Project Stage 2 – Rule based wrapper construction

Our two web sources are:

- https://www.goodreads.com/
  - Goodreads is a "social cataloging" website that allows individuals to freely search its database of books, annotations, and reviews.
- https://www.amazon.com
  - Amazon is an e-commerce company which started as an online bookstore.

Thus, these two repositories proved excellent to extract structured data.

First we went through the HTML DOM tree of the webpage. We found out particular s, s, etc. and their associated classes which were being used to display the book title, author name, rating and price. Once we had identified this structure, we started crawling the pages and used BeautifulSoup to obtain the DOM tree of the page. Each attribute had different classes in their associated html tag. Then, we just had to extract the text/value of the particular attribute from the page and store it in csv format.

While extracting the attributes, we were faced with few exceptions despite having understood the grammar of the page. For example, certain books in amazon.com did not conform to observed pattern. Hence, we skipped those records during extraction.

The entity we extracted are books. The attributes are:

1. Title - (name of the book)
2. Author
3. Rating
4. Format - (eg., paperback, hardcover, etc.)

Each table contains 3000 tuples.

Open source tools used:

- Bs4 (Beautiful Soup) - Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.