CSE 158: Web Mining and Recommender Systems November 24, 2020

Assignment 1 - Predictions from Steam Game Reviews

Cameron Thomas
Kaggle Username: cat028

1. Task - Play Prediction

Given a (user,game) pair predict whether the user would play the game.

1.1. Approach

I assume that users with similar tastes will play similar games and that users are more likely to play a game if it is considered popular. My initial approach was to build a model based off the baseline in Homework 3 Question 4.

1.2. Feature Design

I adjusted the popularity and Jaccard features and added a cosine feature. Previously, popularity was defined using a threshold of the Nth percentile of popular games. Since the best threshold is determined from the test set which is randomly sampled, the model may not perform well on unseen data. In my new model, I define popularity as the number of users who play that game normalized by total played. Instead of a filter, popularity is now a weight. I modified the Jaccard feature to return the average similarity instead of the maximum similarity. If one game in a list has high similarity and all other similarities are zero, then this game would receive a mistakenly high overall similarity which would affect the accuracy of the prediction. Taking the average Jaccard similarity accounts for this edge case. I added a cosine similarity feature to further boost model performance. Then I ranked the scores ((Avg. Jaccard Sim * Max Cosine Sim) / Popularity) in a new dictionary to make the predictions.

1.3. Things That Did Not Work

I tried removing certain features but the multiplication above produced the highest accuracy. I also tried adding another Jaccard feature that found the similarity of games but it failed to improve model accuracy. If I had more time I would attempt a Pearson similarity feature and a user activity weight.

1.4. Results

Model Accuracy: 0.74085
Score: 0.73870
Leaderboard rank: 22/672

2. Task - Category Prediction

Predict the category (Action, Strategy, RPG, Adventure, or Sport) of a game from a review.

2.1. Approach and Implementation

Instead of a bag-of-words, I used a TF-IDF vectorizer from Sklearn for logistic regression. I tried using various C parameters, an SVM, and a random forest classifier. The SVM with C=1 produced the highest accuracy.

2.2. Results

Accuracy: 0.7587
Score: 0.77200
Leaderboard rank: 13/401