Netflix Movie Rating Prediction

Chi-Jen Chien
North Carolina State University
Raleigh, North Carolina, USA
cchien@ncsu.edu

Danielle Hancock
North Carolina State University
Raleigh, North Carolina, USA
dnhancoc@ncsu.edu

Eddy Huang
North Carolina State University
Raleigh, North Carolina, USA
whuang25@ncsu.edu

ACM Reference Format:
Chi-Jen Chien, Danielle Hancock, and Eddy Huang. 2021. Netflix Movie Rating Prediction. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DATA SET

We plan to combine two data sets from Kaggle: The Netflix Prize Data [5], which contains user ratings of various movies, and a dataset taken from IMDb [4], which provides additional information about the movies, such as genre, director, actors, etc.

2 PROJECT IDEA

We want to answer the question: Which aspects of a movie (such as genre, director, year released) contribute the most to users' ratings? Additionally, we plan to build a classifier that predicts a user's rating of a movie based on their previous ratings and data about those movies. A great deal of previous work has been done with the Netflix Prize Data, which contains only user rating data and movie titles/years. Therefore, these approaches were by necessity based on *user* (Collaborative Filtering Algorithm) similarity rather than *movie* similarity, as the data did not include any information about the movies beyond their title and release year. Our approach will be novel because it combines two datasets and focuses on similarity between the movies themselves, rather than the users who rated them.

3 SOFTWARE

We plan to implement two classifiers (probably KNN and Decision tree) using the in-built libraries of sklearn and TensorFlow or PyTorch. We will also need to write a program to clean and preprocess the data.

4 RELEVANT PAPERS

Papers can be found in the References section. [1] [2] [3]

5 WORK DIVISION

Data cleaning and preprocessing (Eddy)
KNN classifier (Danielle)
Decision tree classifier (Chi-Jen)

6 MIDTERM MILESTONE

We plan to have finished the data cleaning and preprocessing, as well as at least the KNN classifier.

REFERENCES

[1] Robert M. Bell and Yehuda Koren. 2007. Lessons from the Netflix Prize Challenge. 9, 2 (2007). https://doi-org.prox.lib.ncsu.edu/10.1145/1345448.1345465
[2] Robert M. Bell, Yehuda Koren, and Chris Volinsky. 2010. All Together Now: A Perspective on the Netflix Prize. *CHANCE* 23, 1 (2010), 24–29. https://doi.org/10.1080/09332480.2010.10739787 arXiv:https://doi.org/10.1080/09332480.2010.10739787
[3] Ted Hong and Dimitris Tsamis. 2006. Use of KNN for the Netflix Prize. (2006).
[4] Stefano Leone. 2020. IMDb movies extensive dataset. https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
[5] Netflix. 2019. Netflix Prize data. https://www.kaggle.com/netflix-inc/netflix-prize-data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn