Luxin Wang, Yudi Wang, Hayk Hovhannisyan, Hung-Ting Chen
GitHub: https://github.com/LuxinWang00/ECE143-group-9/projects/1

Best Sellers in Books

Problem:

·What are the trends in best selling books?
·What drives the price of book?
·How are discounts given by seller?

Background and Real World Applications:

According to a report on June 29, 2018, *Leisure reading in the U.S. is at an all-time low*, the share of Americans who read for pleasure on a given day has fallen by more than 30 percent since 2004. Thus, it is interesting to focus on the best sellers books. From the list page of the website we can collect major information (name, rating, price, number of customer reviews, format) of 100 books per year; besides, from the detailed page for each book, we can figure out other possible meaningful information such as category, length of recommendation, total pages, publishing year and language. By analyzing and plotting them, we can deduce different aspects such as: trends of hottest category changes in the past 20 years, relationship between review numbers and the sales, the way rating/price affects sales, etc. Combining these results should be sufficient to answer the aforementioned problem: What is the most important factor in the book selling nowadays?

Since the popularity in reading books is at all time low we will attempt to provide data to give guidance to the publishing industry, to tell them what is the most dominant factors for today's readers, in hopes it will be used in their selling strategy. At the same time, this project will also summarize the changes of the hottest category that attract reader's attention which the publishing company should focus on in order to do a better publishing plan for the coming year.

Draft of Proposed Solution:

We are going to do this by scratching data from a Singapore books selling website (https://opentrolley.com.sg/Home.aspx), to find best sellers in different categories. The data in the website has been updated this February. From the amazon web page we are going mainly focus on these aspects: name, author, price, original price, pages, publisher, publish type, publish date, available numbers, detail_url. Our intention is to discover possible correlations between these categories, and find their relation to do mentioned questions in the problem section above. Lastly we will utilize python graphical tools to reveal these possible underlying connections among them.

Project Steps:

1. Extracting and Cleaning up data:
    (a). Data extracting: To capture data from Opentrolley website, we decide to use the request package and the XPath method to scraping necessary information from the website HTML document.
    (b). Data cleaning: The raw data we scraped directly from the web page need to do some preprocess to improve the quality. This process includes: drop the unnecessary information; deal with the missing data; transform the price string into numbers; use publish year instead of the exact day; add the category tag.
2. Data Visualization: Using what offered in the course webpage (Data Visualization Using Python Part) and also other guidance material from the internet to do the data visualization process.
3. Analyze and Conclude Data: Based on the graph to draw conclusions and answer the problem we proposed.

Dataset:

Scratched from: https://opentrolley.com.sg/Home.aspx

Data Scraping:

Justin: Literary collection, Business and Economics, Law, Religion
Yudi: Fiction - 164 in total, Medical, Philosophy, Psychology, Bible
Luxin: Cooking, Science, Comic and Graphic novel, Drama
Hayk: History, Literary collections, technology and education