Movie 2 Book Recommender System

By John McGowan

Throughout my life, I have always intended to read more books, but taking recommendations based on other books I have not yet read proves ineffective due to my lack of reading history. However, I *have* watched an extensive number of films via theatres and Netflix, so I wonder: could the taste I have established in films transfer over to help determine which books I might like? An effective book recommender based on movie tastes could prove quite lucrative to the book publishing industry, drawing on peer recommendations and user’s own personal input to drive book sales and higher user engagement with the more active book reading community.

A movie-to-book recommendation engine seeks to deliver a way for less active readers to breach the reading barrier, using both content-based filtering and item-to-item collaborative filtering to find book titles agreeing with film taste. Using the Movielens 20M dataset and Goodreads (Goodbooks 10k) book review dataset, I will build a similarity model that first focuses on genres and keywords NLP of movies/books to recommend titles similar to the user’s choices, then focuses on users that liked similar content to recommend more peer-approved choices. For instance, a user picks five action movies they like, which produces a recommendation for a Tom Clancy novel, which in turn produces a small group of other novels that action readers enjoy similar to the Tom Clancy novel. To keep data sources updated, I will build a web scraper that scrapes the same information available in the dataset using the Goodreads API for new titles and more reviews to retrain the model. The same process will be necessary for new movies using a web scraper on Movielens’ website.

To build the engine, SVD and Cosine similarity will be used to produce similarity scores between titles. Supervised techniques (most likely K Nearest Neighbors) will be used to predict a user’s score of an unread book by considering the most similar movies they have previously rated and weighting their respective similarity to the new item. After predicting their ratings for new items, only the highest scoring novels will be recommended to the user. A variety of metrics will be tested to select the strongest performer, and RMSE will be used to calculate the difference between predicted and actual scores when data is provided to ensure the predictor is accurate.

I anticipate the largest challenges stemming from the project are building the MovieLens web scraper, as they do not have a specific API, and tuning the item-to-item collaborative filter. Since the users are not shared between datasets, great care must be taken to match up genres and keywords which may vary across media platforms. It will be also challenging finding matches of similar items given that most users have rated less than 30 of the 10,000+ available films, but I plan on utilizing matrix factorization to combat this sparsity issue.