Data Science

Lab. Sheet 5 – Data Scraping

Exercise 1
You are required to write a python script that will return a list of the titles of all books available in a given category on the oreilly bookstore site(http://shop.oreilly.com/).

The Getting Data presentation can help you construct the required scripts

The steps are as follows:

Step 1: You should first write a python script which allows you to store the html returned from a search for Data Science books based on a single page selection. You will use the 'get' function from the 'requests' library. The url for the second page of the search is below.

'https://ssearch.oreilly.com/?i=1;page=2;q=data+science&act=pg_2'

Step 2: Using the BeautifulSoup python library for pulling data out of HTML and XML files, you should identiy the paragraph elements with the correct 'class' lable and extract the contents of the anchor tags from these paragraphs. You will then need to clean this text data to get the title of the book.

Step 3: You should next iterate across all pages to extract the complete list of titles.

Step 4: Finally, you should allow for the user to input any search terms to get at the list of books.