Box office Prediction

**Project definition:**

To try and build a model to predict for a given movie its overall worldwide Box Office revenue.

The data set includes past movies with information about them including the cast, the director, plot keywords… And of course, for the training data points we have the classification (total Box office revenue)

**Data Exploration:**

First step to start with is always to explore the data set we are given. This will include independently analyzing each of the features, finding correlations between the features, dimensionality reduction and feature selection.

Features:

The features we have in the data set are:

['id', 'belongs\_to\_collection', 'budget', 'genres', 'homepage',

'imdb\_id', 'original\_language', 'original\_title', 'overview',

'popularity', 'poster\_path', 'production\_companies',

'production\_countries', 'release\_date', 'runtime', 'spoken\_languages',

'status', 'tagline', 'title', 'Keywords', 'cast', 'crew', 'revenue']

Now we’ll go through each of these features and analyze their protentional contribution to the classification/their correlations with other features, and possibly construct new features using the existing ones.

For this task we will create an independent module to deal with every feature on its own, and after we’ve done this for all the features, we will use these modules to construct a new clean and modified dataset on which our model will be trained and tested.

1. Genres feature:

Movies might belong to more than one genre, the values for this feature are a list which its elements are dictionaries with a genre id and a genre name. for example the movie with the id 1 belongs to genres Comedy (genre id 35), Drama(genre id 18), Family (genre id 10751) and Romance (genre id 10749).

In:

import pandas as pd  
train\_set = pd.read\_csv('../data/train.csv')  
print(train\_set["genres"][1])

Out:

[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'name': 'Drama'}, {'id': 10751, 'name': 'Family'}, {'id': 10749, 'name': 'Romance'}]

The genre id will probably not be helpful for extracting information about the movie, it is just an overhead.

We could use the genre id *instead* of using the genre name, we prefer to use the names since it is easier to comprehend.

Since we can’t define any order for genres, a good approach for dealing with such a nominal feature is to transform it to a group of “dummy” binary features. I.e. each feature will indicate whether or not the movie belongs to the specific genre.

For example: The movie with the id 1 will have the value 1 for the newly constructed features genre\_name\_comedy, genre\_name\_drama, genre\_name\_family and genre\_name\_romance, and 0 in all the other “genre\_name\_” feature.

But just before we do that, we should look at our data to check if there are any outliers and genre names that won’t be useful (such that only few movies belong to them)