Homework 3 : Wrapper Learning

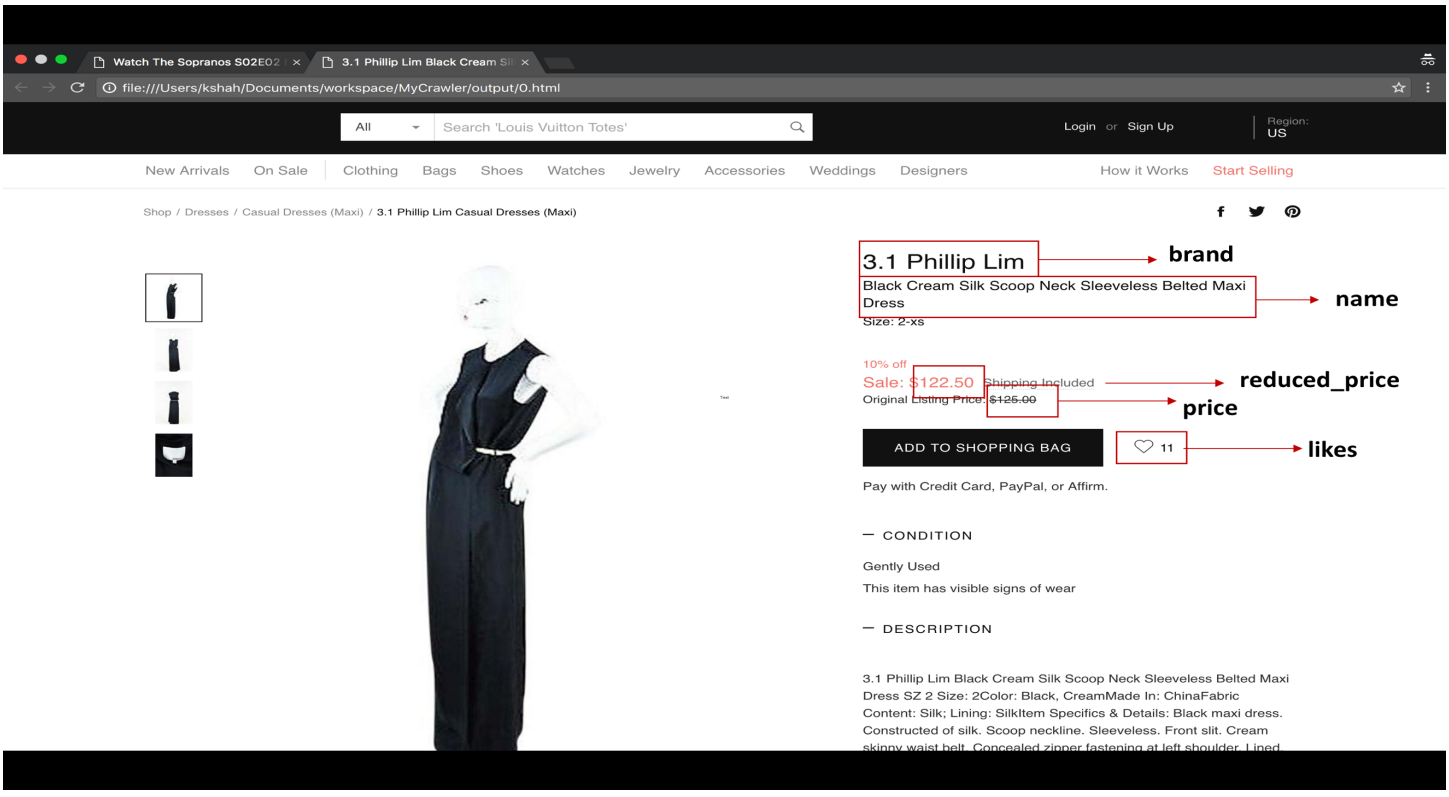
Task 3 – Questions

1. Website - https://www.tradesy.com/
   Description – Website that supports selling and buying clothing, bags, shoes etc. of different brands.

