## General Idea

The movie industry has been producing movies for more than a century already, and it has produced a very extensive catalogue. The whole collection of movies can be found on the IMDB website, with comprehensive information about every movie ever made. Unfortunately IMDB lacks a tool to visualize all of this catalogue entirely, and it is more useful to browse the movies' pages one at a time.

Our project, **MoVis**, aims to solve this shortcoming, providing a tool to explore visually the IMDB dataset. On top of this, our visualization will integrate information from different sources, expanding the original datasets already available on the web.

## Dataset

The main dataset used for the project is the Movie Dataset from Kaggle (https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv). It's a collection of movies taken from the IMDB website containing more than 45000 entries, with 24 different attributes. We select only some of them:

*imdb_id, title, release_date, budget, revenue, runtime, vote_average, vote_count, popularity*

Then we integrate this dataset with a *director* column, computed from the Kaggle *credits.csv* file.

The resulting dataset is joined with another one, the MovieLens 25M Dataset (https://grouplens.org/datasets/movielens/), to retrieve the movies' *genres*.

Finally, we added 4 more columns by doing a web scraping of the IMDB website to obtain the relationship between movies from every */imdbID/movieconnections* page. We consider two movies connected when one is a remake, sequel or prequel of the other, and we differentiate between incoming and outcoming connections. Also, we keep track of the total number of these connections, and which are the connected movies' id.

The final columns are:

*imdb_id, title, release_date, budget, revenue, runtime, vote_average, vote_count, popularity, movielens_id, tmdb_id, genres, director, in_connections, out_connections, tot_connections, connected_movies*

Excluding the *id* columns, we have 14 columns used in the visualization and more than 5000 rows, resulting in an AS index of 70000.

## Dimensionality Reduction

Our project will include a dimensionality reduction technique so that the user can visualize the dataset on a 2D scatterplot.

The analysis of the data done by dimensionality reduction requires a lot of computation, so it is not feasible while the user interacts with the application. For this reason we decided to do a preprocessing phase in which we run the algorithm and store the result in the dataset, adding two new columns, one for each component.

For our project we will use MDS because it allows to define and use different distance functions for different types of attributes, such as Jaccard distance for categorical data and euclidian distance for numerical data.

Possibly we will let the user choose between different implementations of the MDS algorithm which lead to different results.

## Visualizations

Our project will contain the following visualizations:

- a scatterplot, representing the MDS results;
- a bubble plot, used to visualize relations between movies' attributes, using as the third dimension the size of the bubbles. Possibly we'll also add the color of the bubbles to represent another attribute;
- some boxplots, used to plot the distribution of some of the movies' attributes;
- a parallel coordinates chart, used to visualize all of the movies' attributes at once;
- a chord diagram, visualizing movies' genres and their interpolation.

Possibly we'll also implement a navigation bar on the top of the page, with the possibility to filter the dataset, or search for specific movies, or change the color theme of the project, and offer tools to the user to further interact with the data visualization.

## User Interaction

All the visualizations in the project will be related to each other: if the user interacts with one part, the others will change accordingly. More specifically:

- the MDS scatterplot will have a brushing feature to select clusters of similar movies. Possibly there will also be an *onhover* function to display the movie name;
- the bubble plot will also have a brushing feature, allowing to select groups of movies;
- the parallel coordinates chart will have a brushing feature on each column, too;
- finally, the chord diagram will give the possibility to select movies only from a specific genre by clicking on the relative arc. It will be also possible to select more than one genre at a time, by clicking con the links between arcs.

The brushing has the effect of filtering the selected movies in all the visualizations, and as a result the movies not selected will become transparent, and the selected ones will be highlighted.