DATA VISUALIZATION TOOL FOR MOVIE RECOMMENDATIONS

Ritvik Bhagawatula
GT

Ashwin Dubey
GT

Ranveer Thind
GT

Neal Bayya
GT

Nishant Thangada
GT

Sravan Jayanthi
GT

Summary

Our aim is to create a movie recommendation tool with visualizations of movie similarities. With the visualization tool, we will show the relative similarities of the movie content per genre. It is important because our approach deviates from collaborative filtering and focuses on content-based similarity of movies using data points such as popularity and budgets. This allows users to maintain privacy over what movies they have watched and provides less biased results.

Approaches

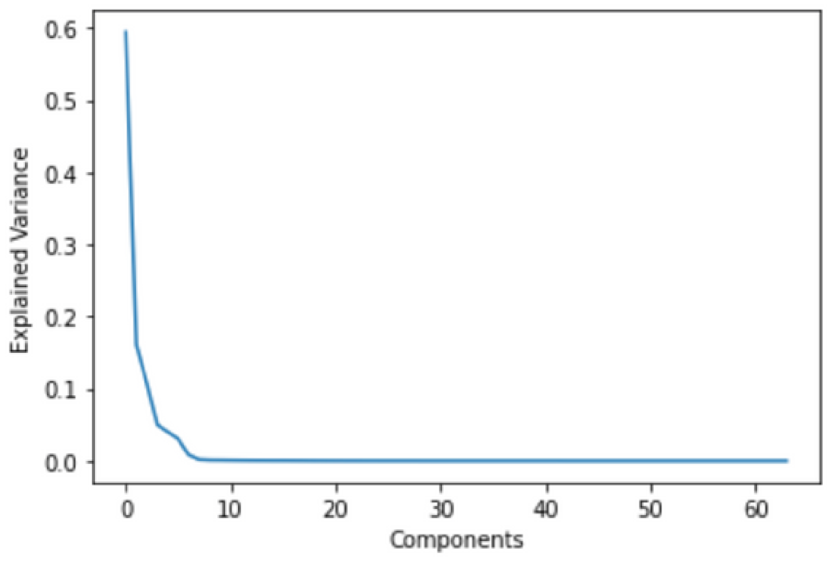
Our approach consists of performing K-Nearest-Neighbors and K-Means algorithms on the TMDB movie dataset. We also ran PCA for dimensionality reduction in order to clean the data. The K-NN algorithm will use the PCA cleaned data to compute “distances” between different movies and the K-Means algorithm generates different clusters based on data where each movie will be a part of a cluster. These two algorithms provide information about similar movies and the visualization uses those predictions to build a display. To solve our problem, K-NN would filter through the features to conduct predictive classification on individual data points and K-means would guarantee convergence into a cluster for a group of movies based on genre. This approach is new in that it will recommend movies on the basis of movie contents and attributes rather than collecting information from users, therefore, we are operating on less assumptions than existing recommendation methods.

Datasets

The data that we got was the TMDB 5000 Movie Dataset which was downloaded from Kaggle; It contains two CSV files. The first CSV, movies, contains information about each movie such as budget, genre, keywords, etc. There are a total of 20 unique columns and 5000 movies in the dataset. The second csv, credits, has information about credits, cast, and crew members. The Movies CSV Size is 5.7MB, Credits CSV Size is 40MB, and the dataset is not temporal since movies span different time periods.

