Notebook

June 24, 2025

1 Problem 0

[2]:

```
import numpy as np
import pandas as pd
import requests
from bs4 import BeautifulSoup
```

2 Problem 1

Part a

The latest Common Crawl dataset can be accessed on Amazon Web Services' Open Data Sets Sponsorships program on the bucket s3://commoncrawl/, located in the US-East-1 (Northern Virginia) AWS Region. The dataset contains billions of websites. Crawl data is free to access by anyone.

Part b

Based on the article, it seems the Common Crawl is more crucial for the pre-training stage of the development for an LLM. Common Crawl Corpus is used by so mmany AI efforts because it is a massive collection of mostly HTML code and text extracted from billions of URLs across the web. This helps LLM builders maximize the quality, size, and diversity of their training data. However, the corpus needs to be altered because it hosts large amounts of content that is undesirable for AI training, like hate speech and pornography.

Part c

Three reasons why the data cannot be said to be the complete internet is because: 1. Common Crawl's technical infrastructure is based in the United States, which skews crawls toward English content, 2. Common Crawl respects robot.txt, a method used by web domain administrators to tell web crawlers which parts of the domain they are allowed to visit. Several large domains use robot.txt to block Common Crawl's crawler from visiting most or all of their content, and 3. The number of domains blocking Common Crawl is likely to increase in the wake of "data revolts" (Frenkel and Thompson 2023), where content creators aim to stop their data from being used to train LLMs.

Part d

One suggestion about the future development of Common Crawl is for Common Crawl to invest in a more curated and values-oriented crawl. As the article states, at the moment, the discovery and selection of URLs to crawl is almost entirely automated, causing the disadvantage that only "popular" URLs are included (popular in the sense that they are often linked to), which makes content from digitally marginalized communities less likely to be included. If Common Crawl invests more in a community-driven approach, focusing on communities that are purposefully overlooked, it would allow for more valued data that doesn't involve problem areas like racism and homophobia.

