دانشگاه صنعتی امیرکبیر

دانشکده‌ی علوم‌کامپیوتر

گردآورنده:

شیده هاشمیان

شماره دانشجویی:

۹۶۱۳۴۲۹

تمرین سوم درس مباحثی در علوم‌کامپیوتر

عنوان : خزش داده ها و جمع آوری دادگان

استاد درس: دکتر اکبری

پاییز ۹۹

۱. مقایسه‌ی دو سایت معرفی شده:

- از نظر تعداد داده:

- سایت **johnlewis**: باتوجه به این که این سایت ارائه دهنده‌ی انواع کالا مانند لباس، ابزار الکتریکی و ... است که یکی از آن دسته‌ها شامل وسایل خانه مانند تخت، مبل، صندلی و میز (که مد نظر این پروژه است) است. که در مقایسه با سایت دیگر داده‌های کمتری دارد. برای مثال برای تخت حدودا ۳۳۸ محصول نتیجه می‌شود.
- سایت **houzz**: این سایت یک سایت مخصوص لوازم خانگی است و انواع محصولات در اسن دسته را شامل می‌شود که بهعث میشود وسعت داده‌ها بیشتر باشد زیرا اختصاصا به این حوزه پرداخته. برای مثال برای تخت حدودا بیش از ۳هزار محصول نتیجه می‌شود.

- از نظر دسترسی:

- سایت **johnlewis**: مسیرهایی که به محصولات مورد نظر این پروژه ختم می‌شود (چه مسیر مستقیم و چه مسیری که با جست‌وجو محصول نتیجه می‌شود) در فایل **robots.txt** این سایت اجزاءی دسترسی به آن‌ها منع شده است و نمی‌توان به آن‌ها دسترسی داشت.
- سایت **houzz**: مسیرهایی از این سایت که محصولات مورد نظر این پروژه ختم می‌شود در حالت مستقیم (در حالت جست‌وجو دسترسی منع شده است) در فایل **robots.txt** این محصول اجازه‌ی دسترسی به آن‌ها داده شده است و می‌توان از آن‌ها برای خزش استفاده کرد.

با توجه به نکته‌های گفته شده در بالا برای مقایسه‌ی این دو سایت، میتوان نتیجه گرفت که سایت **houzz** برای این استفاده سایت مناسبی است.

۲. مقایسه‌ی دو کتاب‌خانه‌ی معرفی شده:

- از نظر کاربرد کلی:

- **Selenium**: این **framework** در اصل برای آزمون خودکار (**automatic test**) اپلیکیشن‌های تحت وب ساخته شده است. با ان حال از آن برای خزش سایت‌ها نیز استفاده می‌شود.
- **Scrapy**: این **framework** برای خزش سایت‌ها ساخته‌شده است. بارزه‌ی شاخص این **framework** است که به‌صورت نامتقارن از چند سایت خزش می‌کند (به این معنا که قبل از تمام شدن یک **task**، **task** دیگر را شروع می‌کند) که این قابلیت آن باعث سریع بودن آن می‌شود.

- نوع داده مورد خزش:

- **Selenium**: برای خرزش داده‌های موجود در **HTML** مناسب بوده، همچنین ساختار آن به شکلی است که کاملا برای خزش داده‌های موجود در **Javascript** نیز مناسب است و به‌راحتی برای این داده‌ها هم می‌تواند مورد استفاده قرار گیرد.

- Scrapy: این framework برای خرزش داده‌های موجود در HTML مناسب بوده، اما برای خزش داده‌های Javascript خیلی مناسب نبوده زیرا روند آن به شکلی است که زمان زیادی می‌برد.

- اندازه‌ی داده مورد خزش:
  - Selenium: این framework برای خزش داده‌ی موجود در یک صفحه، تمام فایل‌های js، css، img را در هر صفحه بررسی می‌کند که اگر تعداد صفحاتی که برای خزش داده‌ی مورد نظر آن را ملاقات می‌کنیم زیاد باشد، این framework عملکرد مناسبی نخواهد داشت.
  - Scrapy: این framework برای خزش تنها آدرس‌های درخواستی را ملاقات کرده که در صورت تعداد صفحاتی که برای خزش داده‌ی مورد نظر نیا به بررسی است زیاد باشد، این framework عملکرد مناسب و سریعی خواهد داشت.
- گسترش پذیری:
  - Selenium: باتوجه به ساده بودن ساختار کلی آن قابلیت گسترش چندانی ندارد و گسترش دادن آن دشوار است.
  - Scrapy: باتوجه به ساختار آن، به راحتی می‌توان آن را گسترش داد. همچنین قابلیت شخصی سازی و پیاده‌سازی (و یا بازنویسی) توابع مختلف توسط برنامه نویس را دارد.

همجنین یک نکته‌ی قابل توجه دیگر این است که برخلاف Selenium برای Scrapy کتاب‌خوانه‌های جانبی زیادی موجود است.

۳. مستند سازی:

- ساختار داده:

این مجموعه داده یک فایل json. است که به‌صورت آرایه‌ای از لغت‌نامه‌هایی به شکل زیر است.

```
{“product_name”: “product_title on the webpage”,
 “images_url_list”: [“first image’s url”, “second image’s url”],
 “description_tags”: [“keyword_1”, “keyword_2”, …]}
```

که در بازه‌ی ۱ و نیم ساعته بیش از ۲۳هزار داده این شکل برای محصولات تخت، مبل، صندلی و میز است که از سایت Houzz می‌شود.

- موارد استفاده:
  - از داده‌های عکس آن می‌توان برای پردازش تصویر و شناسایی اشیاع داخل عکس استفاده کرد (که نوع ساده‌ای از این تسک است باتوجه به این که اکثرا تنها یک شئ در عکس وجود دارد)
  - از داده‌های عکس آن همراه با نام محصولات و استخراج نوع محصول از نام آن، می‌توان برای پردازش تصویر و تشخیص نوع شئ داخل تصویر استفاده کرد.
  - از نام کالا همراه با تگ‌های ویژگی محصولات، می‌توان به‌عنوان داده‌ی یک سیستم پیشنهاد (recommender system) استفاده کرد که با تحلیل پرسش کاربرو تگ‌های ویژگی، نام محصولات مرتبط با نیاز کاربر را به آن نمایش دهد.

- از داده‌های عکس همراه با نام کالاها می‌توان برای آموزش سیستمی که کاربر بتواند با دادن عکسی، نام مرتبط‌ترین محصول به عکس ورودی را از سیستم بگیرد استفاده شود.

۴. نکاتی در مورد فایل ارسالی:

آدرس فایلی که کلاس مربوطه به خزش در آن قرار دارد به‌صورت زیر است
`houzz_scraper/houzz_scraper/spiders/products_spider.py`
و برای اجرای خزنده، لازم به اجرای این فایل است و پس از اجرا، فایل `json` مربوط به داده‌های خزش شده در همین پوشه ای که فایل پایتون قرار دارد ذخیره می‌شود.

یک فایل `data.json` در کنار دیگر فایل‌ها قرار دارد که یک نمونه‌ی کوچکی از تمام اطلاعات بدست آمده توسط خزشگر است.