Springboard--DSC
Capstone Project 2
Project Proposal
Cassie Kluse
August 2020

Problem Statement

Given the user reading history and book ratings, what books will the user most likely want to read next?

Context

To figure out what books users are most likely to enjoy next, I am creating a book recommendation system using Goodreads data. A book recommendation system benefits many people.The world is filled with millions of books, which makes the task of finding the next read daunting. A book recommendation system can help users find their next great read quickly.

Furthermore, a recommendation system can also benefit bookstores and booksellers as it can help them find titles to recommend based on popularity and customer habits. These recommendations can be used when creating displays or company-wide promotions. The system can also be integrated into a bookstore's website, recommending great reads based on what they have bought before and how they rated past purchases. Good recommendations help increase sales and customer loyalty both in-store and online.

Criteria for Success

The recommendation system will be a success if it can accurately recommend titles that the user has not read before and seems to be a good read based on that user's rating/reading history as well as the reading history of similar users.

In order to get the desired results, the best plan of action is to make the system using a User-based Collaborative Filtering Technique.

Scope of Solution Space

I originally planned on using all the data available from my source (more info in the 'Data Source' section below), but because of my computer's limited space and processing power, I decided to focus on one genre :fantasy/sci fi. However, there is still sufficient data to work with as this includes over 3,000,000 reviews for about 250,000 books.

Constraints

One constraint is that the data is not entirely up to date (last updated in May 2019). However, the only way that this should affect my project is that the system won't be able to recommend books that have been published in the past year. Besides that, I don't anticipate any other constraints. In theory, the system should give accurate recommendations if newer books are added to the dataset.

Stakeholders

Potential stakeholders for this type of project include the marketing and sales teams. The marketing team would benefit from this project as it could guide them on decisions regarding displays and company-wide promotions. They could also collaborate with the website developers to make sure customers are shown recommendations while purchasing books from the website. The sales team could provide sales data and trends to improve the system. They would also benefit from this project as they could find ways to use the recommendations to boost sales (such as giving guidance to individual booksellers, who make a significant contribution to instore sales).

Data Source

Originally, I was planning on using data from Goodreads' API. However, the API provided very limited functions that did not meet the data needs for this project. For example, almost all the functions only allow users to get information on one book at a time. There was no easy way to access large amounts of data from the API, and I need information on thousands of books to create a powerful and accurate recommendation system. Therefore, I am going to use data collected by UCSD[1]. This data originated from Goodreads and is quite comprehensive. The datasets include fifteen million reviews for over two million books. However, as mentioned earlier, I will just be using

[1] https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home

data on fantasy/sci fi books to build the system. The system should work well on other genres, though.

Deliverables

All deliverables will be uploaded to my Capstone 2 repository on Github. My deliverables will include separate Jupyter notebooks for each step, as well as a final slide deck and project report.