3 Problem 2

Part a

```
[ ]: r = requests.get('https://httpbin.org/user-agent')
     useragent = r.json()['user-agent']
     headers = {'User-Agent': useragent,
                 'From': 'sp8me@virginia.edu'}
```

```
[5]: url = 'http://books.toscrape.com/'

     r = requests.get(url, headers = headers)
     r
```

```
[5]: <Response [200]>
```

```
[36]: mysoup = BeautifulSoup(r.text, 'html.parser')
      mysoup
```

```
[36]: <!DOCTYPE html>

      <!--[if lt IE 7]>      <html lang="en-us" class="no-js lt-ie9 lt-ie8 lt-ie7">
      <![endif]-->
      <!--[if IE 7]>         <html lang="en-us" class="no-js lt-ie9 lt-ie8">
      <![endif]-->
      <!--[if IE 8]>         <html lang="en-us" class="no-js lt-ie9"> <![endif]-->
      <!--[if gt IE 8]><!--> <html class="no-js" lang="en-us"> <!--<![endif]-->
      <head>
      <title>
          All products | Books to Scrape - Sandbox
      </title>
      <meta content="text/html; charset=utf-8" http-equiv="content-type"/>
      <meta content="24th Jun 2016 09:29" name="created"/>
      <meta content="" name="description"/>
      <meta content="width=device-width" name="viewport"/>
      <meta content="NOARCHIVE,NOCACHE" name="robots"/>
      <!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
      <!--[if lt IE 9]>
              <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></script>
              <![endif]-->
      <link href="static/oscar/favicon.ico" rel="shortcut icon"/>
      <link href="static/oscar/css/styles.css" rel="stylesheet" type="text/css"/>
      <link href="static/oscar/js/bootstrap-datetimepicker/bootstrap-
```

```
datetimepicker.css" rel="stylesheet"/>
<link href="static/oscar/css/datetimepicker.css" rel="stylesheet"
type="text/css"/>
</head>
<body class="default" id="default">
<header class="header container-fluid">
<div class="page_inner">
<div class="row">
<div class="col-sm-8 h1"><a href="index.html">Books to Scrape</a><small> We love
being scraped!</small>
</div>
</div>
</div>
</header>
<div class="container-fluid page">
<div class="page_inner">
<ul class="breadcrumb">
<li>
<a href="index.html">Home</a>
</li>
<li class="active">All products</li>
</ul>
<div class="row">
<aside class="sidebar col-sm-4 col-md-3">
<div id="promotions_left">
</div>
<div class="side_categories">
<ul class="nav nav-list">
<li>
<a href="catalogue/category/books_1/index.html">

                                Books

                        </a>
<ul>
<li>
<a href="catalogue/category/books/travel_2/index.html">

                                Travel

                        </a>
</li>
<li>
<a href="catalogue/category/books/mystery_3/index.html">

                                Mystery
                        </a>
</li>
<li>
<a href="catalogue/category/books/historical-fiction_4/index.html">

                                Historical Fiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/sequential-art_5/index.html">

                                Sequential Art

                        </a>
</li>
<li>
<a href="catalogue/category/books/classics_6/index.html">

                                Classics

                        </a>
</li>
<li>
<a href="catalogue/category/books/philosophy_7/index.html">

                                Philosophy

                        </a>
</li>
<li>
<a href="catalogue/category/books/romance_8/index.html">

                                Romance

                        </a>
</li>
<li>
<a href="catalogue/category/books/womens-fiction_9/index.html">

                                Womens Fiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/fiction_10/index.html">
```

```
                                Fiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/childrens_11/index.html">

                                Childrens

                        </a>
</li>
<li>
<a href="catalogue/category/books/religion_12/index.html">

                                Religion

                        </a>
</li>
<li>
<a href="catalogue/category/books/nonfiction_13/index.html">

                                Nonfiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/music_14/index.html">

                                Music

                        </a>
</li>
<li>
<a href="catalogue/category/books/default_15/index.html">

                                Default

                        </a>
</li>
<li>
<a href="catalogue/category/books/science-fiction_16/index.html">

