توضیحات کد:

برای اجرای پروژه می‌توانید در ترمینال به این صورت عمل کنید:

```
scrapy crawl houzz -o <your file>
```

برای پیاده‌سازی این پروژه از Scrapy استفاده کردم که به صورت پیش‌فرض در صورت وقوع مشکل از از دست دادن کلیه مجموعه داده جمع شده جلوگیری می‌کند:

- spiders: یک Spider برای خزش به نام HouzzSpider پیاده کردم که می‌توان لیستی از آدرس دسته محصولات روی سایت Houzz را به عنوان ورودی در start_urls به آن بدهید تا تمامی محصولات این دسته‌ها را خزش کند.
- items: برای نگهداری و ذخیره‌سازی محصولات خزش شده یک Item به نام HouzzProductItem پیاده‌سازی کردم.
- settings: برای جلوگیری از محدود شدن دسترسی از RandomUserAgentMiddleware استفاده کردم.

مقایسه سایت Houzz و Lewis John:

یک تفاوتی که میان این سایت وجود دارد مسیرهای آن‌ها است که در سایت Houzz مسیرها بسیار کوتاه و با یک قاعده مشخص است که حتی با دانستن آن قاعده می‌توانید چندین مسیر را بسازید و به عنوان مثالی اگر مسیری توی قسمت products تعریف شده هم نباشد آن را به عنوان نتیجه جستجو آن کلمه نشان می‌دهد اما سایت John Lewis چنین قاعده ساده یا مشخصی ندارد و همین باعث می‌شود که مسیرهای سایت Houzz در دسترس‌تر و برای خزش مناسب‌تر باشد.

راجع به مسیر robots.txt هم می‌توان گفت سایت Houzz مسیرهایی که به صورت مشخص اجازه خزش را داده است تا حدی بیشتر از سایت John Lewis است که تنها تعداد اندکی از مسیرها را به صورت مشخص اجازه داده است.

مقایسه Scrapy و Selenium:

اصلی‌ترین تفاوت این ابزار در نحوه کار یا برخورد با یک صفحه web است به این معنی که Scrapy با استفاده از فایل HTML فرایند را پیش می‌برد و داده‌ها را پردازش و ذخیره می‌کند اما Selenium نحوه کارش شبیه به تعامل خودکار با مرورگر است.

با توجه به این تفاوت می‌توان گفت وابسته به حجم داده‌ای که قصد جمع‌آوری آن را داریم یا ساختار و محتویات آن سایت می‌توان میان این دو ابزار انتخاب کرد چون با توجه نحوه کار Scrapy سریع‌تر است و برای مجموعه داده‌های حجیم مناسب‌تر است و یا اگر نیاز به تعامل با قسمت‌های مختلف صفحه یا در واقع Javascript شویم Selenium کارایی بیشتری دارد.

مستندسازی مجموعه داده جمع شده:

هر داده در فرایند خزش به صورت یک Scrapy Item نگهداری شده است که به صورت خودکار با توجه به فرمت خروجی مثل json یا csv متناسب و ذخیره شود.

این Item دارای اجزای زیر است:

- url: شامل آدرس محصولی است که اطلاعات آن را خزش کرده‌ایم و به صورت رشته است.
- title: شامل عنوان محصول است و به صورت رشته است.
- images: شامل آدرس ۲تا عکس اول(در صورت وجود) محصول است و به صورت یک لیست از رشته‌ها است.
- keywords: شامل کلمات کلیدی یا مرتبط با محصول یا در واقع محتویات بخش "This Product Has Been Described As" هست و به صورت یک لیست از رشته‌ها است.

برای هر یک از دسته محصولات سایت Houzz حداکثر حدود ۵۰۰۰ یا ۱۴۰ صفحه را نمایش می‌دهد و وابسته به این که تعداد مسیرهای اولیه‌ای که برای خزش مشخص کرده‌اید می‌توانید حدود ۵۰۰۰ برابر آن داده یا صفحه خزش شده داشته باشید اگر کلات اون دسته را خزش کنید اما در هر صفحه هم ۳۶ محصول به صورت پیش‌فرض نمایش داده می‌شود که در صورتی که تعداد صفحاتی را خزش کنید به ازای هر دسته حدود ۳۶ برابر تعداد صفحات داده خواهید داشت و از این داده می‌توان برای مسائل برچسب‌زنی به اشیا و تولید عنوان به صورت خودکار از روی تصاویر و ایجاد موتور جستجوی تصویری و ... استفاده کرد.