Introduction

This is a scrapy data extraction assignment where you need to use Scrapy Framework for data extraction.
The objective of this task is to evaluate the learning, technical and other skills related to a programming environment.
For the given website the candidate must do the following guidelines to extract the data and store in the required format mentioned below.

1. The candidate must develop a scrapy project for the website provided and extract a minimum of 1000 data items from the website and store the data in a database
2. The project should be well structured and modularized.
3. The coding must go through three important steps:
    a. Crawling
        i. Going through each of the URLs from the URL provided and going through each and every pages as well (Proper Pagination)
    b. Parsing
        i. Parsing is to be done in the last depth, where we find product/person/ property details, all of the required fields mentioned below is to be collected using xpath
    c. Cleaning & Data Structuring
        i. The extracted data should be cleaned properly and the data should be structured in the specified format as explained below.

Prerequisite:

1. **Scrapy framework should be used.**
2. **The data extracted should be in CSV and JSON file format and there must be at least 1000 data items extracted from the website.**
3. **Should submit the code via private git repo and dropbox URL of the output CSV and JSON file.**
4. **The code should be in Python3.**
5. **Code should be optimized and reliable.**
6. **Must follow PEP8 standards for the code.**
7. **Read the below instructions carefully and complete the spider based on the exact requirements.**

