بسم الله الرحمن الرحیم

راهنمای Web ماژول crawler

تیر ۹۷

فهرست مطالب

فصل ۱

۱-۱نگارش ها

در جدول زیر تاریخچه نگارش های مختلف APIآورده شده است.

تاریخ	نسخه	توضیحات
اسفند ۹۶	۱,۰	نسخه اولیه document آماده شد.
تیر۹۷	۱,۱	خروجی تصحیح شد. خطایابی ماژول انجام شد. Progressbar اضافه شد.

۱-۲ توضیحات اولیه

این ماژول جهت بازدید از یک url و ذخیره محتوای آن نوشته شده است.با توجه به عمق در نظر گرفته شده و قابل تنظیم ، این ماژول، لینک های موجود در url اولیه را هم بازدید کرده و محتوای آنها را نیز ذخیره می کند.در کد مربوط به این ماژول از سرویس Hrobot استفاده شده است. محل ذخیره محتوای url ها مسیر و پوشه مشخصی است و در نهایت ماژول لسیتی از ادرس محل ذخیره urlهای بازدید شده را در قالب entity_property برای کاربر ارسال می کند.

۱-۳ نحوه کار ماژول

این ماژول دو آرگومان به نام های url و depth را به عنوان ورودی دریافت می کند و لیستی ازفایل های crawl شده را به صورت entity_property به عنوان خروجی بر می گرداند.جداول زیر ورودی ها و پارامترهای خروجی این ماژول را نشان می دهد. همچنین تصویر یک نمونه خروجی این ماژول در این لینک قابل مشاهده است.

ورودی	توضیحات
url	ادرس urlاولیه که کاربر قصد crawl کردن ان را دارد
depth	عمق درخواستی بری crawl

جدول مربوط به ورودی ماژول

برای بهبود عملکرد ماژول crawler، دو فاکتور optional به ورودی crawler اضافه شده است، برای استفاده از این موارد کافی است به آن را صورت key-value در ورودی ارسال کنید که توضیحات هر یک به شرح زیر است:

کلید	مقدار	نوضیحات
base_url_constraint	True, False	ارسال این مقدار باعث می شود که crawler فقط صفحاتی را ذخیره کند که دارای url پایه مشابه با url پایه مورد درخواست کاربر باشد. در صورتی که این مقدار ارسال نشود مقدار آن false خواهد بود.
link_limit	2000	این عدد مشخص کننده ی تعداد صفحات مجاز برای ذخیره شدن توسط crawler می باشد. اگر تعداد صفحات ذخیره شده از این عدد بگذرد خطا رخ می دهد. در صورتی که این مقدار ارسال نشود مقدار آن برابر ۲۰۰۰ خواهد بود.

جدول مربوط به ورودی‌های اختیاری ماژول

توجه !

دقت شود که در جدول بالا پارامتر depth که عمق درخواستی را نشان میدهد به صورت اختیاری بوده و در صورت تنظیم نکردن عدد ۱ برای عمق crawl به صورت پیش فرض در نظر گرفته می شود.

entity	Output_parameters
File_content	{ "results": [{ "data": "/home/dpe/Documents/inprogress_project", "type": 9, "properties": [{ "url": "http://www.mybabyname.com/", "type": 1 }], "ref": { "task": "crawl", "depth": 1 } }] }

جدول مربوط به پارامترهای خروجی ماژول

۴-۱ محتوای موجود در socket request

برای استفاده از ماژول crawler باید یک درخواست از طریق socket به فریم وورک ارسال شود. بدنه درخواست ارسالی شامل پارامترهای data و type است. پارامتر type مشخص کننده نوع ورودی ماژول(unstructue,file,domain,path,…) است که idمربوط به انواع مختلف ورودی ها و توضیحات مربوطه که در این لینک موجود است، در این فیلد قرار می گیرد. پارامتر data هم همان داده ارسالی است که شامل پارامترهای ورودی ماژول می باشد.

```
{

  "data": {
        "url": "http://www.mybabyname.com/",

        "depth": 2,
       },

  "type": 9
}
```

نمونه body یک socket request

در تصویر نمونه بالا داریم:

data : داده ارسالی برای استفاده در ماژول

url : آدرس urlدرخواستی جهت crawl

depth : عمق درخواستی برای crawl

type : نوع ورودی ماژول را مشخص می کند که می تواند از نوع unstructured,path,file و یا انواع دیگر باشد و باید id ورودی موردنظر در این قسمت قرار گیرد.

توجه !

دقت شود که در جدول بالا پارامتر type شماره id مربوط ورودی ماژول را نشان می دهد. و چون ورودی ماژول url است ، مقدار این پارامتر همیشه 9 می باشد.

۵-۱ جدول خطاها

جدول زیر شامل خطاهایی است که ماژول crawler در پاسخ به کلاینت بر می گرداند. خطا ها با شماره مشخصی نشان داده می شوند.

کد	نام	توضیحات
۱۰۲	InvalidInputError	فرمت ورودی نادست است
۱۱۹	CrawlLimit	صفحات ذخیره شده از تعداد مجاز عبور کرده است.
۱۲۱	TimeoutError	زمانی که hrobot نتواند در مدتی مشخص (۶۰ ثانیه) وارد صفحه ای شود
۱۲۲	ResultNotSetError	زمانی که نتیجه ای برای خروجی set نشده باشد.
۱۱۴	ResultNotFoundtError	زمانی که نتیجه ای یافت نشود.

۵-۱ progressbar

برای نمایش progressbar به صورت real time. محل به روز رسانی progressbar درون ماژول است. بدین منظور هر قسمت از ماژول که قسمتی از نتیجه را محاسبه می کند. قبل یا بعد از آن باید update_progressbar صدا زده شود.

در ماژول crawler از آن جایی که نمی توان با اطمینان از مقدار کار باقی مانده صحبت کرد. در قسمت مربوط به پیام این ماژول عمق فعلی و تعداد صفحات ذخیره شده و تعداد صفحات باقی مانده در این عمق چاپ می شود. درصد پیشرفت نیز حاصل تقسیم تعداد صفحات ذخیره شده بر تعداد لینک های یافت شده تا این لحظه می باشد. بدیهی است که با زیاد شدن تعداد لینک های پیدا شده در یک لحظه ممکن است درصد پیشرفت کاهش یابد.