پروژه Corlwler

شرح پروژه

# ابزار

* Language: Node JS version 12.18.0
* IDE: WebsStorm version 2020.2
* Software Package Manager: NPM version 16.4.7
* Lib: puppetter version 5.1.0
* Lib: fs
* Lib: dns

# نحوی کار

ابتدا به فایل index.js مراجعه کنید، در متود main یک کلاس Scrap رو با پارامترurl سایت مورد نظر را وارد می کنید. متود init را میتوانید با تعداد فایل هایی که در سایت blocklist.de فراخوانی کنید و این برنامه طوری بوده که خزش را به صورت سطحی انجام می دهد. و در متود apply با پارامتر level که تعداد سطح ها را نشان می دهد می توانید؛ تا سطح مورد نظر خزش کنید. و در اخر متود finish را صدا میزنیم که اطلاعات خزش شده به صورت json در file.json ذخیره گردد.

در فایل read.js می توان اطلاعات خزش شده را از فایل خواند.

# شرح کد

این پروژه شامل 7 فایل js هست که هر کدام را توضیح می دهیم.

### index.js

در ان تابع اصلی برنامه قرار داردو برنامه از اینجا خزش را اغاز می کند.

### scrap.js

در این فایل کلاس اصلی خزش هست و کار اصلی برنامه در همین قرار دارد.

جدول 1. ویژگی های کلاس Scrap

|  |  |
| --- | --- |
| توضیح | نام |
| ابجکت نرمالایزر که در ان linkNormailzer قرار دارد. | normalizer |
| Url شروع خزش | baseUrl |
| صفی که url های خزش شده را نشده می دارد. | queueLinks |
| صفی که url های خزش شده را شده می دارد. | queueVisit |
| ابجکتی که اطلاعات استخراج را در خود نگه می دارد. | scrapOb |
| لیستی از ip های بلاک شده | blockList |
| Browser که از اجرای puppetter بورجود می اید ادرسش در این قرار داد. | browser |
| صفحه اصلی برنامه که از baseUrl است. | mainPage |

جدول 2. توابع های کلاس Scrap

|  |  |
| --- | --- |
| توضیح | نام |
| تابع سازنده با دو پارامتر | constructor(baseUrl,normaizer) |
| تابع برای ایجاد اولیه و گرفتن لیست ای پی های لاک و ایجاد اولیه browser و خزش صفحه اصلی | init(n) |
| برای خزش سطحی با پارامتر level | apply(level) |
| try برای پیمایش صف url ها و خزش در هر یک از url ها | try(queueLinks) |
| خزش در صفحه | evaluate(page) |
| بستن browser و نوشتن در فایل | finish() |

### فایل blocklist.js

شامل یک کلاس برای دانلود فایل های ip های بلاک شده از وبسایت blocklist.de

جدول 3. ویژگی های کلاس BlockList

|  |  |
| --- | --- |
| توضیح | نام |
| ادرس browser | browser |
| صفحه blocklist.de | page |
| ارایه از لینک فایل های ip | arrLink |
| ارایه ip ها | arr |
| ابجکتی از ip هل | ob |

جدول 4. توابع های کلاس BlockList

|  |  |
| --- | --- |
| توضیح | نام |
| تابع سازنده | constructor() |
| ایجاد اولیه browser خزش در صفحه و بدست اوردن url | inti() |
| گرفتن ارایه از ip ها که در اون فایل هست. | getLinkWithURL(url) |
| واکشی ارایه ای از ip ها | fetch(n) |
| واکشی ابجکتی ای از ip ها | fetchOB(n) |
| بستن browser و نوشتن بلاک لیست ها | Close() |

### فایل getIPAddress.js

در این فایل شامل دو کلاس هستش که برای بدست اورد ip از url استفاده می شود.

جدول 5. توابع فایل getIPAdress.js

|  |  |
| --- | --- |
| توضیح | نام |
| تابع برای بدست اورد ip از url | getIPAddress(url) |
| بدست اورد دامنه از url | domainFromURL() |

### normalizer.js

شامل یک متود برای نرمالیز کردن link ها

جدول 6 فایل normaizer.js

|  |  |
| --- | --- |
| توضیح | نام |
| نرمال کردن لینک ها | linkNormalizer(links, baseUrl) |

### فایل read.js

در این فایل شامل دو کلاس هستش که برای بدست اورد ip از u

جدول 7. توابع فایل read.js

|  |  |
| --- | --- |
| توضیح | نام |
| تابع برای خواندن فایل که از خزش بدست امده | readFile(‌) |
| تابع برای خواندن فایل که ip های بلاک شده در ان است. | readBlockList() |