                                Science Fiction

                        </a>
</li>
<li>
```

```
<a href="catalogue/category/books/sports-and-games_17/index.html">

                                Sports and Games

                        </a>
</li>
<li>
<a href="catalogue/category/books/add-a-comment_18/index.html">

                                Add a comment

                        </a>
</li>
<li>
<a href="catalogue/category/books/fantasy_19/index.html">

                                Fantasy

                        </a>
</li>
<li>
<a href="catalogue/category/books/new-adult_20/index.html">

                                New Adult

                        </a>
</li>
<li>
<a href="catalogue/category/books/young-adult_21/index.html">

                                Young Adult

                        </a>
</li>
<li>
<a href="catalogue/category/books/science_22/index.html">

                                Science

                        </a>
</li>
<li>
<a href="catalogue/category/books/poetry_23/index.html">

                                Poetry

                        </a>
```

```
</li>
<li>
<a href="catalogue/category/books/paranormal_24/index.html">

                                Paranormal

                        </a>
</li>
<li>
<a href="catalogue/category/books/art_25/index.html">

                                Art

                        </a>
</li>
<li>
<a href="catalogue/category/books/psychology_26/index.html">

                                Psychology

                        </a>
</li>
<li>
<a href="catalogue/category/books/autobiography_27/index.html">

                                Autobiography

                        </a>
</li>
<li>
<a href="catalogue/category/books/parenting_28/index.html">

                                Parenting

                        </a>
</li>
<li>
<a href="catalogue/category/books/adult-fiction_29/index.html">

                                Adult Fiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/humor_30/index.html">

                                Humor
                        </a>
</li>
<li>
<a href="catalogue/category/books/horror_31/index.html">

                                Horror

                        </a>
</li>
<li>
<a href="catalogue/category/books/history_32/index.html">

                                History

                        </a>
</li>
<li>
<a href="catalogue/category/books/food-and-drink_33/index.html">

                                Food and Drink

                        </a>
</li>
<li>
<a href="catalogue/category/books/christian-fiction_34/index.html">

                                Christian Fiction

                        </a>
</li>
<li>
<a href="catalogue/category/books/business_35/index.html">

                                Business

                        </a>
</li>
<li>
<a href="catalogue/category/books/biography_36/index.html">

                                Biography

                        </a>
</li>
<li>
<a href="catalogue/category/books/thriller_37/index.html">
```

```
                                Thriller

                        </a>
</li>
<li>
<a href="catalogue/category/books/contemporary_38/index.html">

                                Contemporary

                        </a>
</li>
<li>
<a href="catalogue/category/books/spirituality_39/index.html">

                                Spirituality

                        </a>
</li>
<li>
<a href="catalogue/category/books/academic_40/index.html">

                                Academic

                        </a>
</li>
<li>
<a href="catalogue/category/books/self-help_41/index.html">

                                Self Help

                        </a>
</li>
<li>
<a href="catalogue/category/books/historical_42/index.html">

                                Historical

                        </a>
</li>
<li>
<a href="catalogue/category/books/christian_43/index.html">

                                Christian

                        </a>
</li>
```

```
<li>
<a href="catalogue/category/books/suspense_44/index.html">

                                Suspense

                        </a>
</li>
<li>
<a href="catalogue/category/books/short-stories_45/index.html">

                                Short Stories

                        </a>
</li>
<li>
<a href="catalogue/category/books/novels_46/index.html">

                                Novels

                        </a>