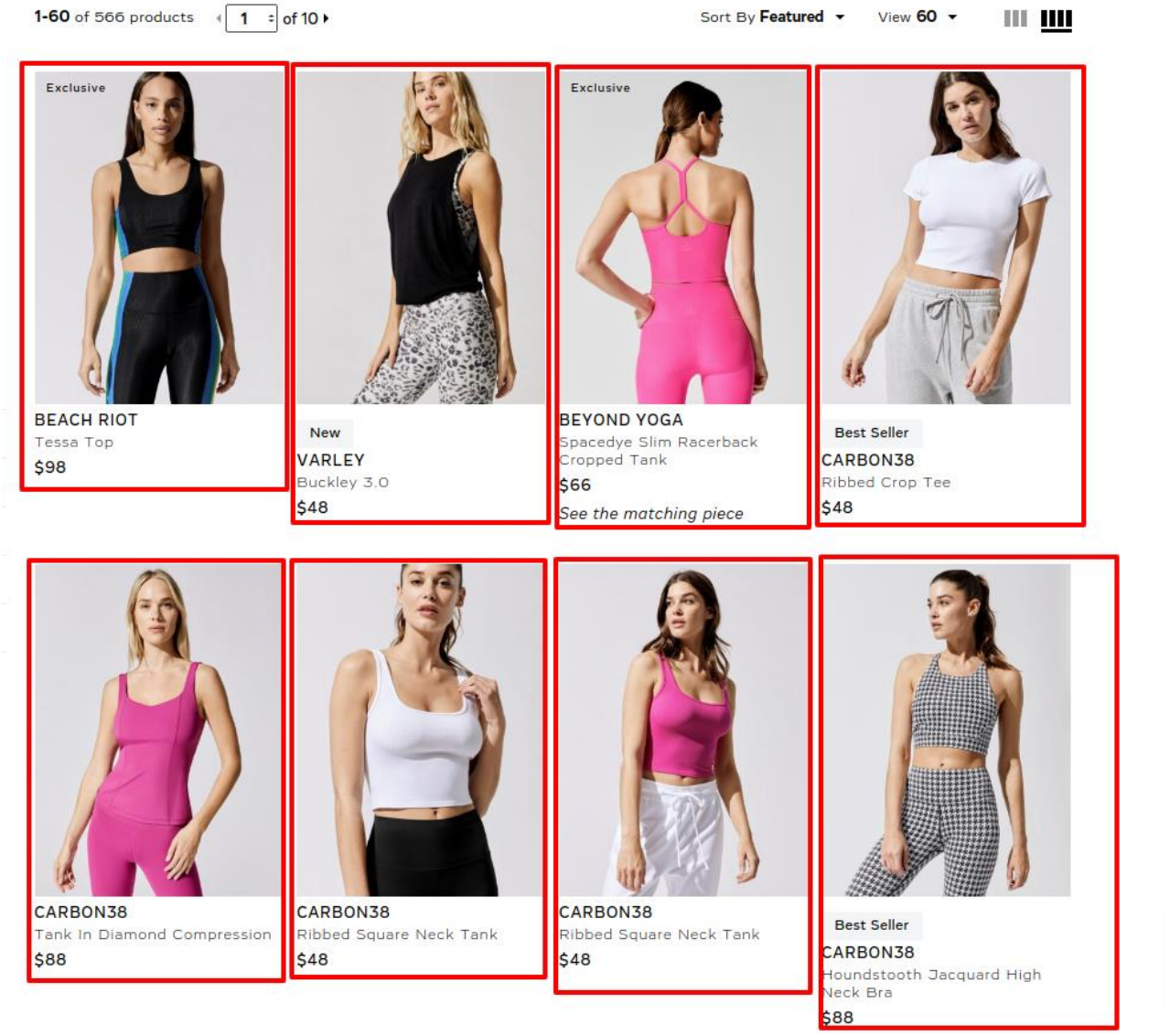
Task

Website to scrape : https://www.carbon38.com/
Total time required to complete: 10 days
Start URL: https://www.carbon38.com/shop-all-activewear/tops [The data extraction process should start from this URL]

