پروژه CORLWLER

شرح پروژه

1) 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

2) نحوی کار

ابتدا به فایل index.js مراجعه کنید، در متود main یک کلاس Scrap رو با پارامترurl سـایت مورد نظر را وارد می کنید. متود init را میتوانید با تعداد فایل هایی که در سـایت blocklist.de فراخوانی کنید و این برنامه طوری بوده که خزش را به صـورت 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.

3) 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

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

جدول 2. توابع های کلاس Scrap

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

فایل blocklist.js

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

جدول 3. ویژگی های کلاس BlockList

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

جدول 4. توابع های کلاس BlockList

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

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

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

normalizer.js

شامل یک متود برای نرمالیز کردن link ها

جدول 6 فایل normaizer.js

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

فایل read.js

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

جدول 7. توابع فایل read.js

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