</li>
<li>
<a href="catalogue/category/books/health_47/index.html">

                                Health

                        </a>
</li>
<li>
<a href="catalogue/category/books/politics_48/index.html">

                                Politics

                        </a>
</li>
<li>
<a href="catalogue/category/books/cultural_49/index.html">

                                Cultural

                        </a>
</li>
<li>
<a href="catalogue/category/books/erotica_50/index.html">

                                Erotica
                        </a>
</li>
<li>
<a href="catalogue/category/books/crime_51/index.html">

                                Crime

                        </a>
</li>
</ul></li>
</ul>
</div>
</aside>
<div class="col-sm-8 col-md-9">
<div class="page-header action">
<h1>All products</h1>
</div>
<div id="messages">
</div>
<div id="promotions">
</div>
<form class="form-horizontal" method="get">
<div style="display:none">
</div>
<strong>1000</strong> results - showing <strong>1</strong> to
<strong>20</strong>.




    </form>
<section>
<div class="alert alert-warning" role="alert"><strong>Warning!</strong> This is
a demo website for web scraping purposes. Prices and ratings here were randomly
assigned and have no real meaning.</div>
<div>
<ol class="row">
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/a-light-in-the-attic_1000/index.html"><img alt="A Light in
the Attic" class="thumbnail"
src="media/cache/2c/da/2cdad67c44b002e7ead0cc35693c0e8b.jpg"/></a>
</div>
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
```

```
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/a-light-in-the-attic_1000/index.html" title="A Light in
the Attic">A Light in the …</a></h3>
<div class="product_price">
<p class="price_color">Â£51.77</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/tipping-the-velvet_999/index.html"><img alt="Tipping the
Velvet" class="thumbnail"
src="media/cache/26/0c/260c6ae16bce31c8f8c95daddd9f4a1c.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/tipping-the-velvet_999/index.html" title="Tipping the
Velvet">Tipping the Velvet</a></h3>
<div class="product_price">
<p class="price_color">Â£53.74</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
```

```
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/soumission_998/index.html"><img alt="Soumission"
class="thumbnail"
src="media/cache/3e/ef/3eef99c9d9adef34639f510662022830.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/soumission_998/index.html"
title="Soumission">Soumission</a></h3>
<div class="product_price">
<p class="price_color">Â£50.10</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/sharp-objects_997/index.html"><img alt="Sharp Objects"
class="thumbnail"
src="media/cache/32/51/3251cf3a3412f53f339e42cac2134093.jpg"/></a>
</div>
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
```

```
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/sharp-objects_997/index.html" title="Sharp Objects">Sharp
Objects</a></h3>
<div class="product_price">
<p class="price_color">Â£47.82</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/sapiens-a-brief-history-of-humankind_996/index.html"><img
alt="Sapiens: A Brief History of Humankind" class="thumbnail"
src="media/cache/be/a5/bea5697f2534a2f86a3ef27b5a8c12a6.jpg"/></a>
</div>
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/sapiens-a-brief-history-of-humankind_996/index.html"
title="Sapiens: A Brief History of Humankind">Sapiens: A Brief History
…</a></h3>
<div class="product_price">
<p class="price_color">Â£54.23</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
```

```
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/the-requiem-red_995/index.html"><img alt="The Requiem Red"
class="thumbnail"
src="media/cache/68/33/68339b4c9bc034267e1da611ab3b34f8.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/the-requiem-red_995/index.html" title="The Requiem
Red">The Requiem Red</a></h3>
<div class="product_price">
<p class="price_color">Â£22.65</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/the-dirty-little-secrets-of-getting-your-dream-
job_994/index.html"><img alt="The Dirty Little Secrets of Getting Your Dream
Job" class="thumbnail"
src="media/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.jpg"/></a>
</div>
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
```

```
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/the-dirty-little-secrets-of-getting-your-dream-
