Data Science - Project Stage 2

Web Data Extraction

Team Members:

- Karan Dharni (dharni@wisc.edu)
- Sudarshan Avish Maru (smaru@wisc.edu)
- Amogh Joshi (asjoshi4@wisc.edu)

Web Data Sources:

The two sources of data we selected for our data extraction task are:

- www.amazon.com
- www.booksamillion.com

We have extracted structured data about books from the following three categories: Space Opera, Dark Fantasy and Psychological Thrillers. Moreover, we have extracted data about new paperback books only.

Extraction Methodology:

The process of extracting the data from both the sources was identical.

Step 1: On applying appropriate filters on the data-source websites, a collection of multiple pages is loaded; and each page contains entries for multiple books. We grab the embedded URLs of the books on each page, store them in a file and then move to the next page. Thus, at the end of this step, we have two files, which contain the URLs of the books.

Step 2: In this step, we read the URLs from the files created in step 1, and crawl each of these URLs individually. We extract various attributes for each book. A tuple consisting of these attributes is then stored in a CSV file.

After successful completion of both steps, we obtain two CSV files, one for each data source, containing the extracted data in tabular form. Each tuple of a table describes the data about one particular book.

Entity Description:

The entity type we have extracted from the above mentioned sources is **Books**. We have created two tables (CSV format), one for each source. These 2 CSV files are named books_amazon_output.csv and books_millions_output.csv.

The schema followed by both the tables is:

{Name, Category, Author, Price, Series, Pages, Publisher, Date, Language, ISBN-10, ISBN-13, Dimension, Weight}

Here's a brief description of the attributes extracted:
Name: Title of the book
Category: Category of the book (We have extracted data for books limited to categories Space Opera, Dark Fantasy and Psychological Thrillers)
Author: Author(s) of the book
Price: Cost of the book
Series: Series of the book
Pages: Number of pages in the book
Publisher: Publication house of the book
Date: Date on which this book was (or will be) released
Language: Language in which this book is written
ISBN-10 and ISBN-13: The 2 ISBNs of a book
Dimensions: Size of the book
Weight: Shipping weight of the book

Some attributes are sparsely populated. For example, in books_millions_output.csv, Dimensions and Weight are sparsely populated, as compared to books_amazon_output.csv.

The number of tuples in the table extracted from **amazon.com :- 2990**
The number of tuples in the table extracted from **booksamillion.com :- 9420**

Description of Tools Used:

We have used '*Beautiful Soup*' and '*Selenium*' tools of the python framework in our project. Selenium is used to get the HTML page source; whereas Beautiful Soup is used to extract the attributes from this source.