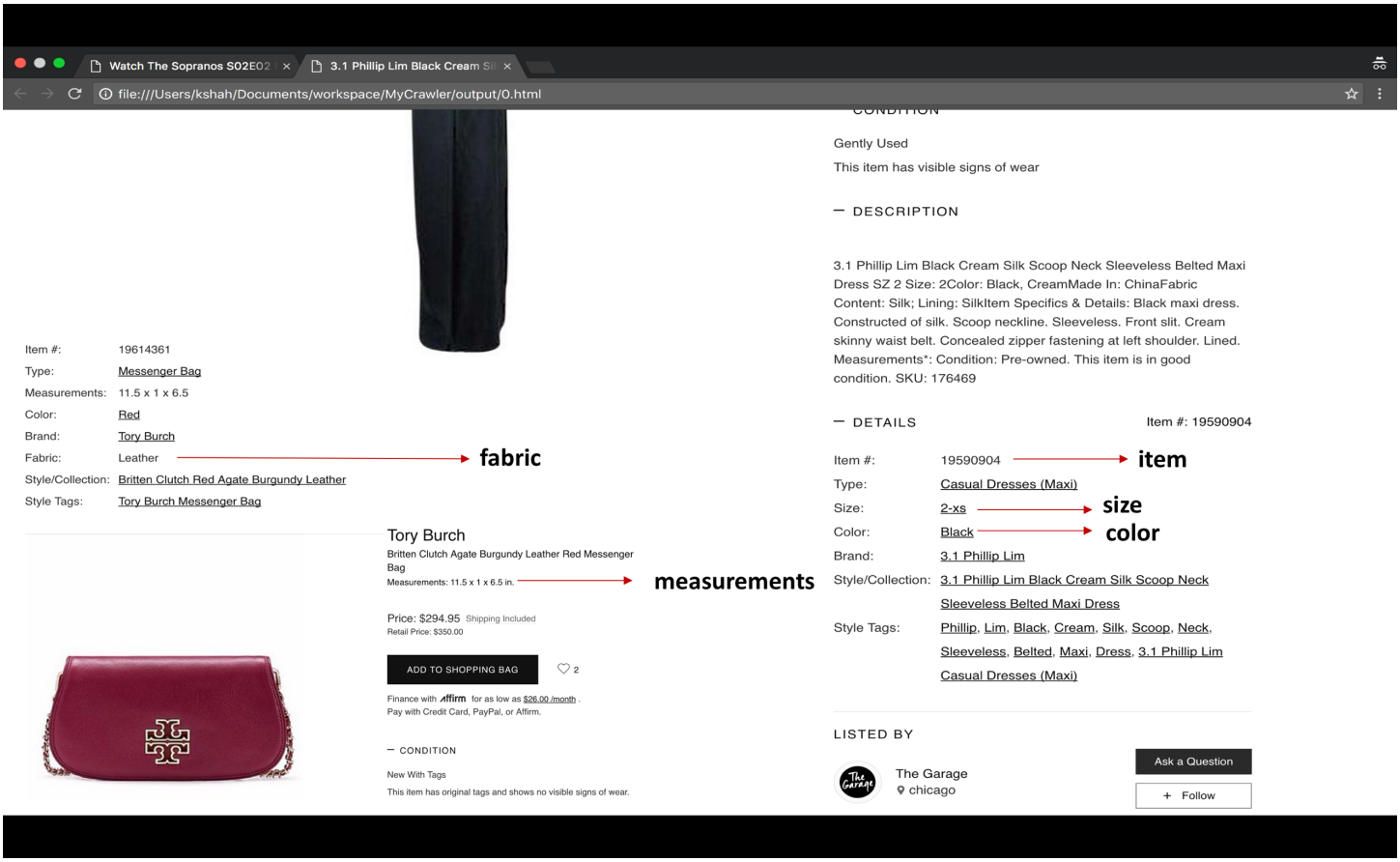
Field	Description
URL*	URL of the page (Type - og: url)
name	Name of the product
brand	Brand of the product
price	Original price of the product
reduced_price	Reduced price of the product (Discounted)
likes	Number of likes
color	Color of the product
measurements	Measurements of item – bag, clutch, shoes, 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 -



Other possible extractions -



2. 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.

3. 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.