job_994/index.html" title="The Dirty Little Secrets of Getting Your Dream
Job">The Dirty Little Secrets …</a></h3>
<div class="product_price">
<p class="price_color">Â£33.34</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/the-coming-woman-a-novel-based-on-the-life-of-the-infamous-
feminist-victoria-woodhull_993/index.html"><img alt="The Coming Woman: A Novel
Based on the Life of the Infamous Feminist, Victoria Woodhull" class="thumbnail"
src="media/cache/3d/54/3d54940e57e662c4dd1f3ff00c78cc64.jpg"/></a>
</div>
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/the-coming-woman-a-novel-based-on-the-life-of-the-
infamous-feminist-victoria-woodhull_993/index.html" title="The Coming Woman: A
Novel Based on the Life of the Infamous Feminist, Victoria Woodhull">The Coming
Woman: A …</a></h3>
<div class="product_price">
<p class="price_color">Â£17.93</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock
```

```
</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/the-boys-in-the-boat-nine-americans-and-their-epic-quest-for-
gold-at-the-1936-berlin-olympics_992/index.html"><img alt="The Boys in the Boat:
Nine Americans and Their Epic Quest for Gold at the 1936 Berlin Olympics"
class="thumbnail"
src="media/cache/66/88/66883b91f6804b2323c8369331cb7dd1.jpg"/></a>
</div>
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/the-boys-in-the-boat-nine-americans-and-their-epic-quest-
for-gold-at-the-1936-berlin-olympics_992/index.html" title="The Boys in the
Boat: Nine Americans and Their Epic Quest for Gold at the 1936 Berlin
Olympics">The Boys in the …</a></h3>
<div class="product_price">
<p class="price_color">Â£22.60</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
```

```
<a href="catalogue/the-black-maria_991/index.html"><img alt="The Black Maria"
class="thumbnail"
src="media/cache/58/46/5846057e28022268153beff6d352b06c.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/the-black-maria_991/index.html" title="The Black
Maria">The Black Maria</a></h3>
<div class="product_price">
<p class="price_color">Â£52.15</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/starving-hearts-triangular-trade-
trilogy-1_990/index.html"><img alt="Starving Hearts (Triangular Trade Trilogy,
#1)" class="thumbnail"
src="media/cache/be/f4/bef44da28c98f905a3ebec0b87be8530.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/starving-hearts-triangular-trade-
trilogy-1_990/index.html" title="Starving Hearts (Triangular Trade Trilogy,
#1)">Starving Hearts (Triangular Trade …</a></h3>
<div class="product_price">
```

```
<p class="price_color">Â£13.99</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/shakespeares-sonnets_989/index.html"><img alt="Shakespeare's
Sonnets" class="thumbnail"
src="media/cache/10/48/1048f63d3b5061cd2f424d20b3f9b666.jpg"/></a>
</div>
<p class="star-rating Four">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/shakespeares-sonnets_989/index.html" title="Shakespeare's
Sonnets">Shakespeare's Sonnets</a></h3>
<div class="product_price">
<p class="price_color">Â£20.66</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
```

```
<div class="image_container">
<a href="catalogue/set-me-free_988/index.html"><img alt="Set Me Free"
class="thumbnail"
src="media/cache/5b/88/5b88c52633f53cacf162c15f4f823153.jpg"/></a>
</div>
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/set-me-free_988/index.html" title="Set Me Free">Set Me
Free</a></h3>
<div class="product_price">
<p class="price_color">Â£17.46</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/scott-pilgrims-precious-little-life-scott-
pilgrim-1_987/index.html"><img alt="Scott Pilgrim's Precious Little Life (Scott
Pilgrim #1)" class="thumbnail"
src="media/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.jpg"/></a>
</div>
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/scott-pilgrims-precious-little-life-scott-
pilgrim-1_987/index.html" title="Scott Pilgrim's Precious Little Life (Scott
Pilgrim #1)">Scott Pilgrim's Precious Little …</a></h3>
```

```
<div class="product_price">
<p class="price_color">Â£52.29</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/rip-it-up-and-start-again_986/index.html"><img alt="Rip it Up
and Start Again" class="thumbnail"
src="media/cache/81/c4/81c4a973364e17d01f217e1188253d5e.jpg"/></a>
</div>
<p class="star-rating Five">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/rip-it-up-and-start-again_986/index.html" title="Rip it
