Movie Recommender

By Michael S. Bull

# Introduction

The Internet is filled with reviews on almost everything imaginable, including products, foods, music, services, and movies, to name a few. These reviews are both helpful to the consumer and producer. For the consumer, the reviews are a way they can provide feedback about their experience with the service or product. This information is then used by other consumers who are looking for an opinion on a particular item before they purchase it. The producer’s use the reviews to see if changes are needed based on customer satisfaction. The producers also use the reviews to recommend other items to users based on various criteria.

Netflix saw the importance of these reviews to their business, such that in October 2006, they offered the Netflix Prize Competition[1]. Bell and Koren summarize in their paper [1] some of the strategies used to win the first Netflix progress prize. They state that the goal of the competition was to decrease the root mean square error of their existing recommender system by ten percent. The first person or team to accomplish this goal would win $1,000,000. This competition shows how vital these reviews are to a company.

# Background

Natural Language Processing (NLP) is widely used to take unstructured reviews and extract meaningful data from them for further analysis and classification into various categories like opinions and sentiments. Leung, Chan, and Chung [2] describe a process of collaborative filtering (CF) that moves beyond just sentiment analysis into a sentimental orientation (SO). This process goes beyond determining a positive, neutral, or negative review based on the context of the data and tries to infer a more granular scale.

In the paper by Khairnar and Kinikar [3], their approach was to mining the data from the reviews used sentiment analysis, opinion words, and feature-based identification. One of the features they used was parts of speech (POS) tagging and text summarization.

Qu, Ifrim, and Weikum took a different approach in their paper on review rating predictions [4]. Their method used a “Bag-of-Opinions.” Bigrams and n-grams made up these opinions, thus allowing them to gather a better context of the words than if they used unigrams.

# Project Description

This project will focus on configuring a movie recommender that will suggest movies for existing customers to watch. The recommender will group users based on similar ratings given to the same films. Once the users are grouped into clusters, and knowing all users in any given cluster have not watched or rated all of the movies, the recommender will determine the next best movie to watch. For example, users “A,” “B,” and “C” all have watched four movie from a set of seven and rated them all similarily. If users “A” and “B” both watched movie three, but user “C” has not, then the recommender could suggest movie three to user “C” to watch next.

# Project Implementation

The recommender will take as input the review dataset provided by IMDb [5]. Data preparation will be in the form of feature selection, such as the use of opinion-lexicon, sentiment analysis, or other data mining techniques or the use of tools such as WordNet and Word2Vec.

An unsupervised learning model will be used to train the classifier, and the project will use the K-Means method to perform the clustering. Once the optimal number of clusters is determined based on the categories of movies, then the users will be grouped into their respective clusters based on their ratings to film in the chosen categories. Due to the limitation of plotting the data in a friendly easy to read format, the dataset will likely be reduced to the top seven or eight categories of movies and the users who have rated the most movies in those categories.

The output will be displayed in a 2D matrix format. Each user (row) and each movie (column) will contain the user’s “perceived rating.” Or in the case the user has not rated that movie, it will include a no rating symbol.

To verify the results, I will use the same K-Means method and clustering groups as used with the training data, except I will use the actual user ratings as the prepared data. By comparing select movie recommendations from both results for a given user, this will show the accuracy and trustworthiness of the preparation and classification of the training data.

References

[1] Bell, Robert M., and Yehuda Koren. “Lessons from the Netflix prize challenge.” *Acm Sigkdd Explorations Newsletter* 9.2 (2007): 75-79. Available Online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.1489&rep=rep1&type=pdf>

[2] Leung, Cane WK, Stephen CF Chan, and Fu-lai Chung. “Integrating collaborative filtering and sentiment analysis: A rating inference approach.” *Proceedings of the ECAI 2006 workshop on recommender systems*. 2006. Available Online: <https://s3.amazonaws.com/academia.edu.documents/30843243/10.1.1.69.1870.pdf?response-content-disposition=inline%3B%20filename%3DLearning_and_adaptivity_in_interactive_r.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200307%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200307T053441Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=18e8fae7daee4fde98df5671b3605617fa4bb826f6a8261a1fd5ab07904400b7#page=70>

[3] Khairnar, Jayashri, and Mayura Kinikar. “Latent semantic analysis method used for mobile rating and review summarization.” *Int. J. Comput. Sci. Telecommun* 4.6 (2013): 61-67. Available Online: <https://pdfs.semanticscholar.org/2ad3/3749f42795c4a85bbfe110f892668bb7a7ab.pdf?_ga=2.85542553.1629128933.1583559208-878685597.1580782731>

[4] Qu, Lizhen, Georgiana Ifrim, and Gerhard Weikum. “The bag-of-opinions method for review rating prediction from sparse text patterns.” *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010. Available Online: <https://dl.acm.org/doi/10.5555/1873781.1873884>

[5] IMDB dataset. Available Online: <https://www.kaggle.com/utathya/imdb-review-dataset>