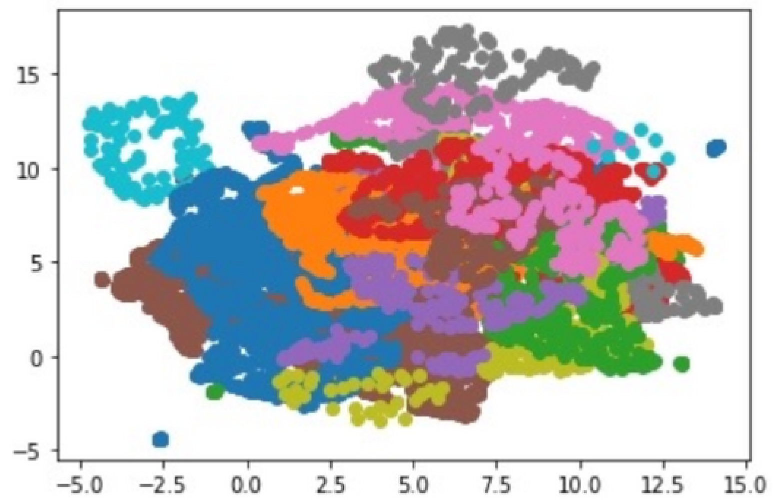
Experiments/Results

We first evaluated our approach by implementing PCA on our preprocessed movies dataset using 64 features.



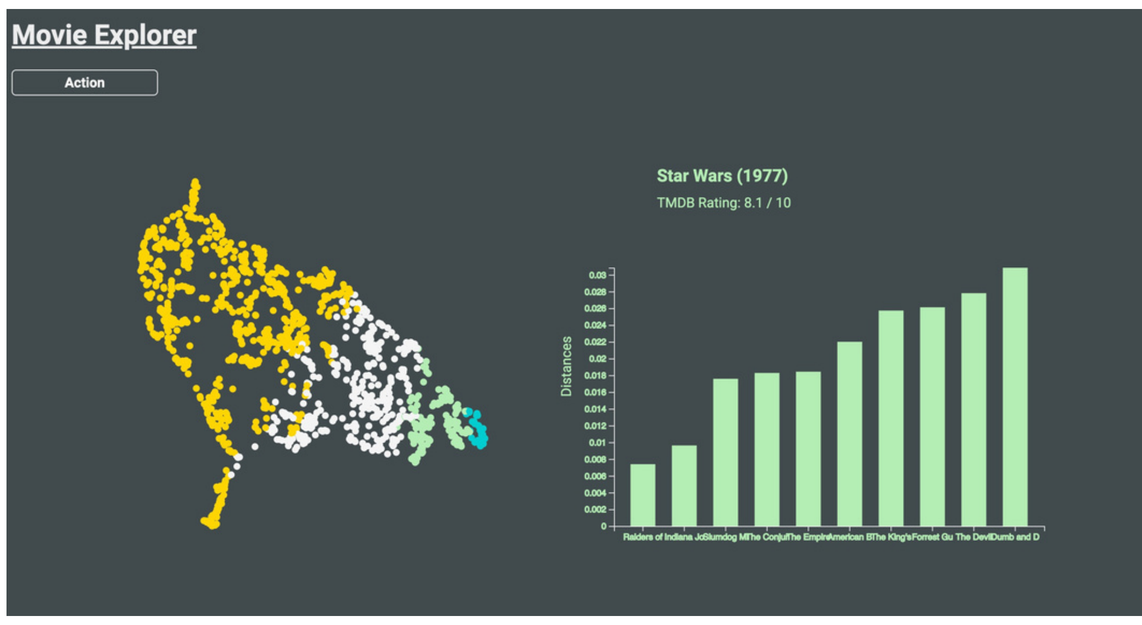
We can use the elbow method along this curve and thus choose to select the top 8 components to be used in our PCA transformation.

We then use the low level representation from PCA in our K-NN. We evaluate this approach by implementing a euclidean distance between movies searched in our prompt to the user and then evaluate across our dataset to find the top 10 nearest neighbors. This method differs from other methods since it is user-agnostic and we can just prompt the user for a movie they are interested in.



We show how using 4 clusters for each genre in K means produces a UMAP visualization. The plot represents the distinct regions where each subclass is separable by genre.

Here, we show a visualization representing the most similar movies related to Star Wars based on relative distances.


Movie Explorer
Action
Star Wars (1977)
TMDB Rating: 8.1 / 10
Distances
0.03
0.028
0.026
0.024
0.022
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
Raiders of Indiana JoSlumdog MThe ConjuThe EmpirAmerican BThe King'sForrest Gu The DevilDumb and D