Up and Start Again">Rip it Up and …</a></h3>
<div class="product_price">
<p class="price_color">Â£35.02</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
```

```
<article class="product_pod">
<div class="image_container">
<a href="catalogue/our-band-could-be-your-life-scenes-from-the-american-indie-
underground-1981-1991_985/index.html"><img alt="Our Band Could Be Your Life:
Scenes from the American Indie Underground, 1981-1991" class="thumbnail"
src="media/cache/54/60/54607fe8945897cdcced0044103b10b6.jpg"/></a>
</div>
<p class="star-rating Three">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/our-band-could-be-your-life-scenes-from-the-american-
indie-underground-1981-1991_985/index.html" title="Our Band Could Be Your Life:
Scenes from the American Indie Underground, 1981-1991">Our Band Could Be
…</a></h3>
<div class="product_price">
<p class="price_color">Â£57.25</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/olio_984/index.html"><img alt="Olio" class="thumbnail"
src="media/cache/55/33/553310a7162dfbc2c6d19a84da0df9e1.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/olio_984/index.html" title="Olio">Olio</a></h3>
```

```
<div class="product_price">
<p class="price_color">Â£23.88</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/mesaerion-the-best-science-fiction-
stories-1800-1849_983/index.html"><img alt="Mesaerion: The Best Science Fiction
Stories 1800-1849" class="thumbnail"
src="media/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.jpg"/></a>
</div>
<p class="star-rating One">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/mesaerion-the-best-science-fiction-
stories-1800-1849_983/index.html" title="Mesaerion: The Best Science Fiction
Stories 1800-1849">Mesaerion: The Best Science …</a></h3>
<div class="product_price">
<p class="price_color">Â£37.59</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
```

```
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/libertarianism-for-beginners_982/index.html"><img
alt="Libertarianism for Beginners" class="thumbnail"
src="media/cache/0b/bc/0bbcd0a6f4bcd81ccb1049a52736406e.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/libertarianism-for-beginners_982/index.html"
title="Libertarianism for Beginners">Libertarianism for Beginners</a></h3>
<div class="product_price">
<p class="price_color">Â£51.33</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
<article class="product_pod">
<div class="image_container">
<a href="catalogue/its-only-the-himalayas_981/index.html"><img alt="It's Only
the Himalayas" class="thumbnail"
src="media/cache/27/a5/27a53d0bb95bdd88288eaf66c9230d7e.jpg"/></a>
</div>
<p class="star-rating Two">
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
<i class="icon-star"></i>
</p>
<h3><a href="catalogue/its-only-the-himalayas_981/index.html" title="It's Only
```

```
the Himalayas">It's Only the Himalayas</a></h3>
<div class="product_price">
<p class="price_color">Â£45.17</p>
<p class="instock availability">
<i class="icon-ok"></i>

        In stock

</p>
<form>
<button class="btn btn-primary btn-block" data-loading-text="Adding…"
type="submit">Add to basket</button>
</form>
</div>
</article>
</li>
</ol>
<div>
<ul class="pager">
<li class="current">

                Page 1 of 50

            </li>
<li class="next"><a href="catalogue/page-2.html">next</a></li>
</ul>
</div>
</div>
</section>
</div>
</div><!-- /row -->
</div><!-- /page_inner -->
</div><!-- /container-fluid -->
<footer class="footer container-fluid">
</footer>
<!-- jQuery -->
<script
src="http://ajax.googleapis.com/ajax/libs/jquery/1.9.1/jquery.min.js"></script>
<script>window.jQuery || document.write('<script
src="static/oscar/js/jquery/jquery-1.9.1.min.js"><\/script>')</script>
<!-- Twitter Bootstrap -->
<script src="static/oscar/js/bootstrap3/bootstrap.min.js"
type="text/javascript"></script>
<!-- Oscar -->
<script charset="utf-8" src="static/oscar/js/oscar/ui.js"
type="text/javascript"></script>
<script charset="utf-8" src="static/oscar/js/bootstrap-datetimepicker/bootstrap-
```

```
datetimepicker.js" type="text/javascript"></script>
<script charset="utf-8" src="static/oscar/js/bootstrap-
datetimepicker/locales/bootstrap-datetimepicker.all.js"
type="text/javascript"></script>
<script type="text/javascript">
            $(function() {



    oscar.init();

    oscar.search.init();

            });
        </script>
<!-- Version: N/A -->
</body>
</html>
```

