# CS 839 Data Science Project Stage 2 Report

Sreejita Dutta, Deepanshu Gera, Rahul Jayan

## I. Web Data Sources

We extracted structured information about books sold by two websites amazon.com and Walmart.com. Both are ecommerce websites involved in the sales of books, consumer electronics, apparels, food and so on.

## II. Approach

We first select a search page for books of a particular genre and then extract the product id for each book. Each product id is unique and used to keep track of products in any e-commerce website's internal database. Almost 3000 product ids per website is extracted and then we construct an url using this product id to visit the relevant page and scrap the required information.

We used selenium package to automate page loading and then extract each attribute using the XPath of the element in the DOM tree of the html page.

For eg. Consider scenario of scraping mystery novels from amazon.com

We start from the search page of mystery novels.
https://www.amazon.com/s/ref=sr_pg_2?rh=n%3A283155%2Cn%3A%211000%2Cn%3A18&page=2

We start scraping from page 2 to reduce the number of dynamic elements in the page. The product id for each book is extracted from here and stored in a file.

Let's say we want to extract information for the book 'Lord of the Flies'. We extract the product id (0399501487) for this after a lot of inspection. From this product id, the product page is constructed as

https://www.amazon.com/gp/product/0399501487

Each attribute can be then extracted using the XPath of the element in the DOM tree. XPath we used to extract name of the book and sale price is given below.

Name: //h1[@id="title"]//text()

Sale Price: //span[contains(@class,"offer-price") and contains(@class,"a-size-medium")]//text()

The same methodology can be extended to scrap information from walmart as well.

## III. Data Description

We have chosen books as the entity. The scrapped table consists of following attributes

1. Name/Title of the book
2. Sale Price
3. Category/Genre
4. Author
5. ISBN
6. Pages
7. Publisher
8. Language
9. Dimension of the book
10. Weight of the book
11. User Rating

Though we have retained the attribute ISBN in the dataset, we would be suppressing that for the entity matching stage.

## IV. Tools

We have mainly used three python packages in our code - Selenium, requests and lxml

Selenium is a package which is used for automating the tests carried out on web browsers. We have used selenium for automated page loading and then to extract structured data from this page.

Requests package allows us to send HTTP requests and access response using python. We can add content like headers, form data, multipart files and parameters via simple python libraries using this package.

Lxml is a high-performance library used for processing XML and HTML pages in python. We use lxml to extract information from an HTML page using XPath.