Designing Book Recommendation System by User and Book Information

By Farzad Radmehr

What is the problem you want to solve?

Picking the best book to read is important in order to not spending long time to figure out if the book is right choice or not. This gets more critical if we spend same amount of time to read each book before finding the right one, specially, considering different genres, Publishers, Book Authors or Year of Publications. We can use this information and reader age and location to recommend the best book to read by user. This can be done by using the ratings that have been given to previous books by readers. This recommendation system can help the reader to pick the best books and can generate the better results over time, when reader(s) give more rates to the book(s).

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

All online book ordering websites or book stores or libraries are the clients. For example, online websites like Amazon (Amazon Kindle), EBay or Barnes and Noble, etc. If this ranking systems be linked to this service providers, they can offer the books to customers. Recommending better options will increase the sale and customer satisfactions and positive reviews. For example, you order a book in Amazon and give the rate and in next order, the website can help you to find the best option based on your previous choices.

What data are you using? How will you acquire the data?

This data was collected from Book_Crossing community with permission from Ron Hornbaker, CTO of Humankind Systems. This data contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books. It is provided both in CSV and SQL format. These files can be downloaded here. The tables and the fields are as below:

Users: User IDs, Location, Age

Books: ISBN, Book title, Book Author, Year of Publication, Publisher, Image_url_S, Image_url_M, Image_url_L

Rating: User IDs, ISBN, Book Rating

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

Part of data wrangling process will be to decide the missing values or duplicate records, and also checking the data type in each columns. In the beginning, I realized that after importing the data, all values are entered in the first column. So I should split them, remove unnecessary characters and fixing the column names. I also study the outliers or other data inconsistency. We also generate some plots to study the values of each column in data cleaning stage. In second part, we study them by more visualization techniques.
In machine learning part, I use Content Based Filtering, Collaborative Based Filtering and Hybrid Based Filtering. I will work on Hybrid Based Filtering to propose the best approach in this stage. And last, but not the least, I will try to develop an app to recommend the best books based on the inputs.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

My deliverables will be a jupyter notebook and slide deck published to my github account. The note will include a report of my findings and related python code.