Part b

[30]:

```
titles = [h3.find('a')['title'] for h3 in mysoup.find_all('h3')]
titles
```

[30]:

```
['A Light in the Attic',
 'Tipping the Velvet',
 'Soumission',
 'Sharp Objects',
 'Sapiens: A Brief History of Humankind',
 'The Requiem Red',
 'The Dirty Little Secrets of Getting Your Dream Job',
 'The Coming Woman: A Novel Based on the Life of the Infamous Feminist, Victoria
Woodhull',
 'The Boys in the Boat: Nine Americans and Their Epic Quest for Gold at the 1936
Berlin Olympics',
 'The Black Maria',
 'Starving Hearts (Triangular Trade Trilogy, #1)',
 "Shakespeare's Sonnets",
 'Set Me Free',
 "Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)",
 'Rip it Up and Start Again',
 'Our Band Could Be Your Life: Scenes from the American Indie Underground,
1981-1991',
 'Olio',
 'Mesaerion: The Best Science Fiction Stories 1800-1849',
 'Libertarianism for Beginners',
 "It's Only the Himalayas"]
```

Part c

[29]:
```
prices = [p.text.replace('Â£', '') for p in mysoup.find_all('p', class_ =␣
 ↪'price_color')]
prices
```

[29]:
```
['51.77',
 '53.74',
 '50.10',
 '47.82',
 '54.23',
 '22.65',
 '33.34',
 '17.93',
 '22.60',
 '52.15',
 '13.99',
 '20.66',
 '17.46',
 '52.29',
 '35.02',
 '57.25',
 '23.88',
 '37.59',
 '51.33',
 '45.17']
```

Part d

[31]:
```
ratings = [p['class'][1] for p in mysoup.find_all('p', class_ = 'star-rating')]
ratings
```

[31]:
```
['Three',
 'One',
 'One',
 'Four',
 'Five',
 'One',
 'Four',
 'Three',
 'Four',
 'One',
 'Two',
 'Four',
 'Five',
 'Five',
 'Five',
 'Three',
 'One',
 'One',
 'Two',
 'Two']
```

Part e

```
[32]: url_2 = 'http://books.toscrape.com/'
      img_url = [url_2 + img['src'].replace('../', '') for img in mysoup.
       ↪find_all('img')]
      img_url
```

```
[32]: ['http://books.toscrape.com/media/cache/2c/da/2cdad67c44b002e7ead0cc35693c0e8b.j
      pg',
       'http://books.toscrape.com/media/cache/26/0c/260c6ae16bce31c8f8c95daddd9f4a1c.j
      pg',
       'http://books.toscrape.com/media/cache/3e/ef/3eef99c9d9adef34639f510662022830.j
      pg',
       'http://books.toscrape.com/media/cache/32/51/3251cf3a3412f53f339e42cac2134093.j
      pg',
       'http://books.toscrape.com/media/cache/be/a5/bea5697f2534a2f86a3ef27b5a8c12a6.j
      pg',
       'http://books.toscrape.com/media/cache/68/33/68339b4c9bc034267e1da611ab3b34f8.j
      pg',
       'http://books.toscrape.com/media/cache/92/27/92274a95b7c251fea59a2b8a78275ab4.j
      pg',
       'http://books.toscrape.com/media/cache/3d/54/3d54940e57e662c4dd1f3ff00c78cc64.j
      pg',
       'http://books.toscrape.com/media/cache/66/88/66883b91f6804b2323c8369331cb7dd1.j
      pg',
       'http://books.toscrape.com/media/cache/58/46/5846057e28022268153beff6d352b06c.j
      pg',
       'http://books.toscrape.com/media/cache/be/f4/bef44da28c98f905a3ebec0b87be8530.j
      pg',
       'http://books.toscrape.com/media/cache/10/48/1048f63d3b5061cd2f424d20b3f9b666.j
      pg',
       'http://books.toscrape.com/media/cache/5b/88/5b88c52633f53cacf162c15f4f823153.j
      pg',
       'http://books.toscrape.com/media/cache/94/b1/94b1b8b244bce9677c2f29ccc890d4d2.j
      pg',
       'http://books.toscrape.com/media/cache/81/c4/81c4a973364e17d01f217e1188253d5e.j
      pg',
       'http://books.toscrape.com/media/cache/54/60/54607fe8945897cdcced0044103b10b6.j
      pg',
       'http://books.toscrape.com/media/cache/55/33/553310a7162dfbc2c6d19a84da0df9e1.j
      pg',
       'http://books.toscrape.com/media/cache/09/a3/09a3aef48557576e1a85ba7efea8ecb7.j
      pg',
       'http://books.toscrape.com/media/cache/0b/bc/0bbcd0a6f4bcd81ccb1049a52736406e.j
      pg',
       'http://books.toscrape.com/media/cache/27/a5/27a53d0bb95bdd88288eaf66c9230d7e.j
      pg']
```

Part f

[33]:

```
df = pd.DataFrame({
    'Title': titles,
    'Price': prices,
    'Star Rating': ratings,
    'Cover JPEG URL': img_url
})

df
```

