ETL (Extract, Transform, Load) Project

Barnes & Nobel Benchmarking Analysis

Prashanth Saseenthar

Flora Ruan

# Background

The ETL Barnes & Nobel benchmarking analysis is for education purpose with the guidance from Boot Camp tutors. Big thanks to Rene Simon and Bryan Nonni.

After assessing the capability and the availability, we decided to merge the book information from Kaggle (origin: New York Times) with the actual pricing information from Barnes & Nobel (B&N) to implement benchmarking analysis for B&N.

# Extract

Source 1: The initial dataset was pulled from Kaggle. It was a JSON file that stored several queries from the New York Times Best Seller List API. The data contained information including the title, author, description of the book, publisher, the date the book was on the best-seller list, the book’s rank, how long it was on the list, and a URL for the book’s Amazon page.

Source 2: (check the video)



Parse the pricing information in the B&N website using the ISBN-10 digit from Source 1. B&N only show ISBN-13, but the book can be searched using ISBN-10, which is the key ID in the SQL database.

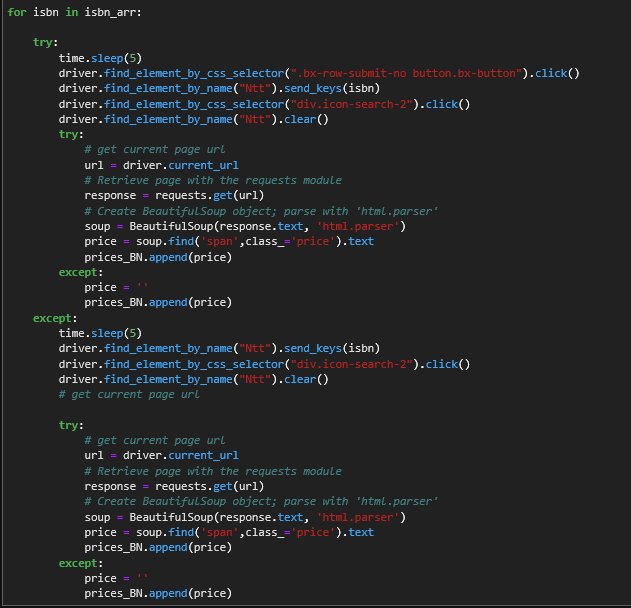
Step 1: Get access to the B&N website.



Step 2: Skip the advertisement by clicking “I will Pass.”

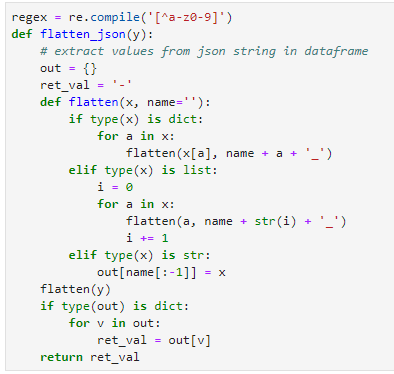
Step 3: Enter ISBN-10 digits to search bar.

Step 4: Search the price on the current page. Clear search bar and continue the next book.

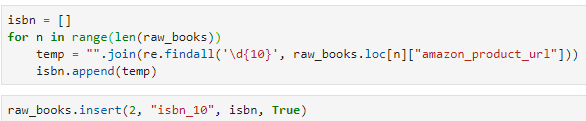


# Transform

Before the Barnes and Noble page could be scraped, the initial raw\_books data frame had to be cleaned. The data initially had extraneous characters for all entries, and the dates were represented in Unix timecodes instead of a standard year, month, and day. Using regular expressions and functions in python, the data was cleaned up to be much more readable.

Figure 1. json flattening functions.

After cleaning the data, the ISBN for each book had to be extracted. The original dataset did not include a separate ISBN column, but the ISBN was available in the Amazon product URL column. Using another regular expression, the ISBN was extracted from the product URL, before being added to the data frame as a new column.

Figure 2. regex for pulling ISBN from Amazon product URL.

# Load

After merging the B&N prices with the NY time table. Upload the table to SQL.