type: application/json' \
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "url": "https://example.com",
      "extract": {
        "schema": {
          "type": "object",
          "properties": {
            "title": {"type": "string"},
            "price": {"type": "number"}
          }
        }
      }
```
```

```
    }'
```

- □□□□`

```
curl -X POST http://localhost:3002/v1/batch/scrape \
    -H 'Content-Type: application/json' \
    -H 'Authorization: Bearer YOUR_API_KEY' \
    -d '{
      "urls": ["https://example1.com", "https://example2.com"],
      "options": {
        "formats": ["markdown"]
      }
    }'
```

□□□ #5
□ □□□ □□□ 7 □□ 2025 18:17:39
□ □□□ □□□ 18 □□ 2025 03:45:38