CS 839: Data Science

Project Stage One

Ankit Jain(ajain64) | Ankit Maharia(maharia) | Prerak Mall(pmall)

In this stage of the project we decided to extract books data from the following two online sources:

1. Amazon
2. Barnes & Noble

We divided the problem statement into following phase of Data Science Pipeline:

1. Data Acquisition
    a. To ensure the overlap among the generated tables we considered the books from the categories Biography, Business & Money, Religion & Spirituality on Amazon and Biography-Teens, Business-Teen, Religion-Teen on Barnes & Noble. Three categories were selected because amazon does not allow to go beyond page 100 for one search category.
    b. In this phase we had to solve two problems. One, get the products on a page. Two, move to the next page.
    c. We initially did it using the urllib library which comes with python core. This was able to help us data from Barnes & Noble with ease, but we ended up with our script being detected as a bot.
    d. To handle the issue of bot detection, we used headless browser and simulated user clicks using headless browser. This was done via the use of Selenium.
2. Construction of Wrapper (Ew)
    a. To do this we first analyzed pages from both the sources and came up with Tw(Target Schema).
    b. After the analysis of html page structure from each source. We used Beautifulsoup to develop to parsers for the same. Beautifulsoup was helpful in building the DOM tree and accessing various nodes of DOM Tree and their attributes.

Output Characteristics:

1. We have extracted books data from two online sources.
2. Output contains two csv files, one for each source. Source one is amazon and Source two is Barnes & Nobel.
3. Each table contains details of books having following schema:
    - Amazon:
      [Title, Author, Type, Pages, Publisher, Language, ISBN_10, ISBN_13, Age_Range, Grade_Level, Product_Dimension, Shipping_Weight, Price]

    - Barnes & Nobel:
      [ISBN-13, Publisher, Publication date, Edition description, Pages, Sales rank, Product dimensions, Age Range, Title, Author, Price, Average Review, Lexile, Series, Sold by, Format]

    Majority of attributes in both the files are same, but since both sources are different each provided some additional information as evident from the above schemas. If same schema is a strict requirement we can always add columns with empty values.

4. Both the csv files source1.csv (Amazon) and source2.csv (Barnes & Nobel) contains more than 3000 tuples.

5. In case of missing values, we have populated that attribute with empty string and to avoid commas in the text we have replaced them with “#”.

Open Source Tools:

In this stage we have used beautiful soup tool that is used to parse the HTML pages and extract desired attribute. Also, to get HTML data we used Selenium to simulate user clicks and get each book page.