Data Madness

Jeroen Vermazeren, i6116870
Xavier Weber, i6094733

Department of Data Science and Knowledge Engineering
Maastricht University

March 27, 2018

1 Abstract

This project was conducted on the MOVIELENS[1] dataset for Education and Development. The used files were a dataset containing movies, and a dataset containing ratings of those movies with userID's. The dataset 'movies' had to be reformatted to make sure that the column genres was spread over multiple columns where every genre has its own column. The research questions are:

1. How to cluster similar movies according to their genre?

2. How to determine which group of users would like a movie?

Principal Component Analysis has been performed on the resulting dataset for two reasons. First to make K-means possible, because the original data contained binary values. Second to reduce dimensions. Then k-means was performed, with k optimised at 7 by using the elbow method. The resulting data contained 7 clusters with similar movies. Then per user, their mean rating for each cluster was computed. If the mean rating for a user is higher than 4.5 out of 5, it is an indication that the user really likes movies of that cluster and therefore it is collected in an "User Group".

For potential movies it is possible to check which "User Group" would like that movie. This is highly useful for a company like Netflix which can then analyse if a potential movie will be liked by which users and how many users. According to those statistics Netflix can decide to buy the movie.

[1] https://grouplens.org/datasets/movielens/