[33]:

```
                                                Title  Price Star Rating  \
0                                A Light in the Attic  51.77       Three
1                                  Tipping the Velvet  53.74         One
2                                          Soumission  50.10         One
3                                       Sharp Objects  47.82        Four
4               Sapiens: A Brief History of Humankind  54.23        Five
5                                     The Requiem Red  22.65         One
6     The Dirty Little Secrets of Getting Your Dream…  33.34        Four
7     The Coming Woman: A Novel Based on the Life of…  17.93       Three
8     The Boys in the Boat: Nine Americans and Their…  22.60        Four
9                                     The Black Maria  52.15         One
10     Starving Hearts (Triangular Trade Trilogy, #1)  13.99         Two
11                              Shakespeare's Sonnets  20.66        Four
12                                        Set Me Free  17.46        Five
13    Scott Pilgrim's Precious Little Life (Scott Pi…  52.29        Five
14                          Rip it Up and Start Again  35.02        Five
15    Our Band Could Be Your Life: Scenes from the A…  57.25       Three
16                                               Olio  23.88         One
17    Mesaerion: The Best Science Fiction Stories 18…  37.59         One
18                       Libertarianism for Beginners  51.33         Two
19                            It's Only the Himalayas  45.17         Two

                                       Cover JPEG URL
0   http://books.toscrape.com/media/cache/2c/da/2c…
1   http://books.toscrape.com/media/cache/26/0c/26…
2   http://books.toscrape.com/media/cache/3e/ef/3e…
3   http://books.toscrape.com/media/cache/32/51/32…
4   http://books.toscrape.com/media/cache/be/a5/be…
5   http://books.toscrape.com/media/cache/68/33/68…
6   http://books.toscrape.com/media/cache/92/27/92…
7   http://books.toscrape.com/media/cache/3d/54/3d…
8   http://books.toscrape.com/media/cache/66/88/66…
```

```
9    http://books.toscrape.com/media/cache/58/46/58…
10   http://books.toscrape.com/media/cache/be/f4/be…
11   http://books.toscrape.com/media/cache/10/48/10…
12   http://books.toscrape.com/media/cache/5b/88/5b…
13   http://books.toscrape.com/media/cache/94/b1/94…
14   http://books.toscrape.com/media/cache/81/c4/81…
15   http://books.toscrape.com/media/cache/54/60/54…
16   http://books.toscrape.com/media/cache/55/33/55…
17   http://books.toscrape.com/media/cache/09/a3/09…
18   http://books.toscrape.com/media/cache/0b/bc/0b…
19   http://books.toscrape.com/media/cache/27/a5/27…
```

Part g

[35]:

```
def books_page_scrape(url):
    r = requests.get('https://httpbin.org/user-agent')
    useragent = r.json()['user-agent']
    headers = {'User-Agent': useragent,
            'From': 'sp8me@virginia.edu'}

    r = requests.get(url, headers = headers)
    mysoup = BeautifulSoup(r.text, 'html.parser')

    titles = [h3.find('a')['title'] for h3 in mysoup.find_all('h3')]
    prices = [p.text.replace('Â£', '') for p in mysoup.find_all('p', class_ =␣
 ↪'price_color')]
    ratings = [p['class'][1] for p in mysoup.find_all('p', class_ =␣
 ↪'star-rating')]
    url_2 = 'http://books.toscrape.com/'
    img_url = [url_2 + img['src'].replace('../', '') for img in mysoup.
 ↪find_all('img')]

    df = pd.DataFrame({
    'Title': titles,
    'Price': prices,
    'Star Rating': ratings,
    'Cover JPEG URL': img_url
})

    return df
```

Part h

[37]:

```
new_df = pd.DataFrame()

for i in range (1,51):
    if i == 1:
        page_url = 'http://books.toscrape.com/catalogue/page-1.html'
    else:
        page_url = f'http://books.toscrape.com/catalogue/page-{i}.html'

    page_df = books_page_scrape(page_url)
    new_df = pd.concat([new_df,page_df], ignore_index = True)

print(new_df.shape)
```

```
(1000, 4)
```

This notebook was converted with convert.ploomber.io