Refer following images for guidelines

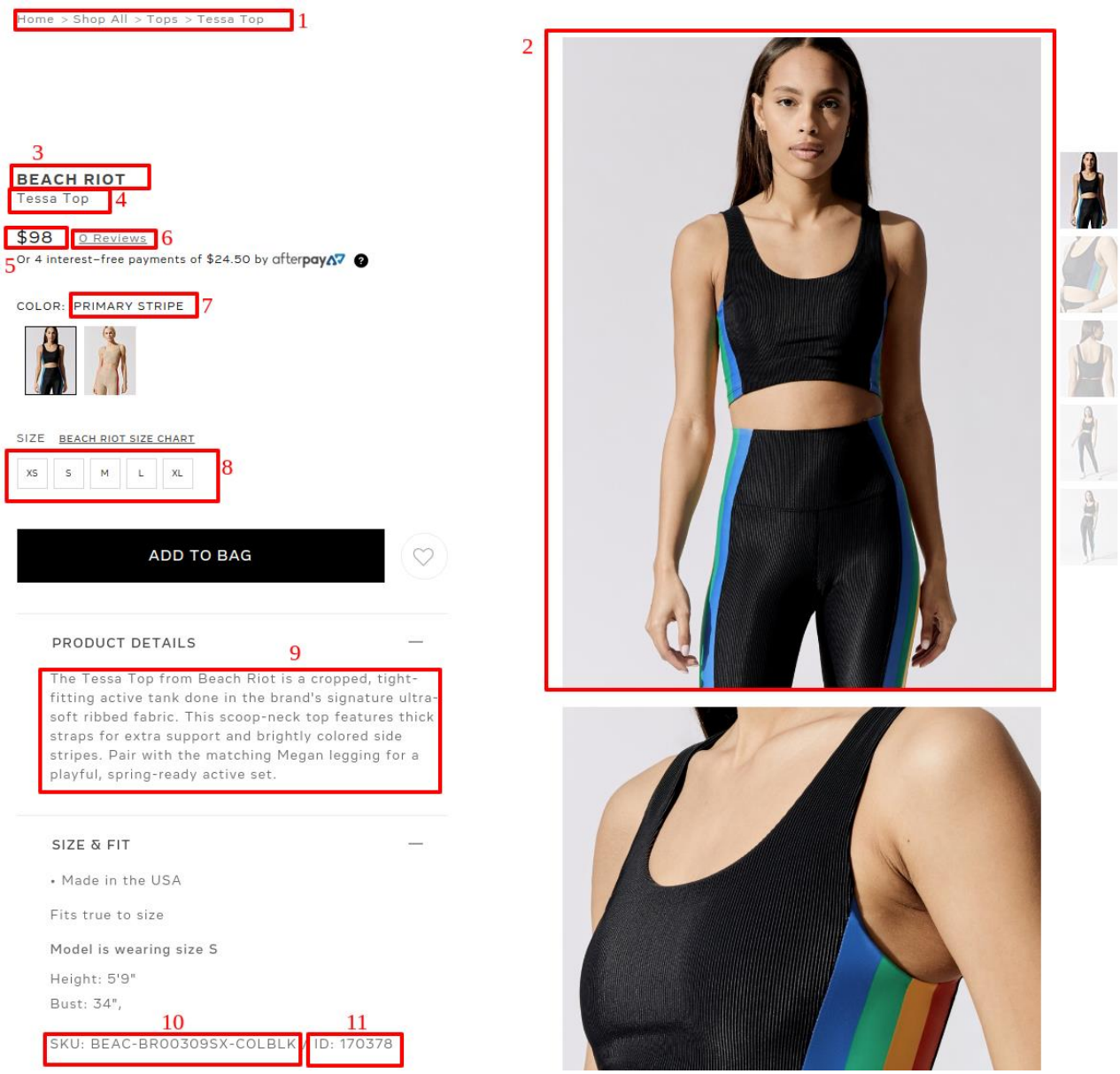
A. Crawling

1. The crawler should go through each and every product URLs as marked in the image, as in the URL above. If there are 60 URLs in the page, it should visit all of them.



2. The crawler should go through each and every page of the profile listing page. i.e. if there are 10 pages, it should visit all 10 pages and every profile URLs in each page.

B. Parsing



The above image can be downloaded [here](#).

Expected data Output:

Eg URL: https://www.carbon38.com/product/tessa-top-primary-stripe

Check the above image for mapping fields to be extracted.

Sl. No	Field Name	Field Type	Example
1	breadcrumbs	*list*	['Home', 'Designers', 'Beach Riot', 'Tessa Top']
2	primary_image_url	*string*	https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2452.jpg
3	brand	*string*	"BEACH RIOT"
4	product_name	*string*	"Tessa Top"
5	price	*string*	"$98"
6	reviews	*string*	"0 Reviews"
7	colour	*string*	"PRIMARY STRIPE"
8	sizes	*list*	['XS', 'S', 'M', 'L', 'XL']
9	description	*string*	"The Tessa Top from Beach Riot is a cropped, tight-fitting active tank done in the brand's signature ultra-soft ribbed fabric. This scoop-neck top features thick straps for extra support and brightly colored side stripes. Pair with the matching Megan legging for a playful, spring-ready active set."
10	sku	*string*	"BEAC-BR00309SX-COLBLK"
11	product_id	*string*	"170378"
12	product_url	*string*	“https://www.carbon38.com/product/tessa-top-primary-stripe”
13	image_urls	*list*	["https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2452.jpg", "https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2467.jpg", "https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2477.jpg", "https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2445.jpg", "https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-biker-shorts-tile-2451.jpg"]

Example Structure of above data item:

```
{
  "breadcrumbs": [
    "Home",
    "Designers",
    "Beach Riot",
    "Tessa Top"
  ],
  "primary_image_url":
"https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-
biker-shorts-tile-2452.jpg",
  "brand": "BEACH RIOT",
  "product_name": "Tessa Top",
  "price": "$98",
  "reviews": "0 Reviews",
  "colour": "PRIMARY STRIPE",
  "sizes": [
    "XS",
    "S",
    "M",
    "L",
    "XL"
  ],
  "description": "The Tessa Top from Beach Riot is a cropped, tight-fitting
active tank done in the brand's signature ultra-soft ribbed fabric. This
scoop-neck top features thick straps for extra support and brightly colored
side stripes. Pair with the matching Megan legging for a playful, spring-
ready active set.",
  "sku": "BEAC-BR00309SX-COLBLK",
  "product_id": "170378",
"product_url":"https://www.carbon38.com/product/tessa-top-primary-stripe",
"image_urls":["https://www.carbon38.com/media/catalog/product/s/u/sund-
su214em40-blusky-biker-shorts-tile-2452.jpg",
"https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-
biker-shorts-tile-2467.jpg",
"https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-
biker-shorts-tile-2477.jpg",
"https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-
biker-shorts-tile-2445.jpg",
"https://www.carbon38.com/media/catalog/product/s/u/sund-su214em40-blusky-
biker-shorts-tile-2451.jpg"]
}
```