Project Update 0

Anne Marijn, Tereza, Clarice & Andrea

Title

Genre-Aligned Book Recommendation System: A Comparative Analysis with Goodreads

Abstract

A max 150-word description of the project question or idea, goals, dataset used. What story would you like to tell and why? What's the motivation behind your project?

This project aims to compare the genre relevance of recommendations generated by three different recommender systems (description-based, rating-based, and combined) with genres of recommendations from Goodreads. Our motivation lies in producing a recommendation system that aligns in terms of genre with Goodreads' system, a widely used platform for book recommendations. Using the extensive Goodreads book dataset from Kaggle, containing over 10 million entries, we will preprocess the data and develop three distinct models. By evaluating the effectiveness of each model in providing personalized book recommendations aligned with user genre preferences, we aim to identify which model performs best in this task. This comparison will shed light on which feature (description or ratings) contributes more significantly to the quality of recommendations, thereby providing insights for optimizing recommendation systems to improve genre alignment.

Research questions

A list of research questions you would like to address during the project.

Is there a significant difference in genre relevance of recommendations generated by our three recommender systems (description-based, rating-based, and combined) compared to genres of recommendations from Goodreads?

Dataset

https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m/data

- The dataset contains over ten million books scraped from the book review website Goodreads.
- Each book entry contains information such as: Id, Name, Nr of pages, Publisher, Language, Author, Rating, ISBN, Count of text reviews, Description
    - Example: 5107, The Catcher in the Rye, 1:133165, 277, 4:808278, total:2610840, 30, 1, Back Bay Books, 44046, 2001, eng, J.D. Salinger, 3.8, 2:224884, 5:891037, 0316769177, 3:553476, 55539, "The hero-narrator of The Catcher in the Rye is an ancient child of sixteen, a native New Yorker named Holden Caulfield. Through circumstances that tend to preclude adult, secondhand description, he leaves his prep school in Pennsylvania and goes underground in New York City for three days. "
- There's 30 files in the dataset. 7 of these contain user ratings, and the remaining 21 files all contain information on books. All files are .csv files.

A tentative list of milestones for the project

Add here a sketch of your planning for the coming weeks. Please mention who does what.

TO DO:

Task	**Person**
Stage 1: Preprocessing	
Preprocessing (loading the data, making any necessary adjustments etc.)	Clarice
Stage 2: Building the RSs	
Model 1: Calculation of similarity scores for "book descriptions" feature	Andrea & Anne Marijn
Model 2: Calculation of similarity scores for "book rating" feature + implementing number of "written reviews" feature as weight to the vectors	Tereza

Model 3: Combination of scores for “book description” and “book rating”	Clarice
<u>Stage 3: Comparison</u>	
Our RS input: book title Our RS <u>output</u>: 3 times top 3 recommended book titles (1 for each model) **To-Do:** ***From our RS, we collect:*** ● The output’s 3 title ● <u>*Manually*</u> ⇒ The output’s books’ 3 corresponding genres as provided by Goodreads ***From Goodreads’ RS, we collect:*** ● <u>*Manually*</u> ⇒ The input title’s 3 corresponding genres as provided by Goodreads ● <u>*Manually*</u> ⇒ The 3 corresponding genres of the top 3 recommended books provided by Goodreads	Clarice, Andrea, Anne Marijn, Tereza

We will add deadlines after having received feedback on this provisional division