

CS 839 Project Stage 2 Report

Zeping Ren, Stephen Sheen, Shiqi Yang

Data Sources

We extracted paper-based book selling information from two online book-selling websites:

- Barnes & Noble: <https://www.barnesandnoble.com>
- Better World Books: <https://www.betterworldbooks.com>

Both websites contain the information about the books. For example, if we look up “Harry Potter and the Chamber of Secrets”, Barnes & Noble Website will show the following information on its web page:

Overview	Product Details	About the Author	Read an Excerpt	More >
Product Details				
ISBN-13:	9780439064873			
Publisher:	Scholastic, Inc.			
Publication date:	08/28/2000			
Series:	Harry Potter			
Edition description:	Reprint			
Pages:	352			
Sales rank:	441			
Product dimensions:	5.20(w) x 7.50(h) x 1.10(d)			
Lexile:	940L (what's this?)			
Age Range:	9 - 12 Years			

And Better World Books shows the following information:

About the Book Find at your local library			
Format	Paperback Book, 341 pages	Language	English
Publisher	Scholastic (Sep. 1st, 2000)	Edition	Unknown
ISBN-13	9780439064873	Dimensions	5.26 x 7.62 x 0.90 inches
ISBN-10	0439064872	Shipping Weight	0.55 lbs
Categories	Children's Fantasy & Children's Magic Books		

Data Extraction Techniques

As mentioned above, both websites display the information with a structured HTML element. Specifically, they use tables to display the details. So we can make use of this to extract the information we need.

Given a web page (as an url), our web crawler converts the web page into an HTML string. If the web page is that of a book, it then extracts the title, the author(s), and the table containing the details of the book. It also extracts any links to other book web pages and adds those into a list of links to explore. Our web crawler repeatedly does this in a depth-first search fashion.

Entity Type

We collect book details information from both websites and obtained the two tables with the following schema:

```
A(title,author,isbn,publisher,publication_date,series,edition,pages,
  sales_rank,dimensions,lexile,age_range)
B(title,author,form,pages,language,publisher,edition,isbn13,dimension,
  isbn10,weight,category)
```

Then we take the intersection of the schema and have the common schema of the following form:

```
A, B(title,author,isbn,publisher,edition,pages,dimension)
```

Our data contains 3600 tuples per table.

Open Source Tools

We used the Python built-in libraries (`urllib`, `html.parser`) to write our web crawler. These two libraries basically provide essential HTML source parsing functionalities.