

## Homework 3 : Wrapper Learning

### Task 3 – Questions

1. Website - <https://www.tradesy.com/>

Description – Website that supports selling and buying clothing, bags, shoes etc. of different brands.

Field	Description
URL*	URL of the page (Type - og: url)
name	Name of the product
brand	Brand of the product
price	Original price of the product
reduced_price	Reduced price of the product (Discounted)
likes	Number of likes
color	Color of the product
measurements	Measurements of item – bag, clutch, shoes, etc
size	Size of item – clothes
fabric	Type of the material for clothing
washlook	Wash look for jeans eg. Dark Rinse
item	Item ID of the product

\*(optionally considered as extraction)

Sample screenshot of website for main extracted features -

The screenshot shows a web browser window displaying a product page on Tradesy. The browser's address bar shows a local file path. The website's navigation bar includes links for 'New Arrivals', 'On Sale', 'Clothing', 'Bags', 'Shoes', 'Watches', 'Jewelry', 'Accessories', 'Weddings', 'Designers', 'How it Works', and 'Start Selling'. The breadcrumb trail indicates the location: 'Shop / Dresses / Casual Dresses (Maxi) / 3.1 Phillip Lim Casual Dresses (Maxi)'. The product is a '3.1 Phillip Lim Black Cream Silk Scoop Neck Sleeveless Belted Maxi Dress'. The price is shown as 'Sale: \$122.50' with a '10% off' discount, and the 'Original Listing Price' is '\$126.99'. The number of likes is '11'. The condition is 'Gently Used' and the description includes details about the fabric, lining, and construction.

Shop / Dresses / Casual Dresses (Maxi) / 3.1 Phillip Lim Casual Dresses (Maxi)

3.1 Phillip Lim  
Black Cream Silk Scoop Neck Sleeveless Belted Maxi Dress  
Size: 2-XS

10% off  
Sale: \$122.50  
Original Listing Price: \$126.99

ADD TO SHOPPING BAG

11

CONDITION  
Gently Used  
This item has visible signs of wear

DESCRIPTION  
3.1 Phillip Lim Black Cream Silk Scoop Neck Sleeveless Belted Maxi Dress SZ 2 Size: 2Color: Black, CreamMade In: ChinaFabric Content: Silk; Lining: SilkItem Specifics & Details: Black maxi dress. Constructed of silk. Scoop neckline. Sleeveless. Front slit. Cream skinny waist belt. Concealed zipper fastening at left shoulder. Lined

## Other possible extractions -

The screenshot shows a web browser with two tabs: 'Watch The Sopranos S02E02' and '3.1 Phillip Lim Black Cream Silk'. The address bar shows a file path: 'file:///Users/kshah/Documents/workspace/MyCrawler/output/0.html'. The left page displays a Tory Burch messenger bag. Red arrows point from the label 'fabric' to the 'Fabric: Leather' field and from 'measurements' to the 'Measurements: 11.5 x 1 x 6.5 in.' field. The right page displays a Phillip Lim dress. Red arrows point from the label 'item' to the 'Item #: 19590904' field, from 'size' to the 'Size: 2-xs' field, and from 'color' to the 'Color: Black' field. The browser's developer tools are open, showing the 'CONDITION' and 'DESCRIPTION' sections of the dress page.

Item #: 19614361  
Type: Messenger Bag  
Measurements: 11.5 x 1 x 6.5  
Color: Red  
Brand: Tory Burch  
Fabric: Leather  
Style/Collection: Britten Clutch Red Agate Burgundy Leather  
Style Tags: Tory Burch Messenger Bag

Tory Burch  
Britten Clutch Agate Burgundy Leather Red Messenger Bag  
Measurements: 11.5 x 1 x 6.5 in.  
Price: \$294.95 Shipping Included  
Retail Price: \$350.00  
ADD TO SHOPPING BAG  
Finance with Affirm for as low as \$26.00/month. Pay with Credit Card, PayPal, or Affirm.  
CONDITION  
New With Tags  
This item has original tags and shows no visible signs of wear.

3.1 Phillip Lim Black Cream Silk Scoop Neck Sleeveless Belted Maxi Dress SZ 2 Size: 2Color: Black, CreamMade In: ChinaFabric Content: Silk; Lining: SilkItem Specifics & Details: Black maxi dress. Constructed of silk. Scoop neckline. Sleeveless. Front slit. Cream skinny waist belt. Concealed zipper fastening at left shoulder. Lined. Measurements\*: Condition: Pre-owned. This item is in good condition. SKU: 176469

Item #: 19590904  
Type: Casual Dresses (Maxi)  
Size: 2-xs  
Color: Black  
Brand: 3.1 Phillip Lim  
Style/Collection: 3.1 Phillip Lim Black Cream Silk Scoop Neck Sleeveless Belted Maxi Dress  
Style Tags: Phillip Lim, Black, Cream, Silk, Scoop, Neck, Sleeveless, Belted, Maxi, Dress, 3.1 Phillip Lim Casual Dresses (Maxi)

LISTED BY  
The Garage  
chicago  
Ask a Question  
Follow

- I chose Crawler4j for scraping data. Link - <https://github.com/yasserg/crawler4j>  
I found it convenient to play with crawler configurations using this tool than scrapy, like what pages the crawler **should** visit or setting the maximum depth.

I chose BeautifulSoup for extracting data from crawled data.

Link - <https://www.crummy.com/software/BeautifulSoup/>

I found the idea of manually extracting information more useful as a part of the learning process than using automatic tools like ScrapingHub. It helped me understand the issues and nuances faced in wrapper learning.

- Wrapper for this assignment was constructed manually. First of all, it was made sure the webpages followed (almost) a similar pattern by checking ?tref=category in the url, thus assuring a single source schema. After examining set of 20-30 web pages, entities like classes, tags, divs were decided for the corresponding features. For example, Number of likes is present in a div with class name idp-love, so use `soup.find(div, class_ = "idp-love")`. BeautifulSoup was used for searching, parsing and navigating the DOM tree to those specific tags. Further, regex string matching and text processing like removing additional spaces gives the exact required values.

P.S : Code for running wrapper. Assuming you're in current directory and webpages are in data folder –  
python ../data/ 1 (1 to process webpages (required first time), 0 to assume dump is already created and just create json). extractions.json will be created outside the current directory.